KAYTUS

KAYTUS Storage Platform KS2000G7&KS3000G7 Technical White Paper

Document Version	1.0
Release Date	06/20/2025
FW Version	7.5.1.x

Notice

This document provides detailed information on the product and helps you to understand and use it. The screenshots involved are only examples. Please refer to the interface shown on the actual device.

Due to product upgrades or other reasons, the contents of this document will be changed at any time, and subject to change without notice. Unless otherwise agreed, this document is only used as a guide, and all statements, information, and suggestions in this document do not constitute any express or implied warranty.

If you have any questions or suggestions about this document, please contact KAYTUS SYSTEMS PTE. LTD.

Contact us

Service Hotline:	+65 6611 8899 (Global) 1800 611 8899 (Singapore)
Address:	150 Beach Road, #14-05/08, Gateway West, Singapore 189720 KAYTUS SYSTEMS PTE. LTD.
Website:	https://www.kaytus.com
Technical Support:	servicesupport@kaytus.com

Security Statement

We take data security and privacy seriously, and pay attention to the security of products and solutions to provide you better services. Before you formally use the storage, read the following security statement.

1. To protect data privacy, when you change the use of storage or eliminate old storage devices, reset the software to the factory default settings, delete information, and clear logs. At the same time, it is recommended that a third-party secure eraser tool be used to fully erase the system drive where the storage software is located.
2. The storage may read and use the personal data of your users (such as email address and IP address of the alert recipients) during the business operation or fault location. The storage provides necessary security protection measures for the processing activities of the whole life cycle of personal data, such as collection, storage, use, transmission, and deletion. At the same time, you are obliged to formulate necessary privacy policies that comply with the laws and regulations of the country or region and take adequate measures to make sure that the personal data of your users are fully protected.
3. We attach great importance to product data security, and the storage provides necessary security protection measures for the processing activities of the whole life cycle of system operation and security data in strict accordance with relevant laws, regulations, and regulatory requirements. As a system operator and security data processor, you are obliged to formulate necessary privacy policies that comply with the laws and regulations of the country or region and take adequate measures to make sure that system operation and security data are fully protected.
4. For the statement of open-source components of the storage, contact our customer service directly.
5. It is necessary for some security features of the storage to be configured independently, such as authentication, transmission encryption, and storage data encryption. These configuration operations may have a certain impact on the performance and usability of the storage. You can decide whether to configure security features according to the application environment.
6. The storage comes with some interfaces/commands for production, installation, and return to factory service, and advanced commands for fault location. Improper use may cause device anomalies or business interruption. It is recommended that you contact our customer service for consultancy when necessary.
7. We have established a comprehensive response and handling mechanism for product vulnerabilities to make sure that security issues are dealt with as soon as possible. If you find any security issues when using the storage, or seek necessary support for product security, contact our customer service directly.

In the above statement, "we" and "us" refer to KAYTUS SYSTEMS PTE. LTD.; KAYTUS SYSTEMS PTE. LTD. has the final right to interpret the statement.

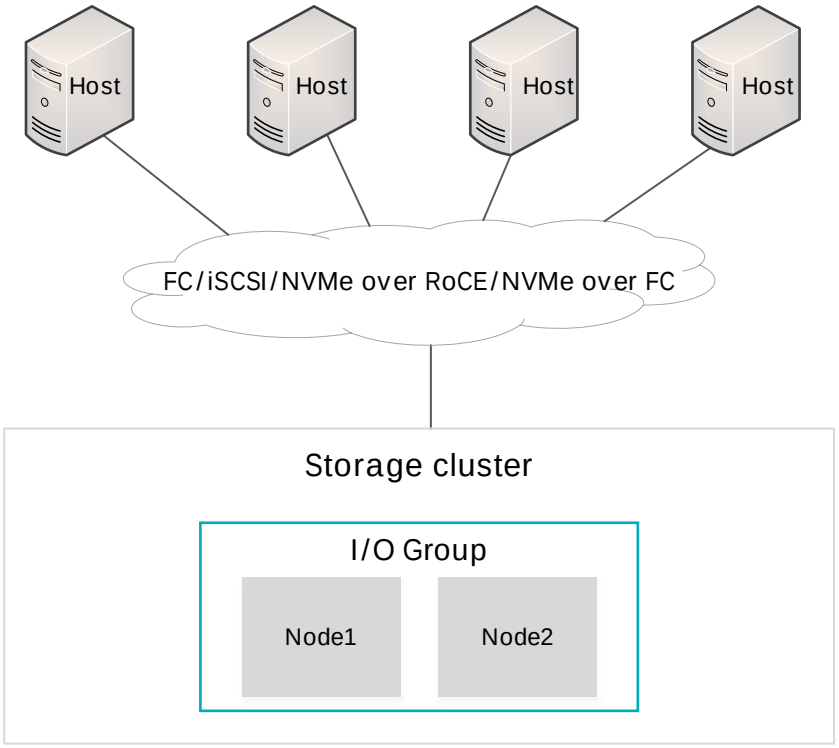
Contents

1 Storage architecture

1.1 Clustered storage

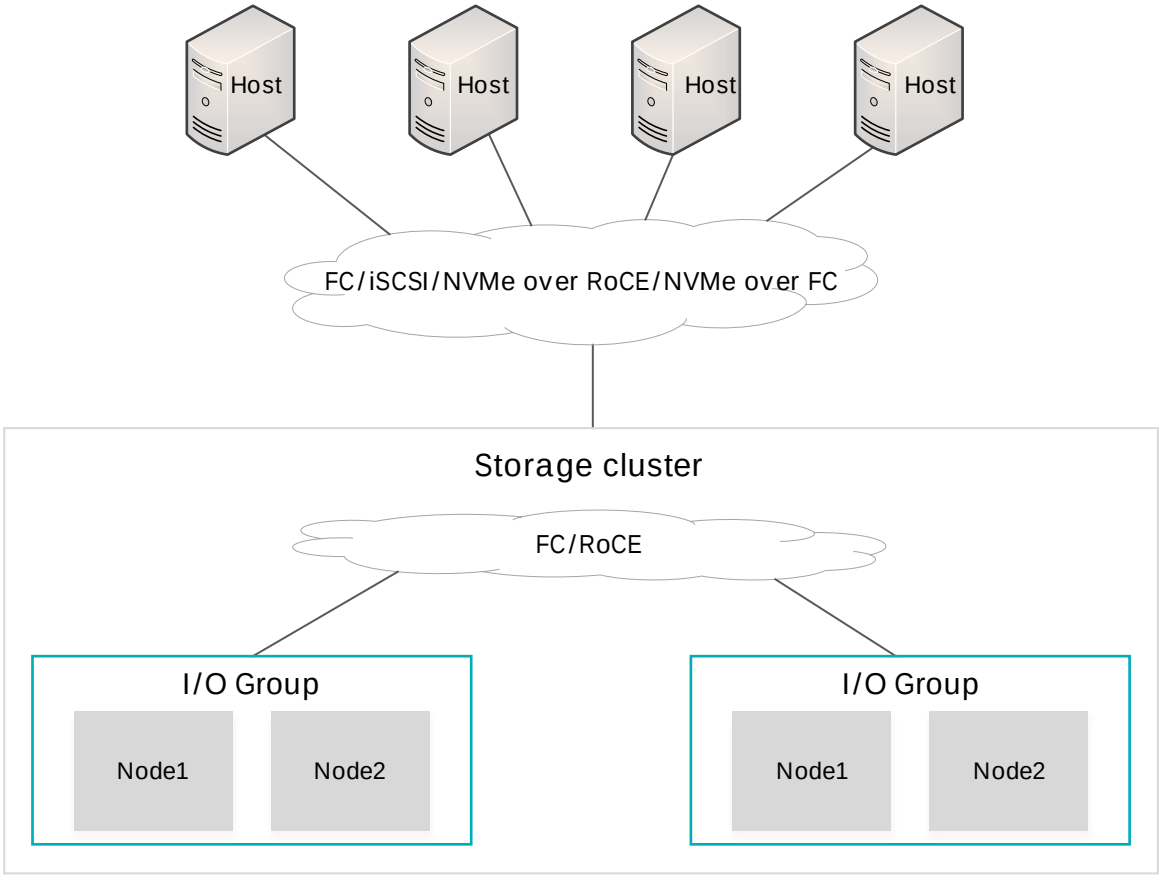
KS2000G7 can only provide storage services for front-end host business as a two-controller clustered system. The application topology of the clustered system is as shown in Figure 1-1.

Figure 1-1 Application topology of clustered system (KS2000G7)



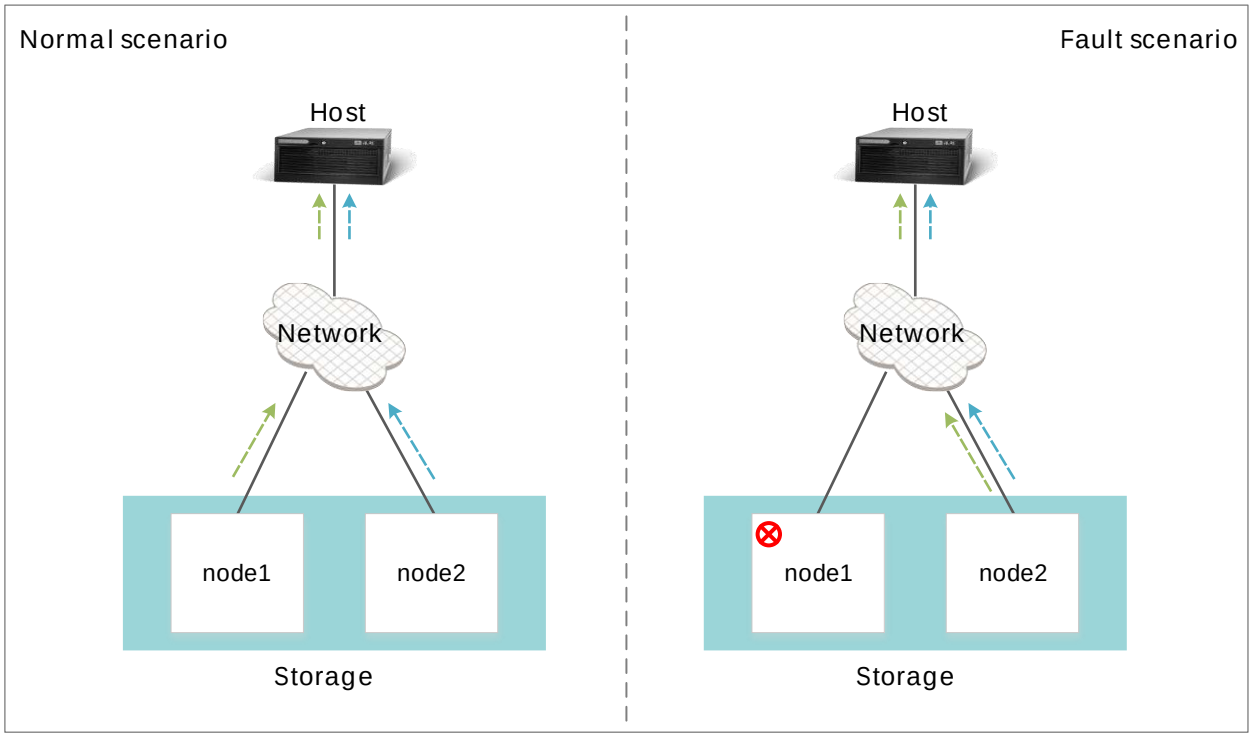
KS3000G7 can provide storage services for front-end host business as a multi-controller clustered system, which is scaled out through FC or RoCE. The application topology of the clustered system is as shown in Figure 1-2.

Figure 1-2 Application topology of clustered system (KS3000G7)



The clustered system provides a continuously available platform for host applications to make sure of business continuity, which means that the host can keep uninterrupted access in the event of any single point of failure in the storage system. During the failure, controllers other than the failed controller will take over the volumes related to the application data and remain online, providing services to the host application. Take two-controller clustered system as an example, as shown in Figure 1-3.

Figure 1-3 Failover under single point of failure



1.2 End-to-end NVMe

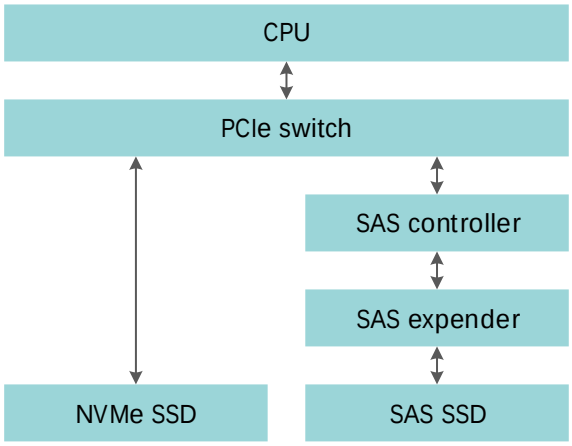
KS3000G7 supports end-to-end NVMe, connecting to front-end hosts through NVMe over RoCE and NVMe over FC, as well as connecting to back-end NVMe SSDs.

Compatible with SAS protocol and NVMe protocol

The combination of SCSI and SAS interfaces still plays an important role in the SAS architecture platform. With the development of SSD media and interface technology, NVMe architecture has gradually become the focus of subsequent storage, and SSD media and storage software architecture have been evolving towards NVMe architecture. The storage is compatible with SAS and NVMe protocols, making sure of a smooth transition of the storage.

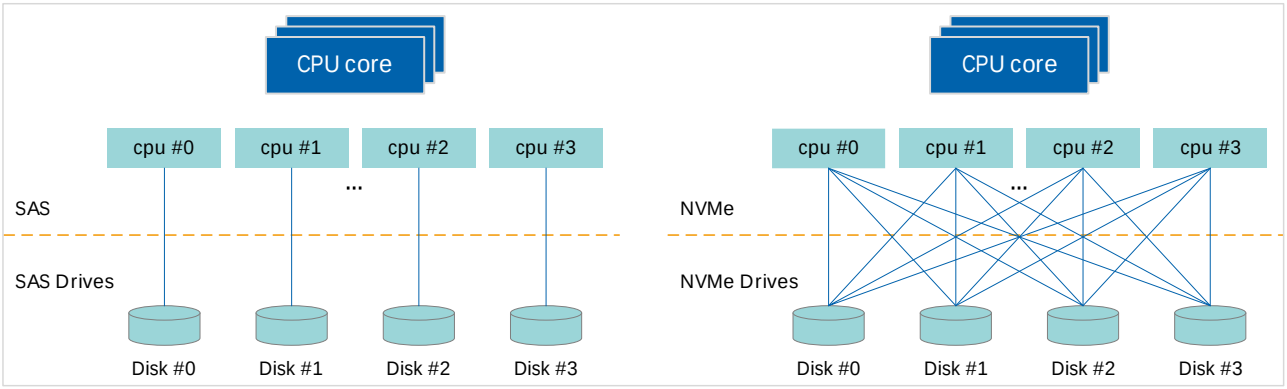
Compared with SAS protocol, the software stack of NVMe protocol is more simple and efficient. Circumventing complex SCSI subsystems, NVMe protocol greatly deceases the number of interactions. By directly connecting to the PCIe bus, the hardware transmission path of the NVMe protocol has lower latency, as shown in Figure 1-4.

Figure 1-4 Physical transmission paths of NVMe protocol and SAS protocol



NVMe protocol supports greater concurrency and queue depth, further releasing the back-end performance. The maximum number of queues in the NVMe protocol is 64 k, and the queue depth per queue is a maximum of 64 k. The multi-queue mechanism is the most important difference between NVMe protocol and SAS protocol, and the high concurrency and low coupling feature brought by the multi-queue mechanism is the key point of the high performance of NVMe protocol. The software concurrency design enables NVMe protocol to bring into play the potential better of multi-core CPUs. The comparison of software architecture between SAS protocol and NVMe protocol is as shown in Figure 1-5.

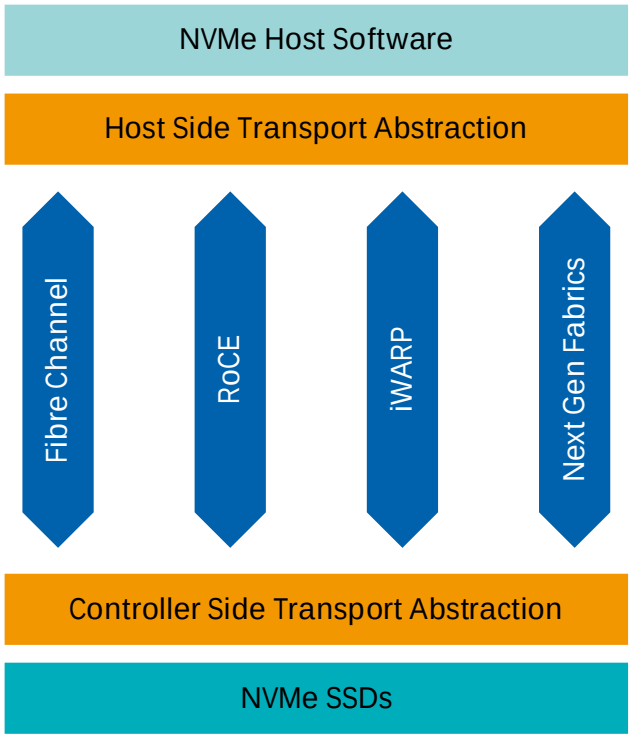
Figure 1-5 Software architecture of NVMe protocol and SAS protocol



NVMe over Fabrics protocol

NVMe over Fabrics defines a universal architecture, which defines a series of storage network structures that support the NVMe block storage protocol, corresponding to network specifications of different transport tier. The common network structures include FC and Ethernet (RoCE and iWARP).

Figure 1-6 Physical transmission paths of NVMe over Fabrics



NVMe over Fabrics enables front-end NVMe interface in the storage system, which can expand the support for a large number of NVMe devices, decrease data transmission latency, and increase the distance between accessible NVMe devices and NVMe subsystems within the data center.

About 90% of network-based NVMe protocols are the same as local NVMe protocols, with the main differences in four aspects, as shown in Table 1-1. Based on the definition of NVMe over Fabrics, NVMe subsystems are identified by NQN. The discovery service is implemented through the newly added discovery and connection commands in the NVMe over Fabrics specification and getting discovery logs. The queue model has added a message-based method. The data only support SGLs for transmission.

Table 1-1 Main differences between NVMe over Fabrics and PCIe

Function	PCI Express (PCIe)	NVMe over Fabrics
Identifier	Bus, device, function (BDF)	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and connection commands
Queue	Memory-based	Message-based
Data transmission	PRP or SGL	SGL only, with key added

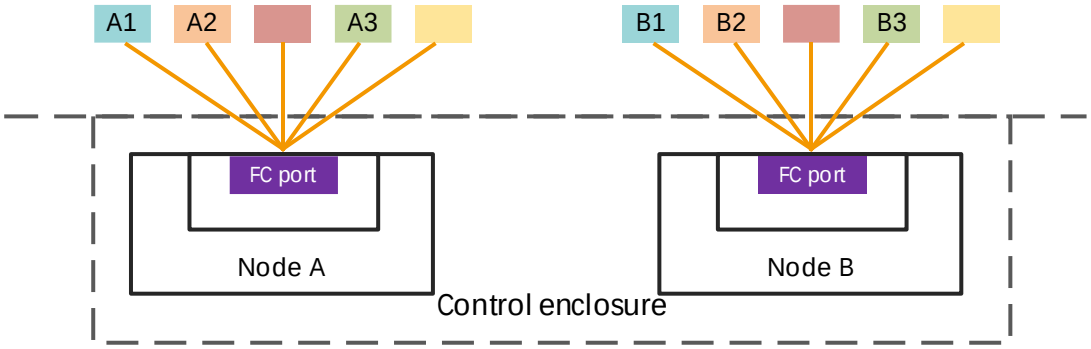
1.3 NPIV technology

KAYTUS storage supports NPIV technology, letting a single HBA card (referred to as an N_Port) to register multiple world wide port names (WWPNs) and N_port IDs. This enables the storage system to virtualize

multiple ports even with only one physical port. To achieve interoperability with other devices, switches must also support NPIV technology.

The NPIV technology of KAYTUS storage enables one FC physical port to virtualize five ports, each having a WWPN for connection with other devices, as shown in Figure 1-7.

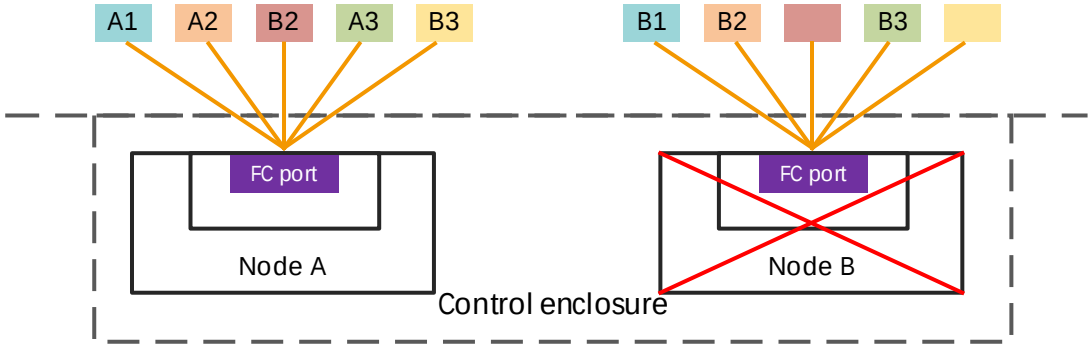
Figure 1-7 NPIV technology for FC ports



The virtual ports can be used in the scenarios that follow.

- Connect the application host as a target port.
- Connect a third-party storage as an initiator port.
- For communication between I/O groups within the cluster.
- In NVMe over FC scenarios, host NVMe and SCSI protocols with different virtual ports, reducing the mutual impact of different businesses.
- In failover scenarios, virtual ports can replace multipath function. The corresponding failover ports of the two controllers have the same WWPN and can seamless switchover in the event of a failure. When controller B is failed, the businesses sent from the host to the ports B2 and B3 on controller B will switch to controller A, as shown in Figure 1-8.

  Figure 1-8 Failover scenario



1.4 Software components

The software provided by KAYTUS storage can manage storage devices and the stored data, helping application servers in data operations. With the coordination of core software modules, host plug-ins, and system management and control software, various backup and disaster recovery functions can be implemented intelligently, efficiently, and economically.

Core software modules on storage system

InVirtualization, InCompression, InRemoteCopy, InMetro, InErase, and InEncryption are license-controlled functions, and other functions can used without a license. To use license-controlled functions, the related license must be activated.

Table 1-2 Core software modules on storage system

Software module	Description
Cluster management	KAYTUS storage can provide storage services for front-end host business as a multi-controller clustered system. The clustered system provides a continuously available platform for host applications to make sure of business continuity, which means that the host can keep uninterrupted access in the event of any single point of failure in the storage system.
InRAID	Distributed RAID (DRAID) is also known as InRAID. Compared with traditional RAID, the hot spare space of DRAID is not on single hot spare drives, but distributes across all member drives, and a single DRAID can support more drives. Currently, it supports DRAID5 and DRAID6.
Storage pool	A storage pool is composed of one or more MDisks. All MDisks in the storage pool are split into extents with the same size. In a normal pool, extents with the same size constitute a volume. In a thin pool, extents with the same size constitute a log volume. Above the log volume, a thin volume is provided by the thin pool. Note: Only KS3000G7 whose per-controller cache is 128 GB and above supports thin pools.
InVirtualization	By logical abstract, InVirtualization can eliminate the differences between different types of storage resources, and get resource integration and utilization. The integrated storage devices can have the same advanced features as the local storage, and the administrator can manage all storage devices at the same time as one storage device, decreasing management costs.
InMigration	InMigration can achieve data migration within an I/O group, including data migration within the same storage pool and between different storage pools, as well as data migration across I/O groups. During data migration, the storage system will rebuild the storage location information of the data. When data migration is completed, the host application can directly connect to the new data volume.
InThin	InThin is a new capacity allocation technology that does not allocate a large enough capacity to the application host in one time, but allocates capacity to the application host for multiple times and in small amounts per time referring to the real capacity demand of the application host. The data generated by

	the application host will increase continuously, and the storage system will allocate part of capacity from the back-end storage pool again when the allocated capacity cannot meet the demand.
InCompression	InCompression is a volume-based compression technology that enables real-time compression of data before the data are written to drives. InCompression is also a hardware-based compression technology, which provides compression function through the QAT compression engine, reducing compression costs and improving the throughput and compression performance of compression devices. Note: Only KS3000G7 whose per-controller cache is 128 GB and above supports InCompression.
InDedupe	InDedupe is a pool-based duplicate data detection and reduction technology, implemented through software, without any hardware. Before data are written to SSDs, real-time detection and deduplication will be done, which can decrease the number of write and the amount of data written to SSDs. InDedupe can save storage spaces, decrease wear on SSDs, and extend the service life of SSDs. Note: Only KS3000G7 whose per-controller cache is 128 GB and above supports InDedupe.
InAutoPartition	InAutoPartition can make sure of the performance of crucial businesses by allocating private cache resources for crucial businesses. At the same time, it improves the utilization of cache resources by providing a cache resource sharing mechanism for non-crucial businesses.
InQoS	InQoS allocates resources of the storage system reasonably through flow control to meet the specific performance requirements of certain applications, which can make sure of the quality of service for crucial businesses.
InLocalCopy	InLocalCopy is a copy function implemented within a cluster, also known as point-in-time copy. On the principle of copy-on-write (COW), InLocalCopy enables the creation of an instantly available consistent copy while the data is online and available.
InRemoteCopy	InRemoteCopy can create a copy mapping between two volumes of the same capacity, making remote backup of data, which can used for data protection or disaster recovery. Two volumes may be in the same I/O group or different I/O groups of the same clustered system, or in I/O groups of different clustered systems.
InMetro	InMetro is implemented based on a clustered system, where two I/O groups can manage each volume at the same time. InMetro uses synchronous remote copy technology, change volume technology, and InMigration technology. Note: InMetro is only available for KS3000G7.

InErase	InErase is used to erase data on a VDisk that are no longer necessary. The erase method is overwriting and the erase unit is vdisk.
InEncryption	InEncryption is used to encrypt data stored in the storage system, providing strong protection for sensitive data. InEncryption can effectively prevent illegal data access or data leakage during storage and transmission, making sure of data security and data integrity. Note: InEncryption is only available for KS3000G7.
Event monitoring	The storage management system can monitor the status of the storage system in real-time, visually show important health information, and send timely notification to the administrator.
DMP	The directed maintenance procedures (DMP) provides an intelligent failure recovery mechanism, including failure diagnosis and recovery steps. It provides guidance for software and hardware failures to quickly and efficiently resolve problems and restore businesses, which can decrease maintenance costs and shorten recovery time.

Plug-ins for application hosts

Table 1-3 Plug-ins for application hosts

Platform	Software	Description
Multipath	InPath for AIX	InPath for AIX is a multipath package, which is developed based on the multipath framework of AIX platform. It provides basic functions such as failover, failback, fault isolation, and path selection strategies, as well as unique functions such as I/O statistics, performance monitoring, and log tracing.
	InPath for Linux	InPath for Linux is a multipath package, which is developed based on the multipath framework of Linux platform. It provides multipath functions such as failover, failback, fault isolation, path selection strategies, and multipath configuration management.
	InPath for VMware	InPath for VMware is a PSPs multipath package, which is developed based on the multipath framework of VMware platform. It provides multipath functions such as failover, failback, path selection strategies, and multipath configuration management.
	InPath for Windows	InPath for Windows is a multipath package, which is developed based on the multipath framework of Windows platform. It provides basic functions such as failover, failback, fault isolation, and path selection strategies, as well as

		unique functions such as I/O statistics, performance monitoring, and log tracing.
OpenStack	Cinder plug-in	Cinder plug-in is a driver for Cinder service in OpenStack. Through Cinder driver, Cinder can connect and use the storage to provide block storage service for OpenStack, improving Cinder performance.
	Cinder-backup plug-in	Cinder-backup plug-in is a driver for backup in Cinder service of OpenStack. It can use the snapshot function of the storage to create backups for Cinder volumes and manage the backups.
	Manila plug-in	Manila plug-in is a driver for Manila service in OpenStack and can manage the storage through Manila service. It provides the NFS and CIFS share services for OpenStack.
Container	K8sCSI plug-in	K8sCSI plug-in enables the storage to provide persistent storage automatically for applications in the Kubernetes cluster, including creating, deleting, mounting, unmounting, and cloning volume, and creating and deleting snapshot.
Microsoft	VSS plug-in	VSS plug-in can use the provider interface of the Microsoft VSS service framework to save the data of local copy mappings into the back-end storage. It can also call the InLocalCopy function of the storage to create and manage local copy mappings.
VMware	SRA plug-in	SRA plug-in is an adapter of the storage, attached to VMware Site Recovery Manager (SRM), enabling SRM to call the InRemoteCopy function of the storage.
	VCP plug-in	VCP plug-in provides a separate server background for integrating storage resources from different models of storage systems for unified and centralized management.

System management and control software

Table 1-4 System management and control software

Software	Description
InStorageManager	The storage system provides a storage management software called InStorageManager, which supports direct access through a browser on PC, enabling effective management, configuration, and maintenance. It also supports access through a browser on mobile phone to do a check of the system status, performance, and resource usage.

Command-line interface (CLI)	The storage system provides CLI-based management and configuration, which supports the use of third-party terminal software to log in with SSH protocol, enabling professional management, configuration, and maintenance.
Service assistant	The storage system provides a service assistant for professional maintenance personnel to manage the storage system.
SSH	SSH is a protocol that provides security for remote login sessions and other network services, effectively preventing information leakage during remote management processes.
SMI-S Server	The storage system supports SMI-S protocol and can be managed and configured through third-party network management software that supports SMI-S protocol.
CIM interface	CIM Client can monitor and manage the storage system by calling the CIM interface provided by the storage, including cluster management, host management, storage resource management (pools, volumes, and hard disks), performance monitoring, and data protection.
Web Server	Web server can parse the HTTP protocol. The storage supports HTTP protocol and can be managed and configured through network management software that supports the HTTP protocol.
SNMP interface	The storage system provides a customized interface that complies with SNMP specifications to be called by third-party systems or software to manage or monitor storage devices. SNMP GET/SET supports v1, v2c, and v3, while SNMP trap supports v2c and v3.
RESTful interface	The storage system provides an interface that complies with REST constraints to be called by third-party systems or software to manage or monitor storage devices.

2 Elastic storage

2.1 InRAID

Feature description

Distributed RAID (DRAID) is also known as InRAID. Compared with traditional RAID (TRAID), the hot spare space of DRAID is not on single hot spare drives, but distributes across all member drives, and a single DRAID can support more drives. Currently, it supports DRAID5 and DRAID6.

Feature principle

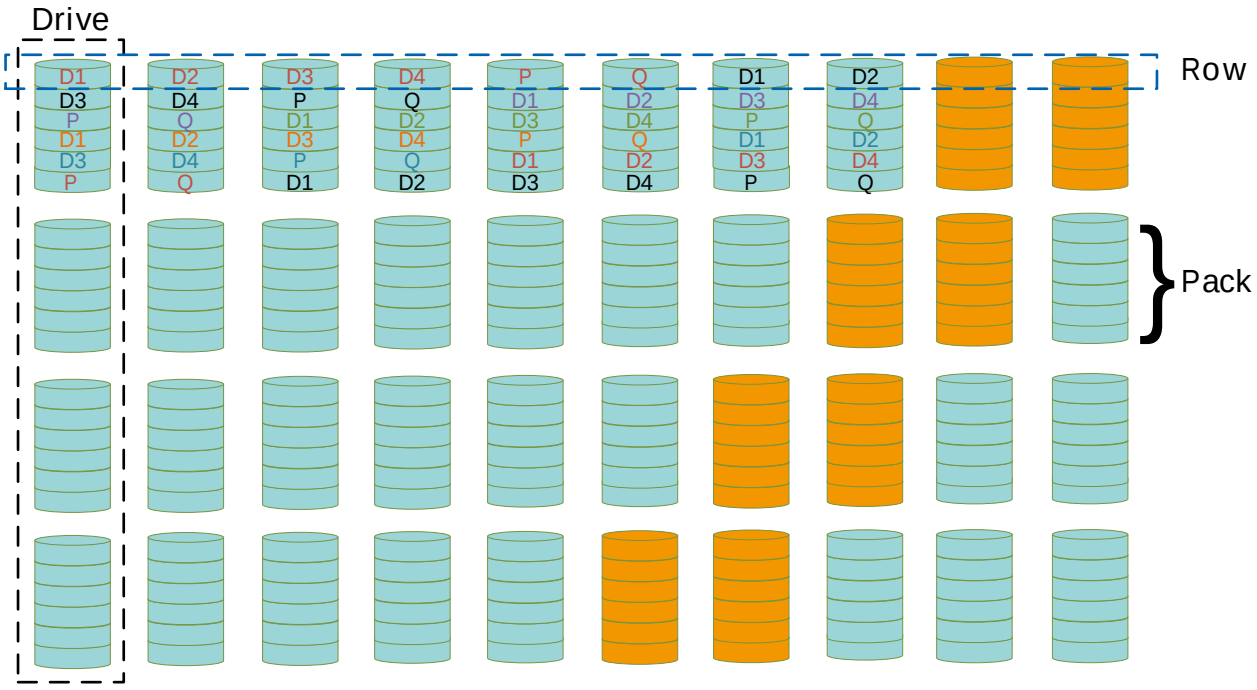
RAID technology combines multiple independent physical drives into a logical drive in different ways, improving the read-write performance of drives and data security. In TRAID, data are distributed on member drives while the hot spare space consists of independent drives. In case of a drive failure, the verification mechanism of RAID enables data on the failed drive to be read from other member drives and placed on a hot spare drive.

In DRAID, the spare drive (or drives) are included in the array as spare space. All drives in the array have spare space reserved that is used if one of the drives in the same array fails. In case of a drive failure, the verification mechanism of RAID enables data on the failed drive to be read from other member drives and placed on the hot spare space. Both data and its parity check information are also distributed on all member drives. Thus, more drives can be selected to create a DRAID.

In DRAID, the distribution of hot spare space depends on the pack, and all drives in the array are divided into rows and packs. The size of a row is the number of all drives that make up a DRAID. A pack is a group of several continuous rows and its size depends on the number of strips in a stripe. Thus, the value of Pack is constant in a DRAID.

Taking DRAID6 as an example, six stripes (4+P+Q) and two hot spare spaces with capacity of two drives are distributed on 10 drivers, as shown in Figure 2-1. The row (array width) is 10 and the pack is six.

Figure 2-1 DRAID6



Technical specifications

The technical specifications of DRAID are as shown in Table 2-1.

Table 2-1 Technical specifications of DRAID

Attribute	Value
Maximum number of DRAID arrays per system	32
Maximum number of DRAID arrays per I/O group	10
Maximum number of member drives per DRAID array	128
Number of member drives per DRAID-5 array	4 to 128
Number of member drives per DRAID-6 array	6 to 128
Number of rebuild areas per DRAID array	1 to 4
Stripe width for DRAID-5 array	3 to 16
Stripe width for DRAID-6 array	5 to 16
Maximum capacity of member drives for RAID-5 array	< 8 TB

The recommended configurations of DRAID are as shown in Table 2-2.

Table 2-2 Recommended configurations of DRAID

Attribute	Value

Array width (number of member drives) of DRAID	48 to 60
Number of rebuild areas with array width less than 36	1
Number of rebuild areas with array width of 36 to 72	2
Number of rebuild areas with array width of 73 to 100	3
Number of rebuild areas with array width of 101 to 128	4

Relationship with other features

DRAID is the basic feature of the storage system and has no conflict with other advanced features.

Benefits

- There are no independent and idle hot spare drives in DRAID. In case of a drive failure, all member drives participate in the rebuilding work, eliminating the write bottleneck of a single drive, effectively improving the rebuilding speed, and reducing the risk of RAID failure in case of a second drive failure.
- All drives of DRAID participate in I/O operations, with no idle drives, improving the utilization and performance of drives.
- The number of drives that make up a DRAID is a maximum of 128, which can meet different performance requirements.

Application scenarios

DRAID is applicable to scenarios with higher requirements for the reliability and performance of storage systems.

2.2 Storage pool

Feature description

In the storage system, the drives of the control enclosures and disk enclosures are combined to form different types of RAID, and a RAID is an MDisk. The storage system supports heterogeneous storage virtualization, and each volume mapped from external storage systems can form an MDisk. For more information, see 2.3 InVirtualization.

The storage system supports normal pools and thin pools. A storage pool is composed of one or more MDisks. All MDisks in the storage pool are split into extents with the same size. In a normal pool, extents with the same size constitute a volume. In a thin pool, extents with the same size constitute a log volume. Above the log volume, a thin volume is provided by the thin pool.

Log volumes, also known as LSA volumes, are unique to thin pools and are used for space management

for thin volumes. Each thin pool contains a data space and a metadata space, which are respectively used to store the user data and generated metadata.

The composition of a normal pool and a thin pool is as shown in Figure 2-2 and Figure 2-3.

Figure 2-2 Composition of a normal pool

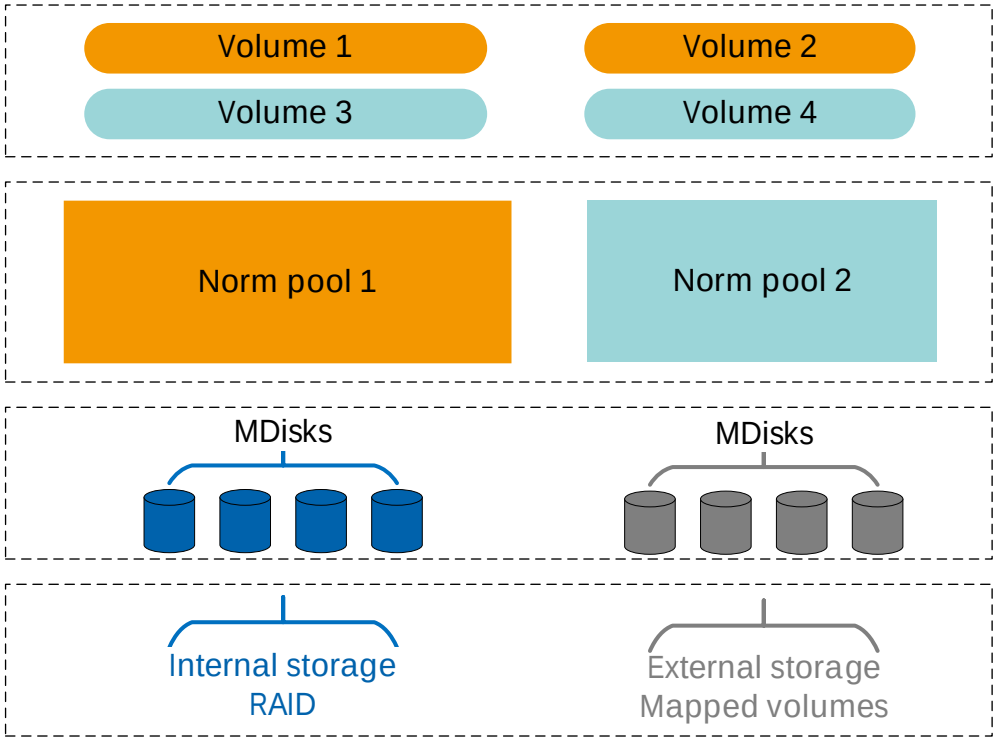
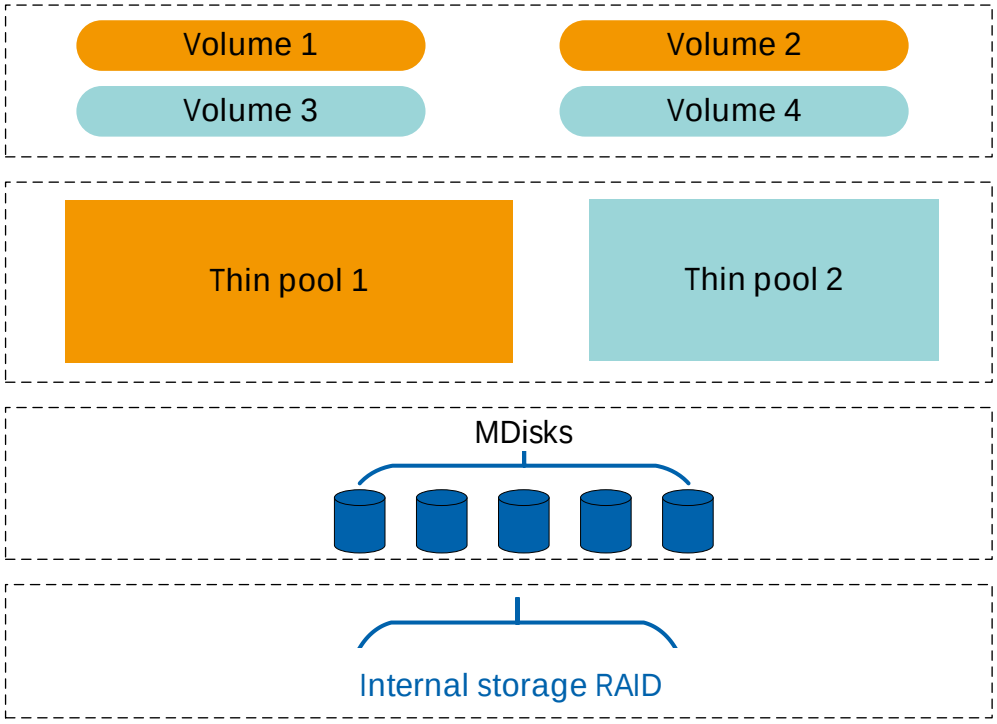


Figure 2-3 Composition of a thin pool



Feature principle

The storage system supports two redundancy policies, including disk redundancy and enclosure redundancy, which can provide data protection in different levels to meet the requirements for redundancy and data recovery capabilities of different business scenarios. The enclosure redundancy policy has tolerance of a single disk enclosure failure, providing higher reliability.

In a normal pool, it is necessary that space for volumes be allocated in advance, which makes sure that volumes can be used normally. For thin volumes, space for volumes can be allocated in multiple times instead of full allocation in one time. Once allocated, the space of each volume cannot be changed. When data are written to volumes each time, it will overwrite existing blocks.

In a thin pool, it is no longer that constant space be allocated to all volumes in advance. A unified allocation mode is used to manage the space and space is allocated based on the amount of data written. When the data are written to the volume each time, the LSA volume in the thin pool will allocate appropriate space based on the data size. Through unified management, the thin pool can control the storage location and organization of data in the array, which can optimize the data write process and improve the write performance.

With the increase of business, the space of a storage pool space will be gradually decreased until used up. If capacity expansion is not completed in time, it may have an unwanted effect on the business. To prevent the total capacity of volumes from greatly exceeding the actual total capacity of the storage pool, the storage system provides overcommitment ratio of storage pools to limit the total capacity of volumes to a specified percentage of the storage pool capacity when they are created.

Technical specifications

The technical specifications of storage pools are as shown in Table 2-3.

Table 2-3 Technical specifications of storage pools

Attribute	Value
Maximum number of MDisks per system	4096
Maximum number of MDisks per storage pool	128
Total manageable capacity of normal pools per system	32 PB
Maximum manageable capacity of a normal pool	32 PB
Total manageable capacity of thin pools per system	204 TB
Maximum manageable capacity of a thin pool	204 TB
Maximum number of normal pools per system	1024
Maximum number of thin pools per system	4
Maximum number of volumes per system	KS2000G7: 2048

	KS3000G7: 4096 - (number of thin pools*6)

Note:

- Storage pools include normal pools and thin pools.
- The maximum number of volumes in a single storage pool is the same as the system.
- The technical specifications of a single I/O group is the same as the system.
- Multiple attributes cannot reach the maximum values at the same time, which should be considered in parallel.

The size of extents has a direct effect on the maximum capacity of a normal volume in a normal pool, and their relationship is as shown in Table 2-4.

Table 2-4 Relationship between extent and capacity of a normal volume

Extent size (MB)	Maximum capacity of a normal volume (TB)
128	16
256	32
512	64
1024	128
2048	256
4096	512
8192	512

The extent size of a thin pool is constant at 2,048 MB, and the maximum capacity of a DRAID MDisk is 272 TB. The size of extents has a direct effect on the maximum capacity of an MDisk in a normal pool, and their relationship is as shown in Table 2-5.

Table 2-5 Relationship between extent and capacity of an MDisk

Extent size (MB)	Maximum capacity of a DRAID MDisk (TB)
128	256
256	512
512	1024 (1 PB)
1024	2048 (2 PB)
2048	4096 (4 PB)
4096	8192 (8 PB)
8192	16384 (16 PB)

Relationship with other features

Storage pool is the basic feature of the storage system and has no conflict with other advanced features.

Benefits

- Support disk redundancy and enclosure redundancy, which can provide data protection in different levels to meet the requirements for redundancy and data recovery capabilities of different business scenarios.
- Support setting overcommitment ratio, which can prevent unwanted business effect from too much overcommitment.
- The extents making up volumes in a storage pool come from different MDisks, which can improve the I/O performance in read-write operations, and increase the reliability.
- Based on the feature of SSDs, storage pool uses sequential write instead of random write to meet the high-performance requirements of business systems.

Application scenarios

Storage pool is the basic function and is applicable to all scenarios.

2.2.1 Full-stripe sequential write

Feature description

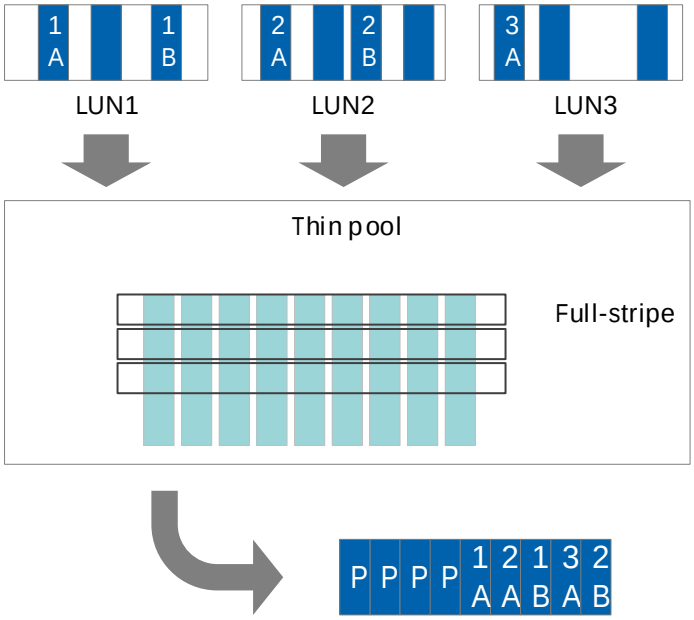
Thin pools can allocate and manage the space in a unified manner, which can enable full-stripe sequential write of the data. When thin pools receive random data from different volumes and write them to SSDs, full-stripe sequential write can change the random data to sequential data and write them to SSDs as a full stripe of RAID in one time. Full-stripe sequential write can effectively prevent the write penalty of write RAID.

Feature principle

The sequential write is achieved through redirect-on-write (ROW) technology. Whether for writing new data or changing existing data, a space will be reallocated sequentially each time the data are written. Thus, regardless of which volume and location the data are written from, the data in a thin pool will be placed in the sequence as they are written. These sequential data will be written to SSDs after they form a full stripe.

The conversion process for writing new data is as shown in Figure 2-4. After 1A, 1B, 2A, 2B, and 3A are written to the three volumes in a thin pool, the data will be placed in a full stripe in the sequence as they are written. The system will do a check of the stripe to see if it is full. If yes, the system writes the data of the full stripe to SSD in one time.

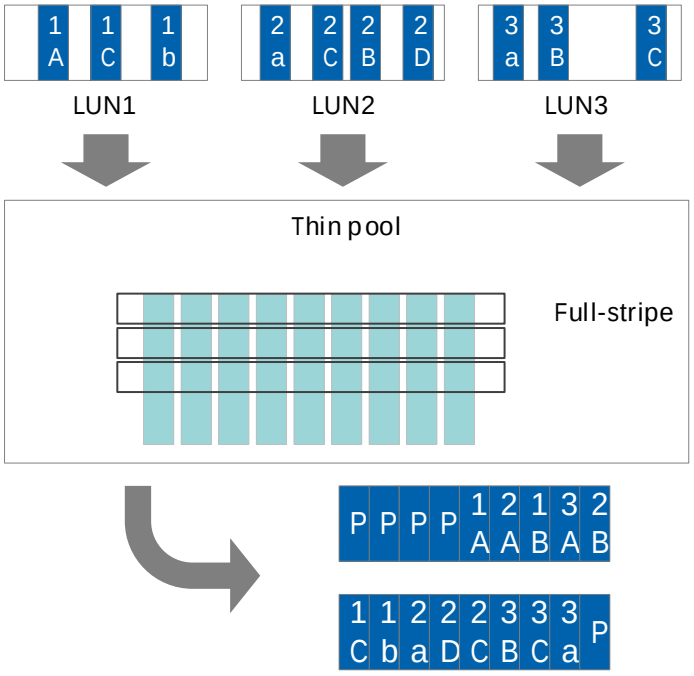
Figure 2-4 Conversion process for writing new data



The conversion process for writing new data and changing existing data is as shown in Figure 2-5. Write 1C, 2C, 2D, 3B, and 3C, and change 1B to 1b, 2A to 2a, and 3A to 3a. The data will be placed in a new full stripe in the sequence as they are written. Whether for newly written data or changed data, new space will be allocated for them.

At the same time, 1B, 2A, and 3A that have been written will become garbage data. With the new data being written successively, the amount of garbage data in the thin pool will also increase. These garbage data will be cleared through a garbage collection mechanism to free the occupied space and make sure that newly written data can get available full-stripe space continually.

Figure 2-5 Conversion process for writing new data and changing existing data



Technical specifications

To get the best write performance and minimum read/write penalty for full-stripe sequential write, the recommended configuration of a thin pool is 8+1 DRAID5 or 8+2 DRAID6.

Relationship with other features

Full-stripe sequential write is the basic feature of thin pools and has no conflict with other advanced features.

Benefits

- Convert random write to full-stripe sequential write, improving write performance.
- Prevent random small I/Os from writing data to drives, reducing the read/write penalty.
- All SSDs can be written at full-block, reducing write amplification and improving performance.
- Realize global balanced write of data in thin pools, preventing repeated write to a fixed position on a volume and frequent write to one SSD, slowing down the internal wear of SSDs.

Application scenarios

Applicable to application scenarios containing a large number of random small I/Os. Full-stripe sequential write can effectively convert random small I/Os into continuous large I/Os, improving write efficiency and overall write performance.

2.2.2 Garbage collection

Feature description

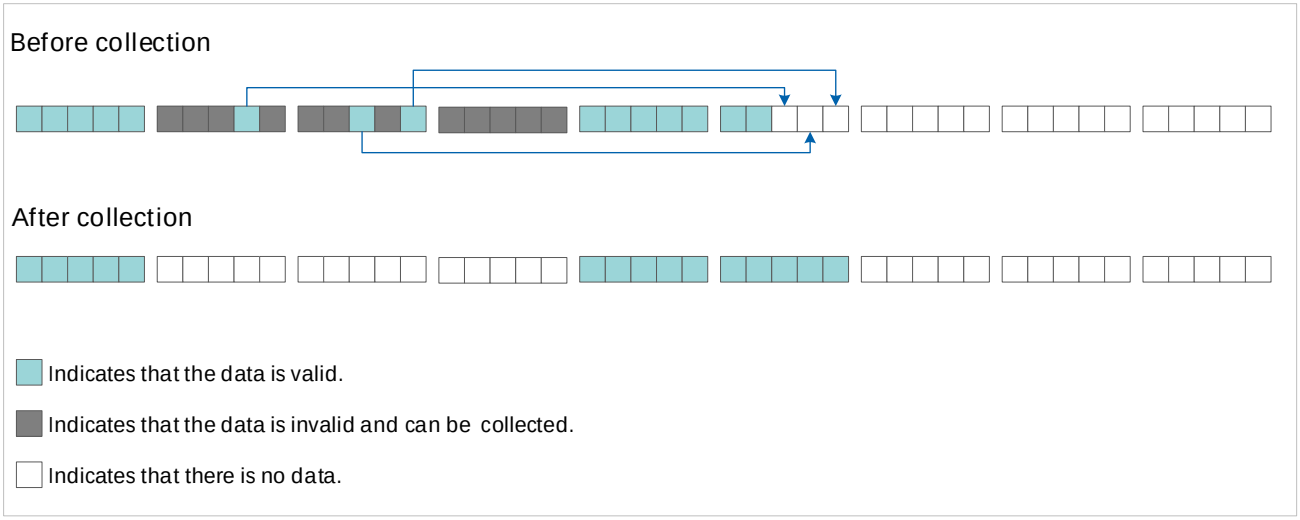
In a thin pool, extents with the same size constitute a log volume. Garbage collection is used to find and reclaim space containing garbage data in the log volume, thus freeing new space available for the log volume.

Feature principle

The log volume uses ROW method for host write I/Os. The latest data of the same LBA is saved in a new location on the log volume. The previously saved data in the log volume become garbage data, the space occupied by garbage data becomes invalid, and new data cannot be written until the space is reclaimed.

The unit of space allocation for host write I/O on the log volume is grain, while the unit of garbage collection on the log volume is block, which may contain multiple grains. During the garbage collection, blocks containing a large amount of garbage data will be found. When the block contains valid data, the valid data will be migrated to other block and its related metadata will be changed. Then, the block space is reclaimed and marked as free space, and the log volume can write data to the block.

Figure 2-6 Garbage collection



To make sure of the efficiency of garbage collection, it is necessary for a thin pool to reserve a certain amount of OP space, which can make the blocks of the log volume have a certain rate of invalidity, thus reducing the amount of data to be migrated during block reclamation.

Garbage collection should make sure of the space requirements for writing data to the log volume. Otherwise, the business will be interrupted if there are no space for the host to write business data. Thus, multiple controllers are necessary for garbage collection to reclaim multiple blocks concurrently to make sure of space reclamation speed.

Data will be migrated and metadata will be changed during garbage collection, which occupies certain processing capability of arrays for host I/Os, thus having effect on the performance of host business. To make sure of space reclamation speed and meet the space requirements of host I/Os, as well as minimize the impact on the performance of host business, flow control is necessary for the intensity of garbage collection.

Relationship with other features

Garbage collection is the basic feature of thin pools and has no conflict with other advanced features.

Benefits

Garbage collection provides a space reclamation function for log volumes in thin pools, which can effectively utilize the role of log volumes and improve the performance of SSDs.

Application scenarios

Applicable to high-performance low-latency scenarios, which can utilize the advantages of SSDs with high bandwidth and low latency.

2.2.3 Cold and hot data separation

Feature description

Cold and hot data separation refers to the ability of a thin pool to identify cold and hot data based on data type and changed frequency when data are written to SSDs, so that cold and hot data can be stored separately.

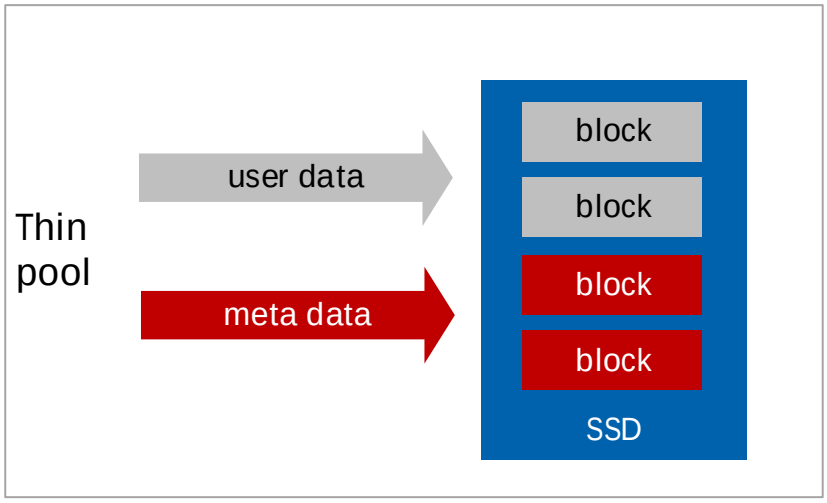
Feature principle

The drives used in a thin pool are all SSDs, which use NAND flash memory. An important feature of NAND flash is that erase operation must be done before write operation. The basic storage unit of NAND flash is page. Pages are grouped into blocks, which are the smallest area than can be erased. Thus, before writing data to a page, it is necessary to erase the block where the page is located. The best case for SSDs to work is that all data in the block are invalid. Only in this way can data migration occur as little as possible during garbage collection within a SSD.

To get this, when data are written to a thin pool, the cold and hot data are identified based on the type and changed frequency of the data, and the identifiers of cold and hot data are added during the writing process. The cold and hot data are sent to SSDs through different channels. For example, user data are usually identified and marked as cold data, while metadata are usually identified and marked as hot data. After receiving the data, SSDs will identify the hot and cold data identifiers and store them in different data partitions. Different data partitions occupy different blocks in SSDs, thus achieving separation of hot and cold data at the block level.

The example of cold and hot data separation is as shown in Figure 2-7.

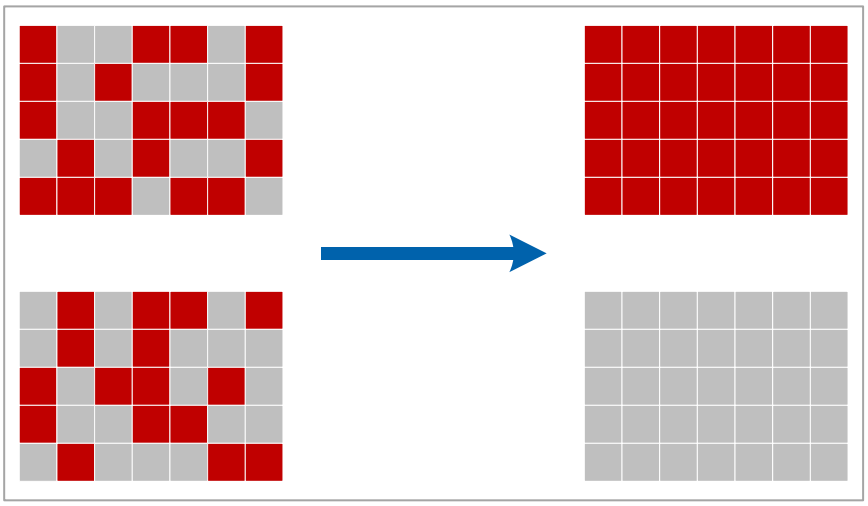
Figure 2-7 Cold and hot data separation



The data in cold data partitions will not be changed frequently, which can prevent frequent garbage collection and data migration within the SSD. The data in hot data partitions will be changed frequently, so most of the data in hot data partitions have become garbage data. Thus, space reclamation can be done directly without too much data migration during garbage collection.

The distribution of cold and hot data after being separated is as shown in Figure 2-8, where the red represents hot data and gray represents cold data.

Figure 2-8 Distribution of cold and hot data



Before cold and hot data separation, the hot and cold data are stored in two blocks in a mixed manner. Because the hot data are changed frequently, all the hot data in the two blocks may become garbage data after a period. However, the cold data are not changed frequently, thus all of them may stay valid. In this case, the valid data in both blocks must to be migrated to other blocks for garbage collection so that both blocks can be reclaimed, which will cause large quantities of data migration.

After cold and hot data separation, the data in the first block are all hot data. After the hot data become garbage data, data migration is not necessary and the block can be directly reclaimed. The data in the second block are all cold and valid data, space reclamation is not necessary for the block, thus data migration is not necessary too.

Relationship with other features

Cold and hot data separation is the basic feature of thin pools and has no conflict with other advanced features.

Benefits

- Store cold and hot data separately, reducing the quantity of data that must be migrated during garbage collection and the number of times for garbage collection.
- Improve garbage collection efficiency, decrease the program/erase (P/E) cycles on SSDs, and extend the service life of SSDs.
- Improve the overall performance of the system, and decrease the effect of garbage collection on performance.

2.2.4 Global wear leveling

Feature description

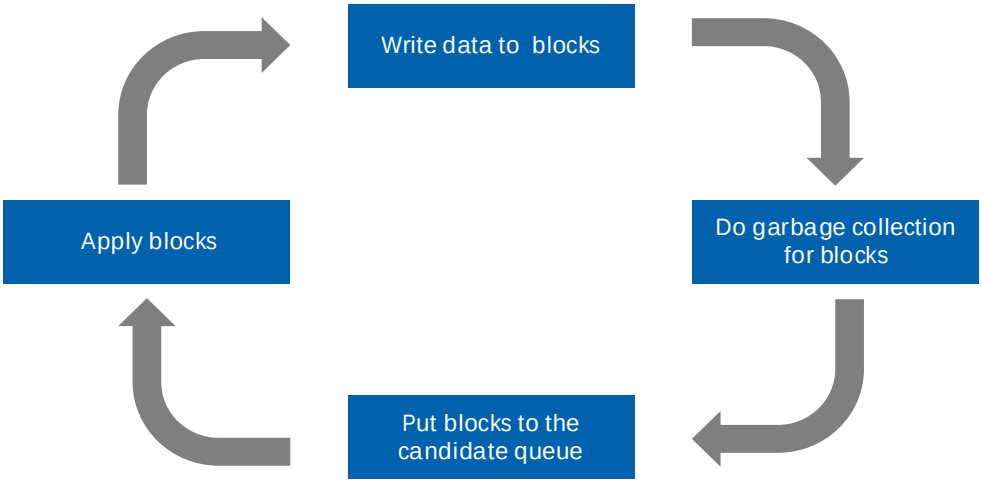
The program/erase (P/E) cycles of spaces in a thin pool are different, which causes certain SSDs to be used more frequently and wear faster than others. Global wear leveling can enable a thin pool to allocate and use all space as equally as possible, leveling the wear degree between all SSDs.

Feature principle

The space of a thin pool will be divided into several blocks, which are used for data writing. The thin pool will record the allocation and use times of each block, and sort the blocks based on their use times. When it is necessary to allocate a block, the block with the least use will be allocated first, thus leveling the use frequency of all blocks.

The allocation and usage process of blocks is as shown in Figure 2-9.

Figure 2-9 Allocation and usage process of blocks



Some blocks in a thin pool are used to store valid data permanently, garbage collection and data writing is not necessary for them. After a period, the use times of these blocks will be different greatly from other blocks, un-levelling wear degree between the blocks. To prevent this, the thin pool monitors blocks that are used less frequently and remain unchanged for a long time, and do garbage collection forcibly on these blocks to enable these blocks to be used again, preventing too much differences in use frequency of blocks.

Relationship with other features

Global wear leveling is the basic feature of thin pools and has no conflict with other advanced features.

Benefits

- Make sure that no SSD is damaged too quickly because of excessive writes, making all SSDs in a thin pool wear leveling.

- Decrease the risk of single SSD failures, thus reducing maintenance costs and improving data security.
- Make all SSDs have balanced service loads and maximize their performance, thus improving the overall performance of the system.

Application scenarios

Applicable to general application scenarios of thin pools.

2.3 InVirtualization

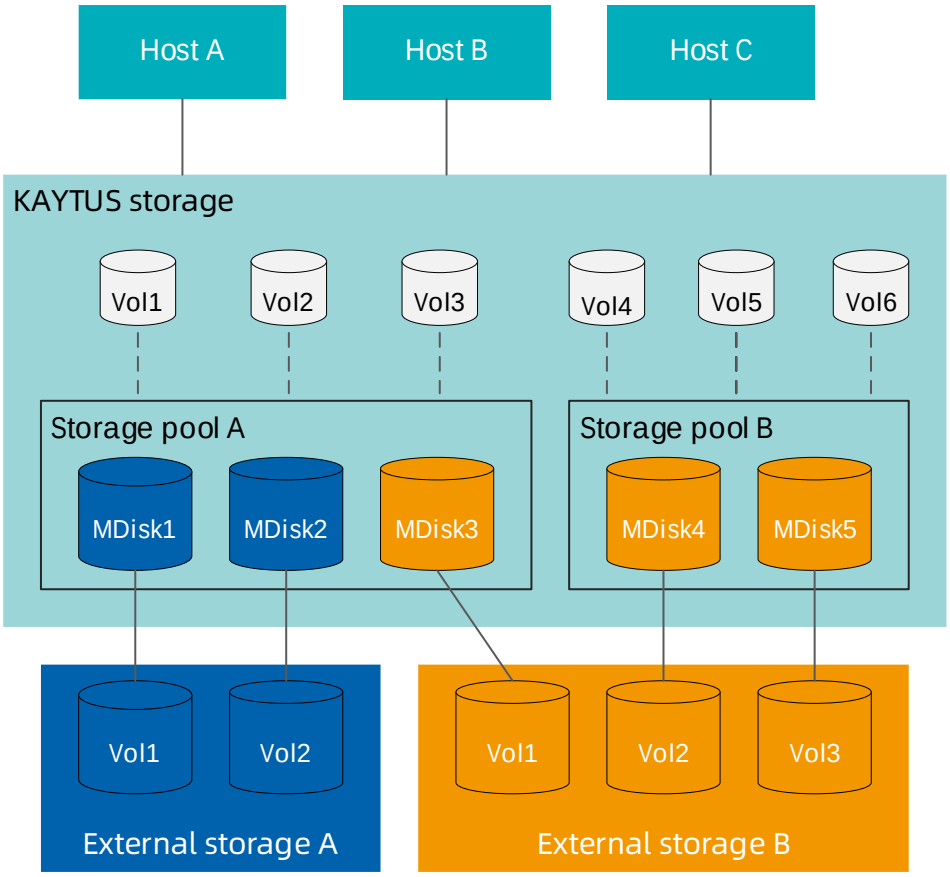
Feature description

By logical abstract, InVirtualization can eliminate the differences between different types of storage resources, and get resource integration and utilization. The integrated storage devices can have the same advanced features as the local storage, and the administrator can manage all storage devices at the same time as one storage device, decreasing management costs.

Feature principle

The virtualization function of the storage is block-level virtualization. The storage subsystem (virtualized external storage system) can separate the RAID array from the storage space and map it to the front-end storage system in the form of volumes through an internal SAN network. The storage system maps the volumes from the storage subsystem to form an MDisk, and then forms a virtual resource pool (referred to as a storage pool) from one or more MDisks. Subsequently, the system allocates volumes out of the storage pool and maps them to the front-end hosts. The block-level heterogeneous virtualization is as shown in Figure 2-10.

Figure 2-10 Block-level heterogeneous virtualization



For the front-end application hosts, only volumes can be identify. After formatted and partitions are created, volumes can be used as local disks. The management of volumes such as expansion, shrinkage, and snapshots are all done in the virtual tier of the storage system, without involving host resources or relying on the back-end virtual storage subsystems. The back-end subsystem can use advanced features of the storage system after its storage space is taken over to the storage pool.

The storage system provides online takeover function and can take over volumes in heterogeneous storage systems and complete online data migration without interruption of host business, making sure of business continuity and data integrity.

Technical specifications

The storage system can connect to the back-end storage subsystem through an FC network, and each clustered system can map a maximum of 4,096 volumes from the back-end storage subsystem.

Relationship with other features

InVirtualization has no conflict with other advanced features.

Benefits

The storage can virtualize and take over mainstream storage arrays in the market, making storage arrays

easy to integrate, manage, and use.

- Eliminate platform differences and enhance the data mobility. Heterogeneous virtualization can effectively solve many problems in heterogeneous storage environments, letting enterprises to select the most applicable vendor and storage array flexibly. It breaks the information island of independent systems, improving the utilization of overall storage arrays and data value.
- Integrate device spaces, simplify storage management, and improve work efficiency. The scattered free spaces of old storage arrays can be centralized for management and use. The data can also be directly mapped to the front-end hosts through the storage system, eliminating differences between heterogeneous arrays. Both block and file services can be managed uniformly.
- Make sure of business continuity effectively through data migration. The virtualized volumes of the storage system can be migrated online, which is completed between the underlying subsystems, without effect on host businesses.
- Improve the resource utilization in data centers and effectively decrease management costs for storage arrays. The old storage devices can use the advanced features of the new storage, such as InCompression, InThin, and InDedupe, improving the utilization of capacity space in the storage pool. The storage management is also simplified, greatly reducing management complexity and costs.
- Improve the system performance. The virtualized volume is striped across multiple arrays and controllers, providing more cache capacity and greatly improving the system performance.

Application scenarios

- Scenarios with multiple heterogeneous devices.

  When the storage environment of the data center is complex, unified space and management are necessary for storage devices in different models and from different manufacturers. Through the virtualization function of the storage system, the spaces can be integrated virtually and different devices can be managed uniformly.

- Scenarios where advanced features are necessary.

  There are many devices in the data center with different purchase time, and some devices have no advanced features. Through the virtualization function of the storage system, the heterogeneous storage devices can be integrated and have the same rich and advanced features.

2.4 InMigration

Feature description

InMigration can achieve data migration within an I/O group, including data migration within the same storage pool and between different storage pools, as well as data migration across I/O groups. During data migration, the storage system will rebuild the storage location information of the data. When data migration is completed, the host application can directly connect to the new data volume.

Feature principle

InMigration is manually done on a volume basis. The data migration has no effect on the front-end business.

Data migration for normal pools

There are three levels of data migration for normal pools.

- Volumes are migrated to the other location in the same storage pool.
- Volumes are migrated to the other storage pool within the same I/O group.
- Volumes are migrated to the other I/O group in the same clustered system.

Data migration within the same I/O group runs within the storage system and is completely transparent to front-end business, making sure of business continuity. When the data migration is completed, the position of the data blocks will be redirected to a new location, without effect on links related to host access. For data migration between different I/O groups, before the migration, it is necessary to make sure that the host can get access to the volumes in both I/O groups at the same time and both I/O groups can get access to the volumes to be migrated at the same time. The migration steps are as shown below.

1. The source-end writes the data in the cache to the drive, and all I/Os that are currently in progress at the host-end are in a write-through mode, directly writing the data to the drive.
1. Add the access permission of the target I/O group to the source volume.
2. Establish a new path so that the host can get access to the migrated target volume.
3. Do volume migration.
4. After the migration, delete the access permission of the source I/O group to the volume.

Data migration for thin pools

Volumes are migrated to the other I/O group in the same clustered system. The data migration between different I/O Groups is done through InRemoteCopy. For more information, see 4.2 InRemoteCopy.

During the data migration, the data from the source volume are copied to the target volume (a volume in a normal pool or a volume in a thin pool). After the synchronization, the target volume fully takes over the business of the source volume, with the volume information such as UID retained.

Relationship with other features

Normal volumes, thin volumes, compressed volumes, deduplicated volumes, and compressed-deduplicated volumes all support data migration.

For the relationship of source volume (master volume) and target volume (auxiliary volume) with data migration in InLocalCopy and InRemoteCopy, see 4.2 InRemoteCopy.

Benefits

- Data migration can change the storage location of volumes and meet the performance requirements of front-end businesses at any time.

- Data migration is completely transparent to the front-end host business, without service interruption.
- Data can be migrated from multiple heterogeneous devices to the same storage device.

Application scenarios

- Balance load and improve IO performance.

  Enterprises have multiple applications, and each application has different performance requirements for storage systems. When businesses with high-performance requirements have been deployed in storage pools with lower performance, host business can be migrated to high-performance storage pools through data migration.

- Migrate data between heterogeneous storage systems.

  There are storage devices from two different vendors and it is necessary to migrate data from storage A to storage B, but the storage devices of the two vendors cannot communicate with each other. In this case, InVirtualization and InMigration can be used to complete data migration.

3 Workload optimization

3.1 InThin

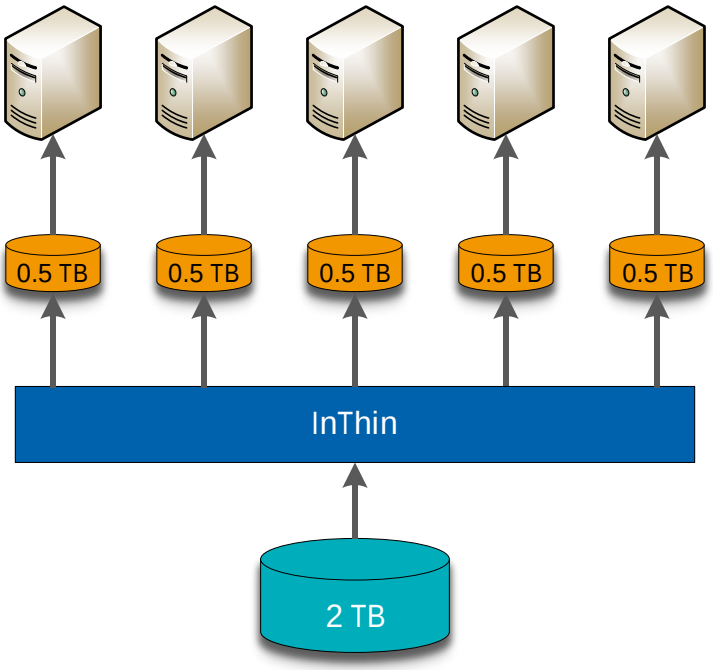
Feature description

InThin is a new capacity allocation technology that does not allocate a large enough capacity to the application host in one time, but allocates capacity to the application host for multiple times and in small amounts per time referring to the real capacity demand of the application host. The data generated by the application host will increase continuously, and the storage system will allocate part of capacity from the back-end storage pool again when the allocated capacity cannot meet the demand.

Feature principle

With InThin, the storage system can supply more storage space to the front-end hosts than is available on the storage system. For example, the available storage capacity for a storage system is 2 TB, but it can map 0.5 TB each to five hosts, with 2.5 TB in total, as shown in Figure 3-1. It is necessary for administrators to monitor the use of real capacity of the storage system and add storage space when the real capacity is not sufficient.

Figure 3-1 Configuration example of InThin



The storage system support thin volumes and fully allocated volumes. Thin volumes are created with real and virtual capacities while fully allocated volumes are created with a one-time allocation of real capacity. Real capacity refers to the physical space allocated to a volume, but for virtual capacity, no physical space will be allocated to a volume. Both normal and thin pools support thin volumes.

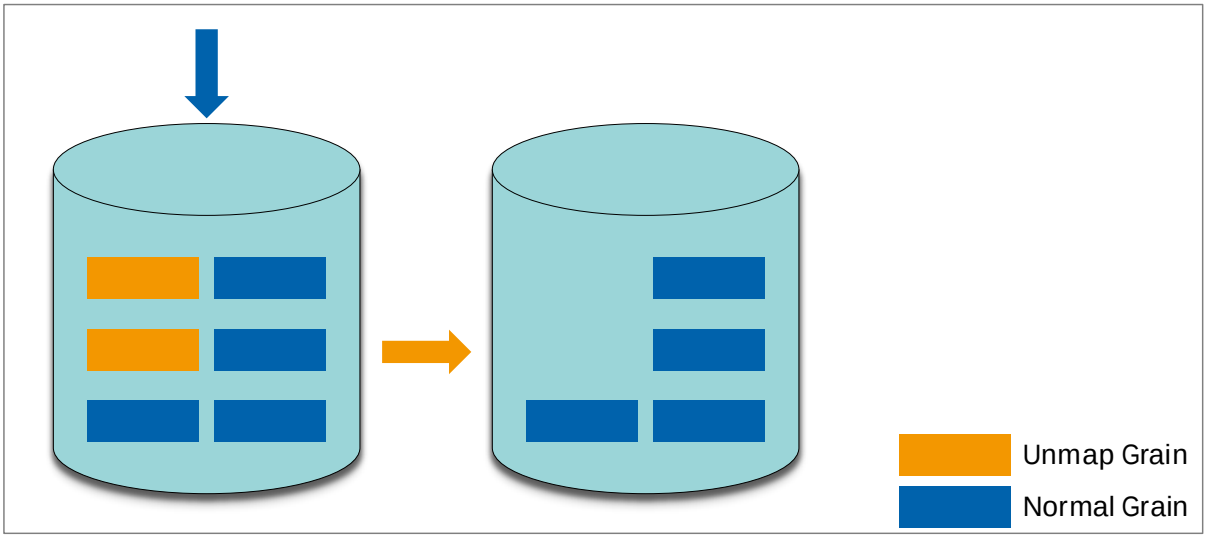
For thin volumes, the initial capacity allocation is as shown below.

- Thin volumes in a normal pool: By default, 2% is allocated as the real capacity and 98% as the virtual capacity. A small portion of the real capacity is used to store initial metadata, and the real capacity can also be set to 0%.
- Thin volumes in a thin pool: The default real capacity allocated is 0%, which means that the capacity allocated to the application host is virtual capacity and no physical space is allocated. In a thin pool, the metadata of thin volumes is stored independently.

During write operation, the data will be written to the thin volume referring to the real capacity with a constant size (grain size).

- Thin volumes in a normal pool: The grain size can be set to 32 KiB, 64 KiB, 128 KiB, and 256 KiB when a thin volume is created in a normal pool. The maximum virtual capacity of thin volumes is depend on grain size. Smaller grain size can save space but cause the index table to be larger, while larger grain size can get better performance. The effect on performance is different in different application scenarios. It is recommended that the grain size is set reasonably based on actual scenarios. When it is necessary to take a snapshot for a thin volume, make sure that the grain size of both the thin volume and the snapshot are the same.
- Thin volumes in a thin pool: The grain size can be set to 8 KiB and 32 KiB when a thin volume is created in a thin pool. Space reclamation is supported, which can identify deleted data and do garbage collection immediately to decrease data migration and thus improve SSD service life and data reduction rate.

Figure 3-2 Unmap example of InThin



The space allocation methods for normal and thin pools are different.

- In a normal pool, thin volumes support both automatic and non-automatic expansion. When auto-

expansion is enabled for a thin volume, the storage system will automatically increase the real capacity of constant sizes to it. Thin volumes with auto-expansion enabled will reserve a certain amount of real free capacity to prevent unexpected errors in case that the amount of data stored by users does not match the real capacity of the thin volume. If auto-expansion is disabled, thin volumes will not have real free capacity. When the actual necessary capacity exceeds the configured capacity of a volume, the volume will go offline immediately.

- In a thin pool, metadata are stored independently and do not occupy the real capacity of a thin volume. Thin pools manage space through unified allocation, it is not necessary to allocate any space to volumes in advance. Volumes in a thin pool can enable compression feature and deduplication feature, making the stored data can be much more than the real capacity through data reduction.

Alarm threshold can be set for a thin volume to prevent it from going offline because of capacity depletion. If the capacity necessary to user exceeds the threshold, the storage system will generate alarm events and send alarm notifications through email or SNMP. This makes sure that capacity can be expanded immediately and business continue to run correctly without interruption.

Technical specifications

The technical specifications of thin volumes are as shown in Table 3-1.

Table 3-1 Technical specifications of thin volumes

Attribute	Value
Maximum number of thin volumes per system	KS2000G7: 2048 KS3000G7: 4096 - (number of thin pools*6)
Maximum capacity of a thin volume	512 TB
Minimum capacity of a thin volume	512 B

The size of extents has a direct effect on the maximum capacity of a thin volume in a normal pool, and their relationship is as shown in Table 3-2. The maximum capacity of a thin volume in a thin pool is independent of extent, and the storage space in a thin pool is managed in a unified method.

Table 3-2 Relationship between extent and capacity of a thin volume

Extent size (MB)	Maximum virtual capacity of a thin volume (GB)
128	16000
256	32000
512	65000
1024	130000
2048	260000

4096	520000
8192	524288

The real capacity of a thin volume in a normal pool is increased by constant grain, and the grain size has a direct effect on its maximum virtual capacity. Their relationship is as shown in Table 3-3. The grain size of a thin volume in a thin pool can be 8 KiB and 32 KiB, and its maximum virtual capacity is 512 TB.

Table 3-3 Relationship between grain and virtual capacity of a thin volume

Grain size (KiB)	Maximum virtual capacity of a thin volume (GB)
32	260000
64	520000
128	524288
256	524288

Relationship with other features

- Thin volumes can be applied in the scenarios of InLocalCopy, InRemoteCopy, and InMetro.
- Thin volumes in a normal pool can be changed to normal volumes in a normal pool, or thin volumes in a thin pool.
- Thin volumes in a thin pool can be changed to normal volumes and thin volumes in a normal pool.

Benefits

- InThin can eliminate almost all free spaces and is compatible with the conversion from random write to sequential write in thin pools, effectively improving the utilization of available storage capacity.
- Thin volumes in a thin pool support unmap space reclamation, which can clear the data that have been deleted by users timely. This can make sure of the efficiency and availability of storage spaces, decrease the data migration at both system-level and SSD-level, and extend the service life of SSDs.
- InThin can decrease the investment cost for storage devices in the early stage effectively, and more drives are necessary to purchase only when storage capacity is not sufficient.
- In a thin pool, compression feature and deduplication feature can be enabled for thin volumes in real-time, achieving data reduction and improving capacity savings.

Application scenarios

- Scenarios that cloud storage spaces should be allocated on demand.

  Cloud storage users purchase a large cloud storage space usually, but only use a small portion of it in the short term. It will waste a large amount of storage space and increase enterprise costs if cloud

storage service providers allocates all storage space in one time. In this case, thin volumes can be used to allocate spaces referring to the actual demand, improving the utilization of storage space, and effectively reducing the cost of cloud storage service providers for storage device purchase, management, and maintenance.

- Scenarios that storage space that should be used reasonably.

  For environments with multiple applications, the initial spatial planning is often not precise. InThin enables the allocation of sufficient virtual space to applications, and then allocates physical spaces referring to the real capacity requirements of applications, thus effectively improving space utilization and reducing the risks caused by unreasonable early planning.

3.2 InCompression

Feature description

InCompression is a volume-based compression technology that enables real-time compression of data before the data are written to drives. InCompression is also a hardware-based compression technology, which provides compression function through the QAT compression engine, reducing compression costs and improving the throughput and compression performance of compression devices.

Feature principle

The storage system uses real-time data compression technology, with constant length input (8K/32K) and variable length output. After compression, the compressed data will be stored on drives with constant length of 8 bytes, effectively improving the utilization of drive space. The compression solution uses hardware compression as the main method and software decompression as the auxiliary method. In case of a failure in hardware compression, software decompression is used to make that the compressed data are readable.

The advantages and features of InCompression are as shown below.

- Pre-decide mechanism

  For chunks that already have a higher compression ratio, they can still be compressed as other chunks. It saves very little space, but still consumes resources such as CPU and cache resources. To prevent resource waste on uncompressible data, the storage system provides a more efficient compression algorithm (pre-decide mechanism). The chunks that are below a given compression ratio will be skipped by the compression engine, thus saving CPU time and cache processing. The chunks that are decided not to be compressed with the compression algorithm, but that still can be compressed, are marked and processed accordingly. The result might vary because pre-decide mechanism does not check the full block, only a sample of it.

- Time-based compression

  RACE provides a time-based compression technology. RACE can do duplicate data retrieval and

compression when data are written to the same compressed block at the same time, bring a higher compression ratio and a reduction in the number of retrieval tables. When the data from host-end are written to RACE, they are compressed and fill up compressed blocks with constant size. Multiple compressed writes can be aggregated into a single compressed block, and the retrieval table is stored within the same compressed block. Because these writes are likely from the same application and with the same data type, the compression algorithm usually detects more duplicate data.

- Transparent compression process

  The compression engine is integrated into the storage system. It does not change the attributes of system data, but only changes the storage of data, and is compatible with different attributes of compressed volumes. When data are written to the storage system, the compression process is transparent and has no effect the host. Based on the compression engine, real-time compression technology uses lossless data compression algorithms, which can dynamically compress data to meet the performance, reliability, and scalability requirements of enterprises.

- Higher-performance compression chips

  The QAT hardware compression function uses higher-performance chips to improve the performance of processing I/O requests such as throughput, IOPS, and latency. With the optimization of compression algorithms, it improves the compression performance.

Technical specifications

The technical specifications of compressed volumes are as shown in Table 3-4.

Table 3-4 Technical specifications of compressed volumes

Attribute	Value
Maximum number of compressed volumes per system	4096 - (number of thin pools*6)
Maximum capacity of a compressed volume	512 TB
Minimum capacity of a compressed volume	512 B
Whether can be enabled or disabled in real-time	Yes

Relationship with other features

- Compressed volumes can be applied in scenarios such as InLocalCopy, InRemoteCopy, and InMetro.
- Deduplication feature can be enabled for thin volumes with compression feature enabled.

Benefits

InCompression can significantly improves and optimizes the compression ratio, number of retrieval tables, and read-write performance, which can save more than twice the storage space, improve the business performance, and effectively decrease the TCO.

- Improve I/O performance.

  Real-time compression technology can compress active primary data and improve I/O performance when data are written to drives.

- Save storage spaces.

  Data compression has been completed when the host writes data to the storage system, thus using less storage space with the same data.

- Decrease the number of storage devices.

  With InCompression, the same storage space can store more data, thus reducing the number of storage devices and the TCO for space, purchase, cooling, management, and maintenance.

- Decrease the wear of SSDs.

  InCompression can decrease the amount of data written to SSDs, effectively reducing the read-write operations to SSDs while making ensure of data security. This can decrease the wear of SSDs, extend the service life of SSDs, and further decrease the TCO.

Application scenarios

- Universal compressed volumes

  Compressed volumes can store multiple types of data, such as directory index data, design data, oil and gas geological data, and seismic data, all of which can be highly compressed. Based on practice, the compression ratio for multi-type data is 50% to 60%, effectively reducing the use of space and providing more space to other data.

- Database application

  The data type of database is mainly tablespace files, and the compression ratio of tablespace files is very high. Taking DB2, Oracle, and Microsoft SQL Server as examples, the expected compression ratio is 50% to 80%.

- Virtualization application

  Virtualization applications have been widely applied in businesses and its market is constantly expanding. More storage spaces are necessary to support virtualization applications to store more virtual server image data and backup data. InCompression supports mainstream virtualization applications such as VMware, Hyper-V, and kernel-based virtual machines, of which the expected compression ratio is 45% to 75%, effectively reducing the demand for storage space.

- Log server

  Log files are crucial data for any company or department. Thus, it is crucial for administrators to log in to the log server of the system quickly to find logs. The effect of compression on log files is obvious because of the unique structure of log files, with an expected compression ratio of a maximum of 90%.

3.3 InDedupe

Feature description

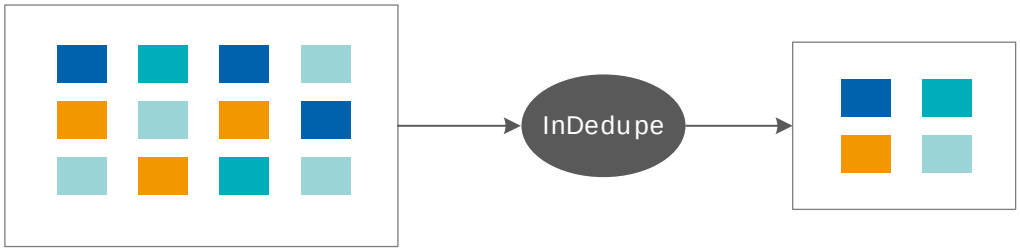
InDedupe is a pool-based duplicate data detection and reduction technology, implemented through software, without any hardware. Before data are written to SSDs, real-time detection and deduplication will be done, which can decrease the number of write and the amount of data written to SSDs. InDedupe can save storage spaces, decrease wear on SSDs, and extend the service life of SSDs.

Feature principle

InDedupe identify duplicate data by the fingerprint value of the data. When the fingerprint values of new data are different from those of the existing data, they are identified and stored as non-duplicate data to drives. When the fingerprint values of new data are the same as those of the existing data, it is necessary to compare the written data byte by byte with the data saved on drives to make sure of the correctness and consistency of the data.

The principle of InDedupe is as shown in Figure 3-3.

Figure 3-3 Principle of InDedupe



InDedupe supports all-zero data recognition. The fingerprint value of all-zero data is constant. When the fingerprint value of the written data is identified as fingerprint value of all-zero data, it is necessary to do a check of the written data byte by byte to see if they are all-zero data to make sure of data security. When they are all-zero data, LBA will be mapped to PBA of all-zero data that does not occupy real storage space, thus saving storage space and making sure of accuracy of user data statistics. When LBA is mapped to PBA of all-zero data, the returned data of read operations is zero.

To meet different application scenarios and get performance balance, InDedupe can be enabled and disabled in real-time. For thin volumes with the deduplication feature enabled, duplicate data detection and deletion are done within the thin pool during data write process. After the deduplication feature is disabled, the data that has been deduplicated can still be read and written, but the new written data will be stored directly without duplicate data detection and deletion. For thin volumes with the deduplication feature disabled, duplicate data detection and deletion are not done during data write process.

To meet the high performance requirements of production environments, the storage system also supports background deduplication technology. The front-end application data will be stored to SSDs without deduplication, and then the system will complete duplicate data detection and reduction in background when its workload is low. Background deduplication can effectively shorten the I/O path and

improve the performance of the front-end system. Because background deduplication will use system resources such as CPU and memory, the storage system can control the concurrency of background deduplication tasks through sensing system workload (such as I/O concurrency), thus making sure that background deduplication tasks be completed successfully while the unwanted effect on business performance be minimized.

Technical specifications

The technical specifications of deduplicated volumes are as shown in Table 3-5.

Table 3-5 Technical specifications of deduplicated volumes

Attribute	Value
Maximum number of deduplicated volumes per system	4096 - (number of thin pools*6)
Maximum capacity of a deduplicated volume	512 TB
Minimum capacity of a deduplicated volume	512 B
Whether can be enabled or disabled in real-time	Yes

Note: The size of deduplication data blocks is the same as the grain size.

Relationship with other features

- Deduplicated volumes can be applied in scenarios such as InLocalCopy, InRemoteCopy, and InMetro.
- Mirrors can be created for deduplicated volumes for data migration.
- Compression feature can be enabled for thin volumes with deduplication feature enabled.

Benefits

InDedupe optimizes technologies such as duplicate data detection, data fingerprint value calculation, and all-zero data detection, which can effectively improve business performance and decrease the TCO.

- Save storage spaces.

  When the host writes data to the storage devices, duplicate data will not be written to SSDs. Thus, the storage space occupied will be smaller than the capacity of the written data.

- Decrease the number of storage devices.

  With InDedupe, the same storage space can store more data, thus reducing the number of storage devices and the TCO for space, purchase, cooling, management, and maintenance.

- Decrease the wear of SSDs.

  InDedupe can prevent repeated write operations caused by duplicate data, effectively reducing the read-write operations to SSDs while making ensure of data security. This can decrease the wear of SSDs, extend the service life of SSDs, and further decrease the TCO.

Application scenarios

Whether to enable InDedupe or not is based on the actual business scenario. In business scenario where the data duplicate ratio is high and the data reduction is obvious after InDedupe is enabled, it is recommended that InDedupe should be enabled. In business scenario where the data duplicate ratio is low, the data reduction is not obvious after InDedupe is enabled, and the requirement for performance is high, it is recommended that InDedupe should be disabled.

- Database application

  In typical database scenarios, the duplicate data ratio is usually between 2:1 and 4:1. The data reduction rate will not be very high, thus it is recommended that InDedupe should be disabled.

- Virtualization application

  The duplicate data ratio in VSI scenarios is usually between 5:1 and 9:1, and the duplicate data ratio in VDI scenarios is usually above 10:1 and even as high as 20:1. In virtualization scenarios, the data reduction rate will be very high, thus it is recommended that InDedupe should be enabled.

3.4 InAutoPartition

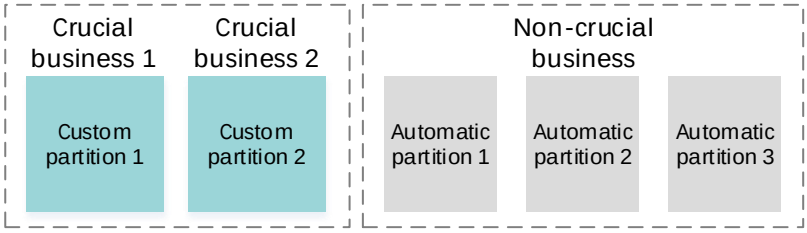
Feature description

InAutoPartition can make sure of the performance of crucial businesses by allocating private cache resources for crucial businesses. At the same time, it improves the utilization of cache resources by providing a cache resource sharing mechanism for non-crucial businesses.

Feature principle

Cache partitions are achieved on a pool basis, and only supports parent pools but not child pools. For crucial businesses, the capacity of cache partitions can be customized, which are reserved by the system and will not be occupied by other cache partitions. For non-crucial businesses, the cache partitions can be set to an automatic mode, and the real capacity of each partition will vary with changes in the business load. There are limits for the total amount of cache resources that can be used by all automatic cache partitions.

Figure 3-4 Principle of InAutoPartition



Technical specifications

The technical specifications of InAutoPartition are as shown in Table 3-6.

Table 3-6 Technical specifications of InAutoPartition

Attribute	Value
Maximum number of cache partitions per system	128
Total capacity of all custom cache partitions per I/O group	No more than 100%

Relationship with other features

InAutoPartition has no conflict with other advanced features.

Benefits

- Isolate cache resources.

  It achieves the independence of cache resources, prevents mutual interference between different business applications, and makes sure of the product quality.

- Make sure of the performance of crucial businesses.

  The cache resources of crucial businesses cannot be occupied by non-crucial businesses, making sure of the performance and reliability of crucial businesses.

- Improve the cache utilization of non-crucial businesses.

  When cache resources are relatively abundant, business applications are permitted to occupy as much cache as possible to improve the performance. When cache resources are scarce, they will be dynamically balanced, making performance of non-crucial businesses be balanced.

- Restrict the resources use of slow drive.

  Prevent the possibility that individual partitions occupy excessive cache resources through restricting the upper limit of cache resources occupied by slow drives, make sure of the overall service quality of the system.

Application scenarios

InAutoPartition can be applied in the performance grading scenarios of member and non-member Internet services. It allocates the business of member and non-member users into different cache partitions to meet the requirements of different users.

- Allocate sufficient private cache resources to the partitions where member users are located to make sure of the performance, service quality, and reliability of member users.
- Non-member users have low performance requirements and significant fluctuations in business load. In the partitions where non-member users are located, cache resources can be utilized to a certain extent to improve the performance when the business load is low. When the business load is high, cache resources will be balanced to balance the performance.

3.5 InQoS

Feature description

InQoS allocates resources of the storage system reasonably through flow control to meet the specific performance requirements of certain applications, which can make sure of the quality of service for crucial businesses.

Feature principle

Flow control refers to a management program based on storage-end I/O queues, which can distribute storage resources. The storage system supports flow control from the front-end to the storage-end and end-to-end flow limit within the storage system. The flow control function can be enabled or disabled online referring to actual requirements and the maximum target or minimum target can be set for business objects based on the business characteristics such as IOPS, bandwidth, and response time. After the flow control function is enabled, the storage system will decrease the process speed of some low-priority businesses, thus decreasing the occupation of system resources by low-priority businesses and making sure of the performance of high-priority businesses.

InQoS can make sure of the service quality of data business through two methods, including maximum traffic control and minimum performance assurance. The controlled objects and configuration items are as shown in Table 3-7.

- Maximum traffic control: The traffic control is done through token allocation and I/O queue management for controlled objects, which makes the storage system allocate resources dynamically.
- Burst traffic control: The system collects the unused resources during off-peak hours and uses them during traffic bursts to break the upper limit for a short period. To get this, the long-term average traffic of the business should be below the upper limit.
- Minimum performance assurance: Through the traffic control of all business objects, the system releases sufficient resources and provides them to objects that have not get the lower limit target, making sure of the performance of these objects as much as possible.

Table 3-7 Controlled objects and configuration items

Feature	Controlled objects	Configuration items
Maximum traffic control (including burst traffic control)	Volume, volume group, host	IOPS and bandwidth
Minimum performance assurance	Volume, volume group, host	IOPS, bandwidth, and maximum response time

Technical specifications

The technical specifications of InQoS are as shown in Table 3-8.

Table 3-8 Technical specifications of InQoS

Attribute	Value
IOPS upper limit	50 to 33554432
IOPS lower limit	50 to 33554176
IOPS burst limit	51 to 33554688
Bandwidth upper limit	0 MBps to 131072 MBps
Bandwidth lower limit	0 MBps to 131071 MBps
Bandwidth burst limit	2 MBps to 131073 MBps
Maximum burst time	0 s to 600000 s
Maximum response time	0 us to 1000000 us

Note: Each attribute can be set to 0, which is the default value and indicates that there is no limit.

Relationship with other features

InQoS has no conflict with other advanced features.

Benefits

- Do flow limit on low-priority businesses, reducing the system resource utilization and making sure of the performance of crucial (high-priority) businesses.
- Do overload control for crucial businesses to make sure of high stability and reliability of the storage system.
- Restrict the traffic of end-to-end businesses within the storage system to make sure of the business performance from front-end to storage-end.
- Provide emergency capabilities for time sensitive businesses to make sure of response time during peak hours.
- When non-critical businesses overload, protect critical businesses actively to make sure of their performance.

Application scenarios

InQoS can applied in different scenarios and make sure of the business performance of crucial businesses and high-level users.

- Maximum traffic control and minimum performance assurance are applicable to scenarios where multiple businesses are deployed. Maximum traffic control can make sure of the performance of crucial businesses through the limit of resources that allocated to non-critical businesses. Minimum performance assurance can make sure of the performance of controlled objects that have been

configured with this policy.

- As an enhancement of maximum traffic control, burst traffic control is applicable to business scenarios that are highly sensitive to latency, letting them break the upper limit for a short period.

4 Data protection

4.1 InLocalCopy

Feature description

InLocalCopy is a copy function implemented within a clustered system, also known as point-in-time copy. On the principle of copy-on-write (COW), InLocalCopy enables the creation of an instantly available consistent copy while the data is online and available.

InLocalCopy is done at the block level, supporting multiple target volumes and cascaded copy. Multiple target volumes can be created for a source volume and copies can be created for a copy. In different local copy mappings, a volume can be both a source volume and a target volume. Without effect on the initial local copy mapping, a copy of the target volume can be create. A consistency group can contain multiple volumes, enabling operations for multiple volumes such as creating copies and deleting copies at the same time. A consistency group can make sure of the point-in-time consistency of multiple volume and data integrity of copies.

For different applications, InLocalCopy provides three type of copies, including snapshot, clone, and incremental backup. To prevent accidental or intentional deletion of snapshot data, security snapshots can be created through setting security period, which cannot be deleted within the security period. After the end of security period, snapshots can be manually deleted or automatically deleted by the system.

Feature principle

Before a local copy mapping is created, the controller cache is in write-through mode by default. Because the local copy module is below the upper cache, do a check of the controller cache mode to make sure that the data view on the host-end is the same as that of the local copy module. In different status of a local copy mapping, the related status of the source volume and the target volume are as shown in Table 4-1.

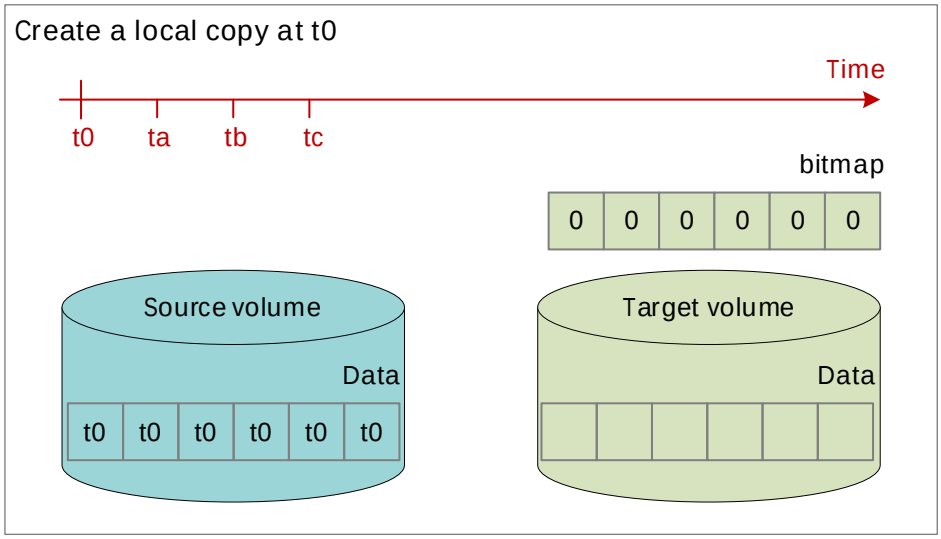
Table 4-1 Status of source volumes and target volumes related to status of local copy mappings

Status of local copy mappings	Source volumes		Target volumes	
	Online/Offline	Cache status	Online/Offline	Cache status
Idle or Copied	Online	Write-back	Online	Write-back
Copying	Online	Write-back	Online	Write-back

Stopped	Online	Write-back	Offline	N/A
Stopping	Online	Write-back	Online: Back-end copying completed Offline: Back-end copying not completed	N/A
Suspended	Offline	Write-back	Offline	N/A
Preparing	Online	Write-through	Online, but not accessible	N/A
Prepared	Online	Write-through	Online, but not accessible	N/A

After the local copy of the specified volume is enabled, a bitmap will be created, and there will be an available copy relationship between the source volume and the target volume. Each target volume has a bitmap that records the progress as data are copied from the source volume to the target volume, that is, which data have been or have not been copied to the target volume. After the active data blocks are copied from the source volume to the target volume, the related values in the bitmap will also be updated synchronously. 0 indicates that the data have not been copied from the source volume to the target volume, and 1 indicates that the data have been copied from the source volume to the target volume.

Figure 4-1 Creation of a local copy mapping



As shown in Figure 4-1, the establishment time of the local copy mapping is t0, and the values in the created bitmap are all 0, indicating that the data have not been copied to the target volume. When the host reads or writes data, the storage system will determine how to do the read-write operations referring to the status of the bitmap, as described below.

- Read data from the source volume.

  The storage system directly reads data from the source volume and return them to the host-end.

- Write data to the source volume.

  For COW snapshots, make sure that the initial data have been copied to the target volume before a

write operation is done, the storage system will first check the previous bitmap relationship. If the related value in the bitmap is 1, the write operation will be done directly. If the related value in the bitmap is 0, the data will be copied to the target volume first, and then the write operation will be done. As shown in Figure 4-2, before the write operations (write a, write b, and write c) write data to the source volume, the data at the point-in-time t0 of the source volume will firstly be copied to the target volume. For ROW snapshots, the host I/O will be redirect-written to the other location, and the target volume will manage the data from the initial location, as shown in Figure 4-3.

Figure 4-2 Principle of writing to source volumes (COW)

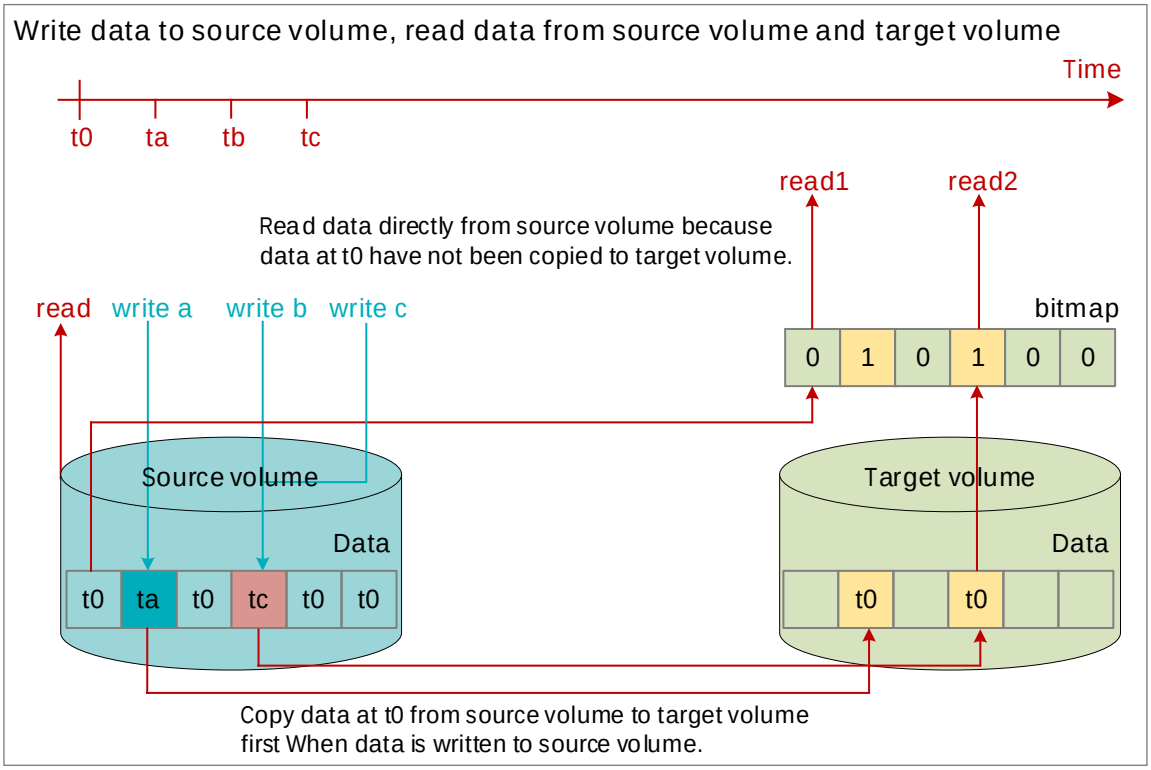
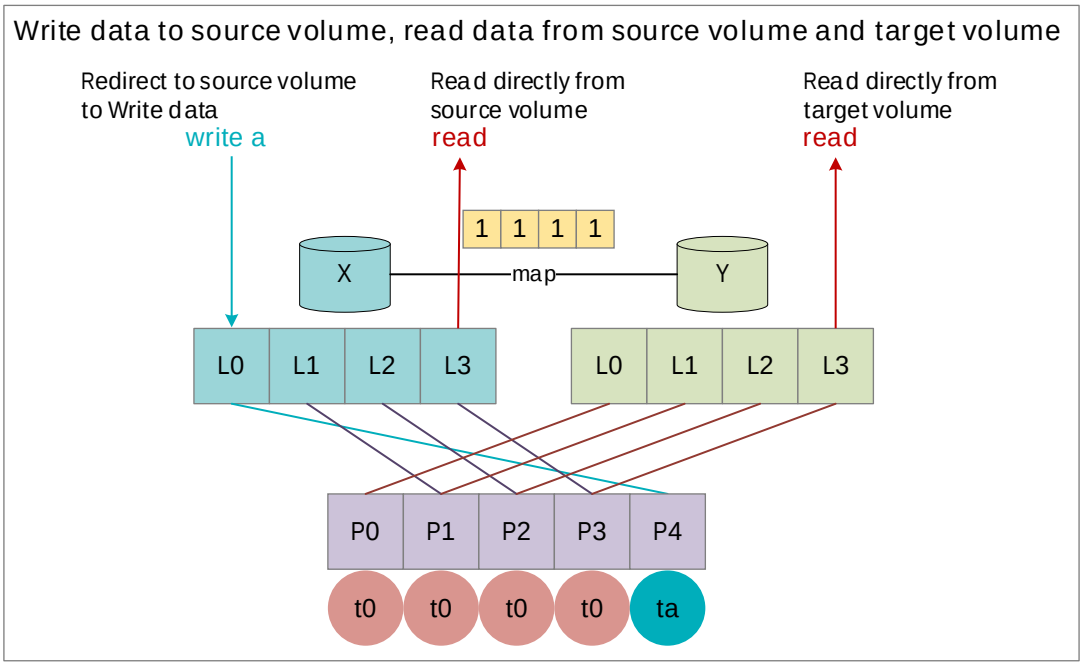


Figure 4-3 Principle of writing to source volumes (ROW)



- Read data from the target volume.

  For COW snapshots, the storage system will do a check of the previous bitmap relationship before reading the data. If the related value in the bitmap is 1, the data will be directly read from the target volume. If the related value in the bitmap is 0, the data will be directly read from the source volume. As shown in Figure 4-2, for the read operation named read 1, the related value in the bitmap is the 0, indicating that the data have not been copied to the target volume, and thus the data will be read from the source volume. For the read operation named read 2, the related value in the bitmap has been changed to 1, indicating that the data have been copied to the target volume, and thus the data will be directly read from the target volume. For ROW snapshots, the data will be read directly from the target volume.

- Write data to the target volume.

  For COW snapshots, it is necessary for the storage system to update the bitmap correctly. To prevent the write operation to the target volume from being overwritten, if a data block has not yet been copied, the related value in the bitmap should be updated to 1. As shown in Figure 4-4, for the write operation named write x, the related value in the bitmap is 1, indicating that the data have been copied to the target volume, and the write operation will be done directly. For the write operation named write y, the related value in the bitmap is 0, indicating that the data have not been copied to the target volume, and thus the related value in the bitmap will be updated to 1 first before the write operation. For ROW snapshots, the data will be redirect-written to the target volume, as shown in Figure 4-5.

Figure 4-4 Principle of writing to target volumes (COW)

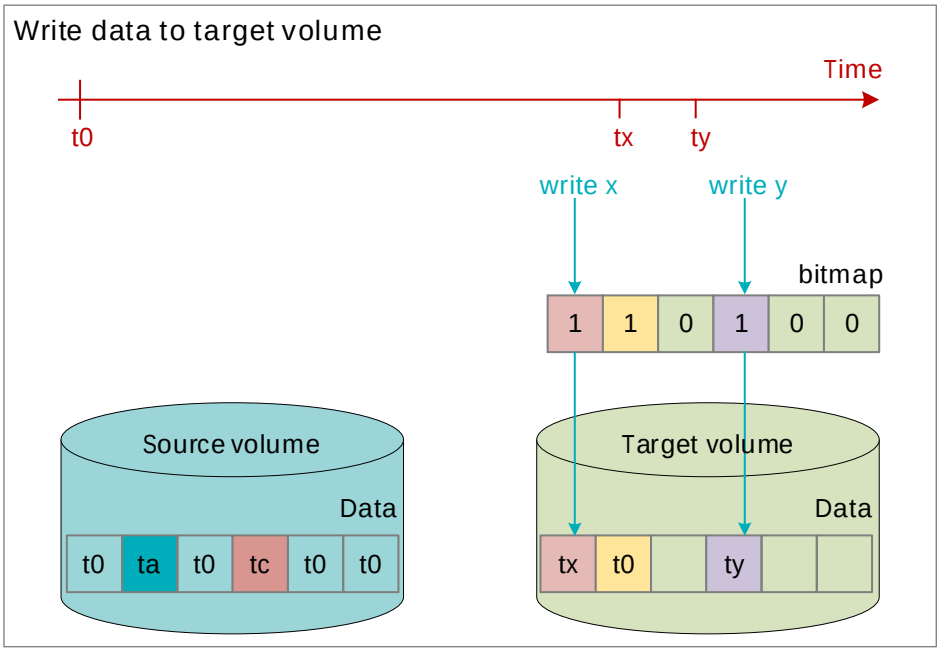
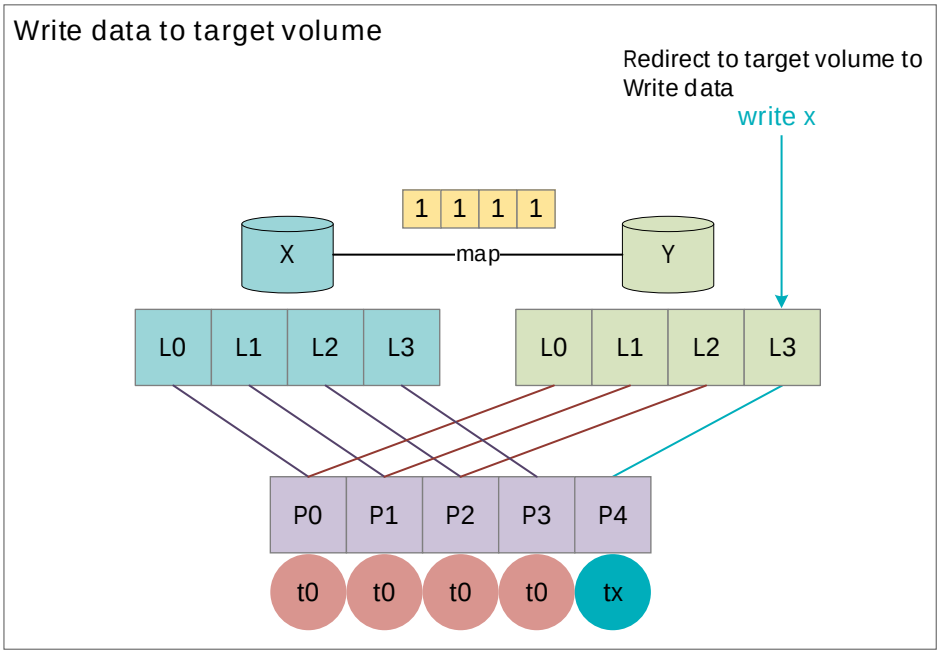


Figure 4-5 Principle of writing to target volumes (ROW)



A snapshot can be changed to a clone, and the copy process priority of clone and incremental backup can be changed, thus changing the copy rate.

- Snapshot

  A snapshot refers to a point-in-time mapping view of the source volume, and it is not intended to be an independent copy, but is dependent on the source volume. After a snapshot is created, the storage system will create a thin volume as target volume automatically, without real capacity. Before the data change, the data space occupied by the target volume is almost zero. Capacity will be allocated to the target volume only after data have been written to the source volume. When a

snapshot is created, a permanent local copy mapping will be established between the snapshot and the source volume data. Snapshot is usually used for data retrieval.

- Clone

  Clone is a full copy of the source volume. After the copy operation is completed, the clone is independent from the source volume and its local copy mapping with the source volume will be broken. Thus, operations on a clone will have no effect on the source volume. After a clone is created, the data will be copied in the background at specified rate, creating a target volume with exactly the same as the source volume. When the copy progress reaches 100%, the local copy mapping between the clone and the source volume will be automatically deleted. Clone is usually used for scenarios where it is necessary to test or change the data, without effect on the data of the source volume.

- Incremental backup

  The background copy rate of an incremental backup is bigger than 0 and the local copy mappings will not be deleted after the copy is completed.
  Incremental backup refers to creating a full copy for the first time of the source volume and then creating incremental copy based on data changes, completing the copy from the source volume to the target volume with minimum data copy. After the copy operation is completed, the local copy mapping between the incremental backup and the source volume will be updated and be always kept. The attributes of the target volume created by incremental backup are also the same as that of the source volume, and incremental copy will be done at the specified rate. Incremental backup is usually used for the data backup and data recovery.

Technical specifications

When the background copy is enabled, the data in the source volume will be copied to the target volume. The target volume becomes a clone volume that is independent from the source volume after the copy is complete. The background copy rate is an attribute of a local copy mapping that can be set as a value from 0 to 150 and the value can be changed dynamically. The relationship between background copy rate and the set value is as shown in Table 4-2.

Table 4-2 Relationship between background copy rate and the set value

Value	Background copy rate (second)
0	0 (disabled)
1 to 10	1 MB
11 to 20	2 MB
21 to 30	4 MB
31 to 40	8 MB
41 to 50	16 MB
51 to 60	32 MB

61 to 70	64 MB
71 to 80	128 MB
81 to 90	256 MB
91 to 100	512 MB
101 to 110	1 GB
111 to 120	2 GB

The technical specifications of InLocalCopy are as shown in Table 4-3.

Table 4-3 Technical specifications of InLocalCopy

Attribute	Value
Maximum number of local copy mappings per system	KS2000G7: 2048 KS3000G7: 4096
Maximum number of consistency groups in per system	499
Maximum number of target volumes per source volume	256
Maximum number of local copy mappings per consistency group	KS2000G7: 2048 KS3000G7: 3072
Total capacity of target volumes per I/O group	8192 TB

Relationship with other features

For InLocalCopy, there are no limits on the types of the source volume and the target volume, which can be any combination of normal volumes, thin volumes, compressed volumes, deduplicated volumes, and compressed-deduplicated volumes. The capacity of the source volume and the target volume must be the same.

A volume can be used in local copy and remote copy at the same time, for information about the interactivity, see 4.2 InRemoteCopy. The support for related operations on the source volume and the target volume for local copy is as shown in Table 4-4.

Table 4-4 Support for volume operations (InLocalCopy)

Operation	Whether supported
Volume in image mode	Supported.
Extent migration within a storage pool	Supported, without effect on host I/O.
Volume migration between storage pools	Supported, without effect on host I/O.

Volume migration between I/O groups	Supported with limits. The local copy mapping cannot be in preparing state, and the bitmap will remain in the initial I/O group.
Capacity expansion	Supported.
Capacity shrinkage	Not supported.
Change of the preferred node within a I/O Group	Supported (only attributed volume).

Benefits

- InLocalCopy supports multi-target local copy mappings, permitting to create multiple target volumes for a source volume, enabling data analysis and test without effect on the source data.
- InLocalCopy supports cascaded local copy mappings, permitting to create copies of volume copies, and a volume can be both a source volume and a target volume in different local copy mappings. The cascaded local copy mappings have no effect on the initial local copy mappings, making sure of data consistency between multiple volumes.

Application scenarios

InLocalCopy is completed at the block level and operates below the host operating system and its cache. Thus, it is not apparent to the host. InLocalCopy can meet the requirements of crucial business, such as creating a copy of a volume while the volume is online and in use and making sure of data consistency. The mainly applicable scenarios are as shown below.

- Quickly create consistent copies of dynamically changing data.
- Quickly create consistent copies of production data for data movement or migration between hosts.
- Quickly create copies of production datasets for application development and test.
- Quickly create copies of production datasets for data auditing and data mining.
- Quickly create copies of production datasets for quality assurance.

4.2 InRemoteCopy

Feature description

InRemoteCopy can create a copy mapping between two volumes of the same capacity, making remote backup of data, which can used for data protection or disaster recovery. Two volumes may be in the same I/O group or different I/O groups of the same clustered system, or in I/O groups of different clustered systems.

For InRemoteCopy, there are two volumes respectively as master volume and auxiliary volume in a remote copy relationship. The master volume is located in the production center that is the master-end, and the auxiliary volume is located in the disaster recovery center that is the auxiliary-end. The master volume is

used to store the production data written by the host-end and provides the access to the business application data. The auxiliary volume is used to store the backup data of the master volume and updates synchronously with the master volume for disaster recovery.

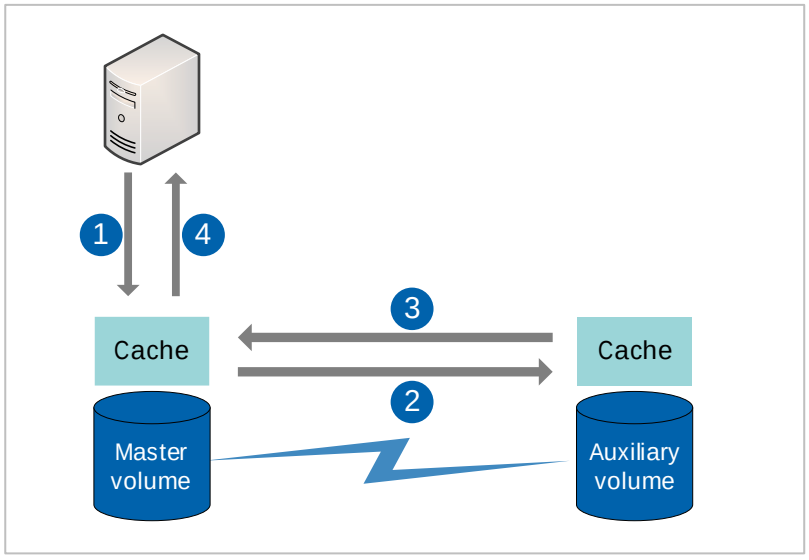
InRemoteCopy provides three copy modes for different disaster recovery scenarios, including synchronous remote copy, asynchronous remote copy, and periodic asynchronous remote copy. In synchronous remote copy, data can be written to both the master volume and the auxiliary volume at the same time, making sure of real-time consistency of data between the master volume and the auxiliary volume. In asynchronous remote copy, there will be certain latency for the data to be written to the auxiliary volume, and the data in the master volume and the auxiliary volume cannot be synchronized in real-time, with a risk of data loss. However, asynchronous remote copy can transmit data over longer distance and requires less bandwidth.

Feature principle

In synchronous remote copy, only after the business data are written to the master volume and the related data are synchronized to the auxiliary volume, the host-end receives a confirmation signal for the completion of the write operation and proceed with the next write operation. After the first copy operation is completed, the data of the master volume and the auxiliary volume are synchronized and updated in real-time, making sure of data consistency between the master volume and the auxiliary volume. The process of synchronous remote copy of a write operation is as shown in Figure 4-6.

1. The host-end completes a write operation, writing the data into the cache of the master volume.
2. The data in the cache of the master volume are synchronously copied to the cache of the auxiliary volume.
3. The auxiliary volume sends feedback that data copy operation has been completed to the master volume.
4. The master volume sends feedback that the write operation has been completed to the host-end.

Figure 4-6 Synchronous remote copy

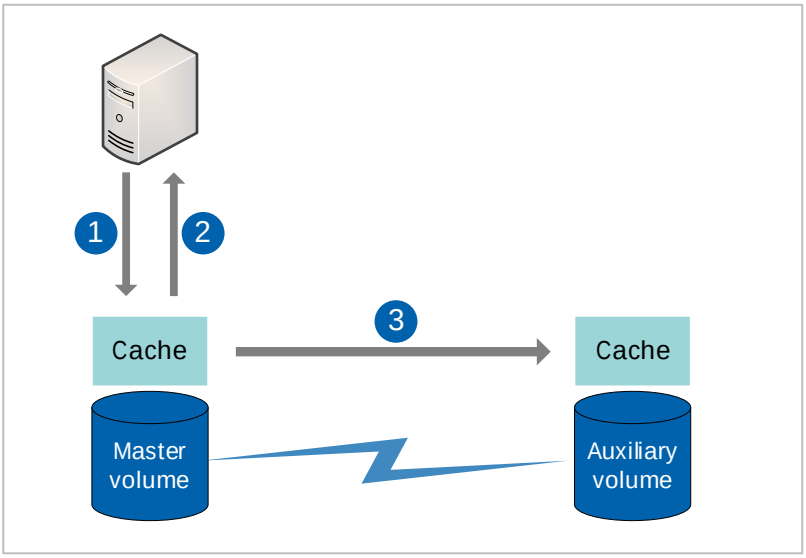


In asynchronous remote copy, the data of the auxiliary volume at each point-in-time cannot be the same

as the data of the master volume in real-time, with a risk of data inconsistency between the master volume and the auxiliary volume. However, the algorithm in this mode can keep the consistency of a mapping on the auxiliary-end through identifying the current active I/O of the master volume. When multiple I/O operations occur at the same time, the auxiliary-end will do multiple I/O operations in sequence, thus long-distance link latency will has no effect on the copy operation. The process of asynchronous remote copy of a write operation is as shown in Figure 4-7.

1. The host-end completes a write operation, writing the data into the cache of the master volume.
2. The master volume sends feedback that the write operation has been completed to the host-end.
3. All I/O data are transmitted sequentially from the master volume to the auxiliary volume.

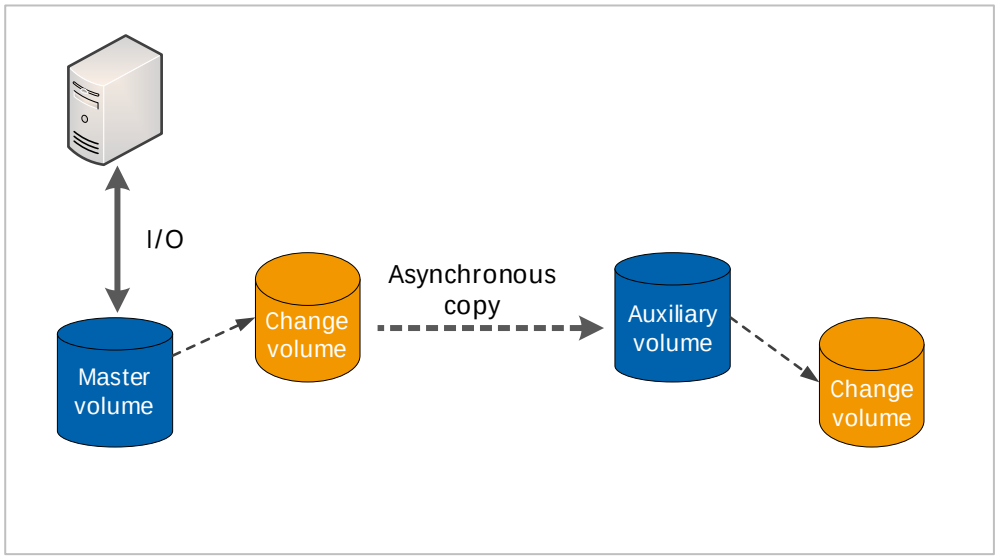
Figure 4-7 Asynchronous remote copy



In periodic asynchronous remote copy, when a remote copy relationship is created, a change volume will be created automatically for the master volume and the auxiliary volume respectively. The change volume uses the snapshot function of InLocalCopy, but it cannot be used like a snapshot. In this mode, a series of snapshot copies will be transmitted from the master-end to the auxiliary-end within a constant cycle (the default value is 300 seconds, adjustable between 60 seconds and 24 hours), applicable to low-bandwidth network environments. The process of periodic asynchronous remote copy is as shown in Figure 4-8.

1. Create a periodic asynchronous remote copy relationship, initiate the synchronization, and create a change volume for the master volume and the auxiliary volume.
2. The master-end starts the asynchronous remote copy to the auxiliary-end once automatically in constant cycle. The process of each cycle is as shown below.
   a. Start the snapshot relationship between the master volume and the change volume.
   b. Start the snapshot relationship between the auxiliary volume and the change volume.
   c. Transmit the data on the change volume of the master volume to the auxiliary volume, and wait until complete.
   d. Stop the snapshot relationship between the master volume and the change volume.
   e. Stop the snapshot relationship between the auxiliary volume and the change volume.

Figure 4-8 Periodic asynchronous remote copy



For InRemoteCopy, a partnership between two storage systems located in different locations must be created and configured, and then two storage systems can communicate through the remote copy link. The remote copy link can be FC link and IP link, and the IP link is used more widely. The technologies that follow are used in the IP remote copy.

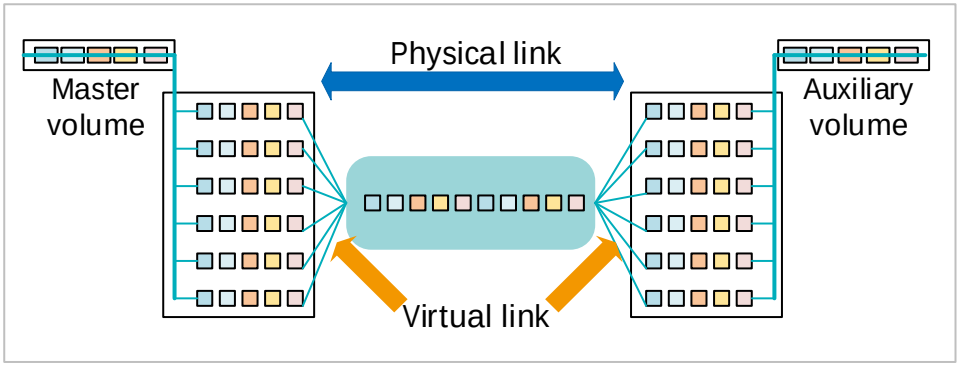
- Compression technology

  Both software compression and hardware compression are supported, reducing the actual amount of data that to be transmitted on the copy link. Software compression consumes the CPU resources of the system, while hardware compression decreases the CPU resource consumption through InCompression.

- Link virtualization technology

  Without link virtualization, after the master-end sends the copied data to the auxiliary-end, the master-end must wait for the feedback of the completion of data copy from the auxiliary-end before the next data transmission. For long-distance IP network, network latency is a key factor that restricts the transmission rate, with higher latency causing lower transmission rate. Link virtualization technology can virtualize a physical link into multiple links that will transmit data together, as shown in Figure 4-9, thus reducing the effect of network latency.

  Figure 4-9 Link virtualization technology



If the data packet is lost from any virtual link, the data will be sent again. The number of virtual links is controlled by the AI engine, which can adjust the number of virtual links accordingly by monitoring the link performance during data transmission. The virtual links that have been set will be saved in the controller and the saved virtual links will be used to start the link again when a link is stopped.

Technical specifications

The technical specifications of InRemoteCopy are as shown in Table 4-5.

Table 4-5 Technical specifications of InRemoteCopy

Attribute	Value
Maximum number of partnerships per system	3 (three FC partnerships, or two FC partnerships and one IP partnership)
Maximum number of remote copy relationships (synchronous, asynchronous, and periodic asynchronous) per system	KS2000G7: 2048 KS3000G7: 4096
Maximum number of periodic asynchronous remote copy relationships per system	KS2000G7: 1024 KS3000G7: 2048
Maximum number of consistency groups per system	256
Maximum number of remote copy relationships per consistency group	KS2000G7: 1024 KS3000G7: 2048
Total capacity of remote copy volumes and InMetro volumes per I/O group	16 PB

Relationship with other features

For InRemoteCopy, there are no limits on the types of the master volume and the auxiliary volume, which can be any combination of normal volumes, thin volumes, compressed volumes, deduplicated volumes, and compressed-deduplicated volumes. The capacity of the master volume and the auxiliary volume must be the same.

For a volume used in local copy and remote copy at the same time, the interactivity is as shown in Table 4-6.

Table 4-6 Interactivity of local copy volumes and remote copy volumes

Volume	Whether supported
Source volume in a local copy and master volume in a remote copy	Supported. Local copy and remote copy are completed independently.
Source volume in a local copy and auxiliary	Supported.

volume in a remote copy	
Target volume in a local copy and master volume in a remote copy	Supported with limits. The local copy mapping and remote copy relationship must be in the same I/O group.
Target volume in a local copy and auxiliary volume in a remote copy	Supported with limits. The local copy mapping must be in idle_or_copied state.

The support for related operations on the master volume and the auxiliary volume for remote copy is as shown in Table 4-7.

Table 4-7 Support for volume operations (InRemoteCopy)

Operation	Whether supported
Volume in image mode	Supported.
Extent migration within a storage pool	Supported, without effect on host I/O.
Volume migration between storage pools	Supported, without effect on host I/O.
Volume migration between I/O groups	Not supported.
Capacity expansion	Supported.
Capacity shrinkage	Not supported.
Change of the preferred node within a I/O Group	Supported (only attributed volume).

Benefits

- Long-distance disaster recovery

  Synchronous remote copy supports the deployment distance a maximum of 300 km between two storage systems, enabling real-time disaster recovery over long distances, making sure of data consistency between the master volume and the auxiliary volume.

- Low performance cost

  Asynchronous remote copy copies data to the auxiliary-end in the write sequence of I/Os, with a certain latency and supports data transmission over longer distances. Asynchronous remote copy can decrease the effect on the business performance of front-end hosts and shorten the waiting time of host applications, making sure that the throughput of the system increases with the expansion of system capacity, and guaranteeing the consistency of constantly growing data groups.

- High bandwidth adaptability

  Periodic asynchronous remote copy makes sure that consistent data are always kept on the auxiliary-end, and makes sure of the data consistency during continuous synchronization. This mode is

designed specifically for low-bandwidth asynchronous copy, permitting the bandwidth to change with different data latency requirements.

- Efficient performance optimization

  Combined with compression technology and link virtualization technology, InRemoteCopy not only decreases the amount of data transmitted and the requirements for bandwidth, but also increases the number of virtual links to transmit data together, making the data transmission rate close to full bandwidth.

Application scenarios

- Disaster recovery scenarios

  Applicable to intra-city disaster recovery scenarios in different server rooms, buildings, and locations, and remote disaster recovery scenarios with a distance of a maximum of 8,000 km. InRemoteCopy provides three modes, which can be selected referring to the site distance and the application performance requirements.

- High network latency and low bandwidth scenarios

  Low network bandwidth, congestion or overload will have effect on the operation of host-end businesses and the data consistency between master-end and the auxiliary-end, and even cause errors. For such scenarios, IP asynchronous remote uses network optimization technology to decrease the risks caused by network problems.

4.3 InMetro

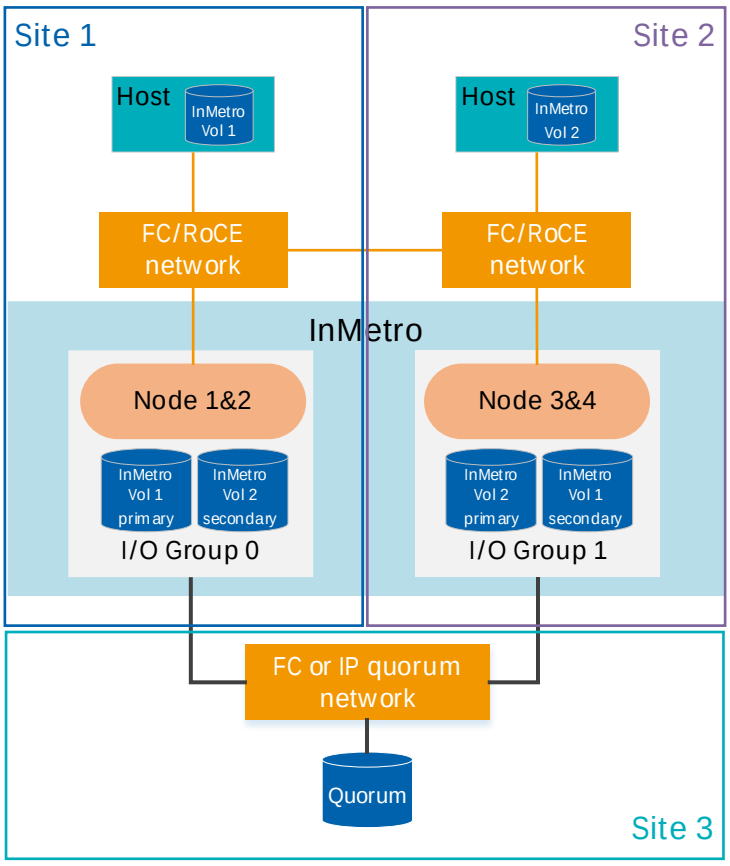
Feature description

InMetro is implemented based on a clustered system, where two I/O groups can manage each volume at the same time. InMetro uses synchronous remote copy technology, change volume technology, and InMigration technology.

Feature principle

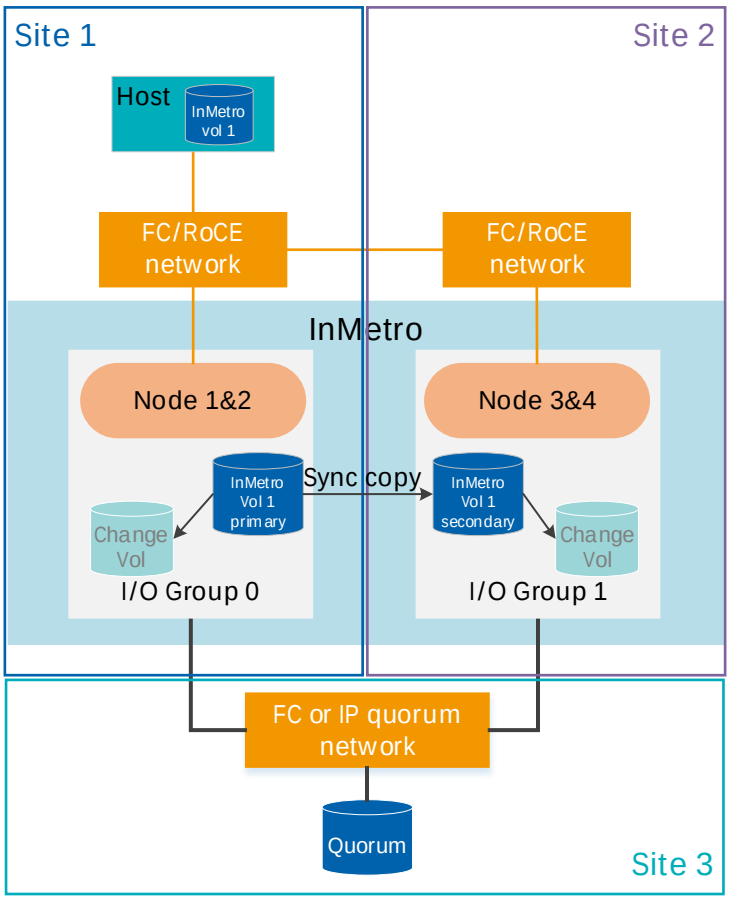
The site deployment of InMetro is as shown in Figure 4-10, hosts and storage systems should be deployed on both sites. The controllers of each I/O group belongs to one site, but each volume is visible and accessible for both sites (I/O groups) as a single object. From the specified site, the host can get access to the related volume copy and transfer to the other copy in case of a failure.

Figure 4-10 Site deployment



The two I/O groups on different sites synchronized through synchronous remote copy. Normally, the primary volume provides data access services, while the secondary volume is not mapped to the host and keeps in offline. The data can be synchronized within the clustered system in real-time through synchronous remote copy, and the synchronous remote copy relationship is automatically controlled by the storage system and cannot be started or stopped manually. The data synchronization between I/O groups is as shown in Figure 4-11.

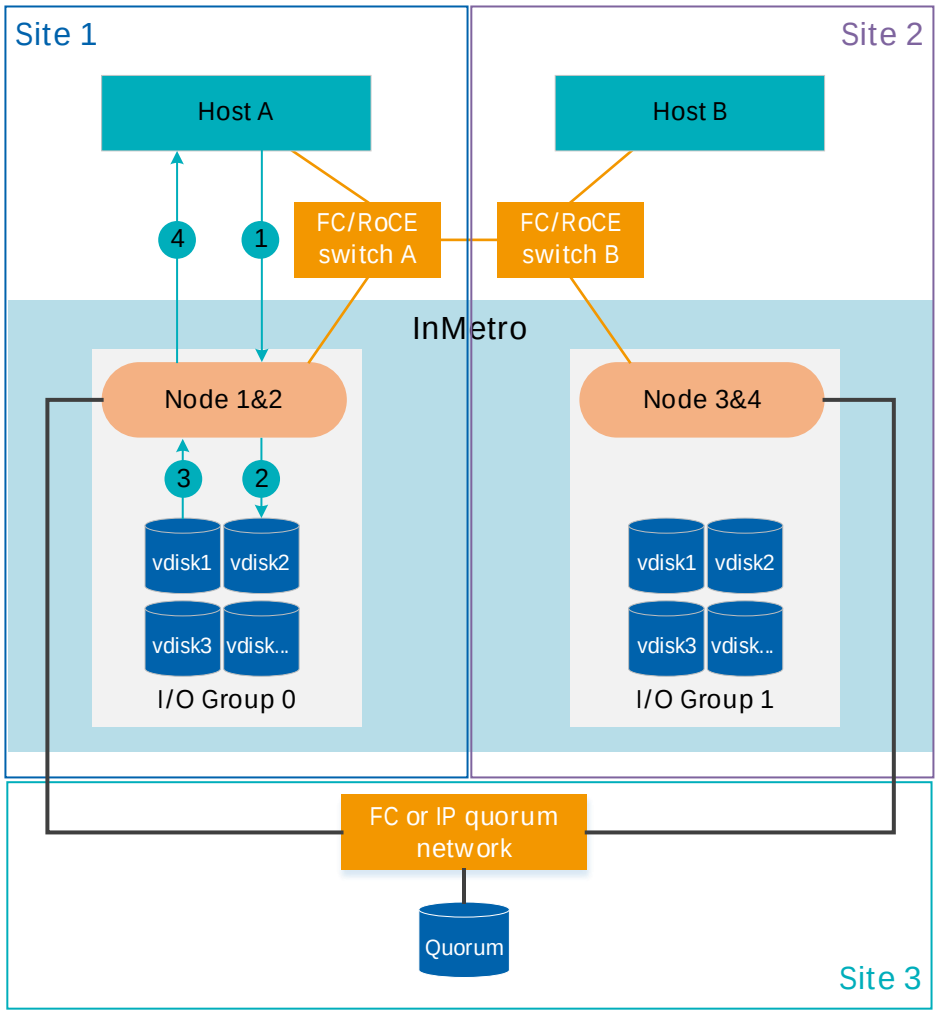
Figure 4-11 Data synchronization between I/O groups



The read process of InMetro is as shown in Figure 4-12, and the descriptions are as shown below. When the data to be read is in the cache, skip steps 2 and step 3.

1. The host-end sends a read request.
2. The storage-end sends a read request.
3. The storage-end reads data.
4. Data are sent to the host-end.

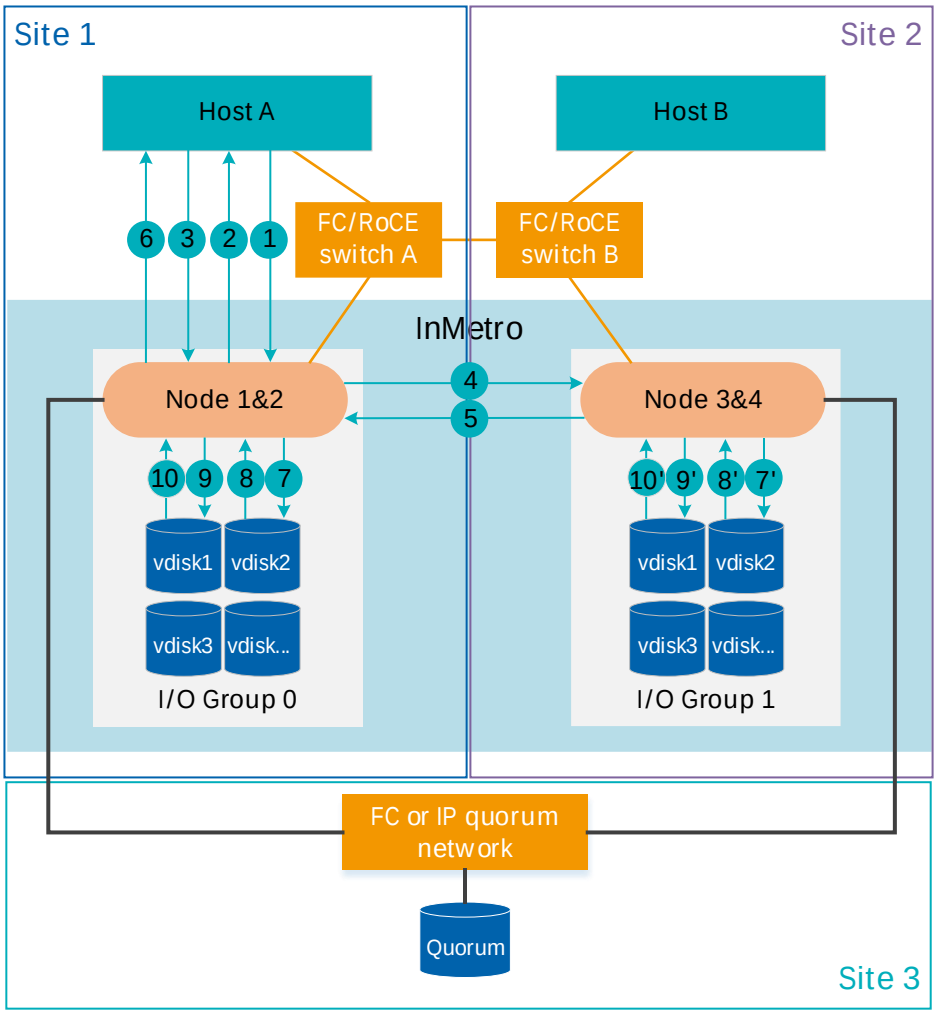
Figure 4-12 Read process of InMetro



The write process of InMetro is as shown in Figure 4-13, and the descriptions are as shown below. Step 1 to step 6 are synchronous operations, which will increase latency and have effect on the performance of host applications. In step 4 and step 5, data are transmitted to the remote storage-end only once during the I/O write. Step 7 to step 10 are asynchronous operations, which have no effect on the performance of host applications.

1. The host-end sends a write request.
2. The local storage-end sends feedback that the data can be written to the host-end.
3. The host-end writes data to the cache of the local storage-end.
4. The cached data of the local storage-end are copied to the remote storage-end.
5. The remote storage-end sends feedback that the data have been written to the local storage-end.
6. The local storage-end sends feedback that the write operation has been completed to the host-end.
7. The storage-end sends a write request.
8. The drive sends feedback that the data can be written to the storage-end.
9. The storage-end writes data to the drive.
10. The drive send feedback that the write operation has been completed to the storage-end.

Figure 4-13 Write process of InMetro



Technical specifications

The technical specifications of InMetro are as shown in Table 4-8.

Table 4-8 Technical specifications of InMetro

Attribute	Value
Maximum number of InMetro volumes per system	1024
Maximum capacity of remote copy volumes and InMetro volumes per I/O group	4 PB (per-controller cache = 64 GB) 16 PB (per-controller cache ≥ 128 GB)

Relationship with other features

Relationship with other features are the same as InRemoteCopy. For more information, see 4.2 InRemoteCopy.

Benefits

- In the scenario of InMetro, all controllers of an I/O group are in the same enclosure, and there are the copied data of each other in the two I/O groups. Thus, it can make sure of complete read-write performance on the host-end even though there is only one I/O group.
- InMetro supports consistency groups of InMetro volumes, and the data of each volume can be managed consistently, applicable to more disaster recovery scenarios.
- InMetro can make sure of business continuity in scenarios such as hardware failures, device power failures, communication connection failures between controllers, and natural disasters.
- InMetro supports incremental synchronization of differential data and provides consistent point-in-time data protection.

Application scenarios

- Scenarios where business cannot be interrupted

  InMetro can effectively make sure of the business continuity with no data loss, applicable to scenarios with high business continuity requirements. The same data can be stored on two I/O groups at the same time, independent of each other. If any failure occurs at one site, the other site can quickly take over the business of the failed site.

- Scenarios with high reliability requirements

  InMetro supports multiple failure scenarios such as overall site failure, I/O group dual-controller failure, and storage subsystem failure. I/O group can seamlessly switch to backup resources and is completely transparent to the upper-level applications. Even though the site fails, the remote I/O group still has a cache-write protection mechanism to prevent performance degradation caused by write-through mode.

4.4 3DC

Note: Only KS3000G7 whose per-controller cache is 128 GB and above supports 3DC.

The three data center (3DC) is a disaster recovery solution, where the three centers refer to the production center, intra-city disaster recovery center, and remote disaster recovery center. The data from the production center are synchronously copied to the intra-city disaster recovery center, and are asynchronously copied to the remote disaster recovery center. The intra-city disaster recovery center usually has the same business processing capability as the production center, and applications can switch to the intra-city disaster recovery center without data loss, keeping continuous business operation. In case of a large-scale disaster such as a natural disaster earthquake, causing the intra-city disaster recovery center and production center to be unavailable at the same time, the remote disaster recovery center can take over the business.

Compared with intra-city disaster recovery center or remote disaster recovery center, the 3DC mode combines the advantages of the two modes and can adapt to larger-scale disaster scenarios. For small-scale regional disasters and large-scale natural disasters, the disaster recovery system can respond quickly, making sure of business security without data loss and getting better RPO and RTO.

With InMetro and InRemoteCopy (synchronous/Asynchronous), 3DC supports multiple networking schemes.

- Synchronous and asynchronous in series (A->B->C)
- Synchronous and asynchronous in parallel (C<-A->B)
- InMetro and asynchronous in series (A<->B->C)
- InMetro and asynchronous in parallel (C<-A<->B)

Figure 4-14 Network schemes of 3DC

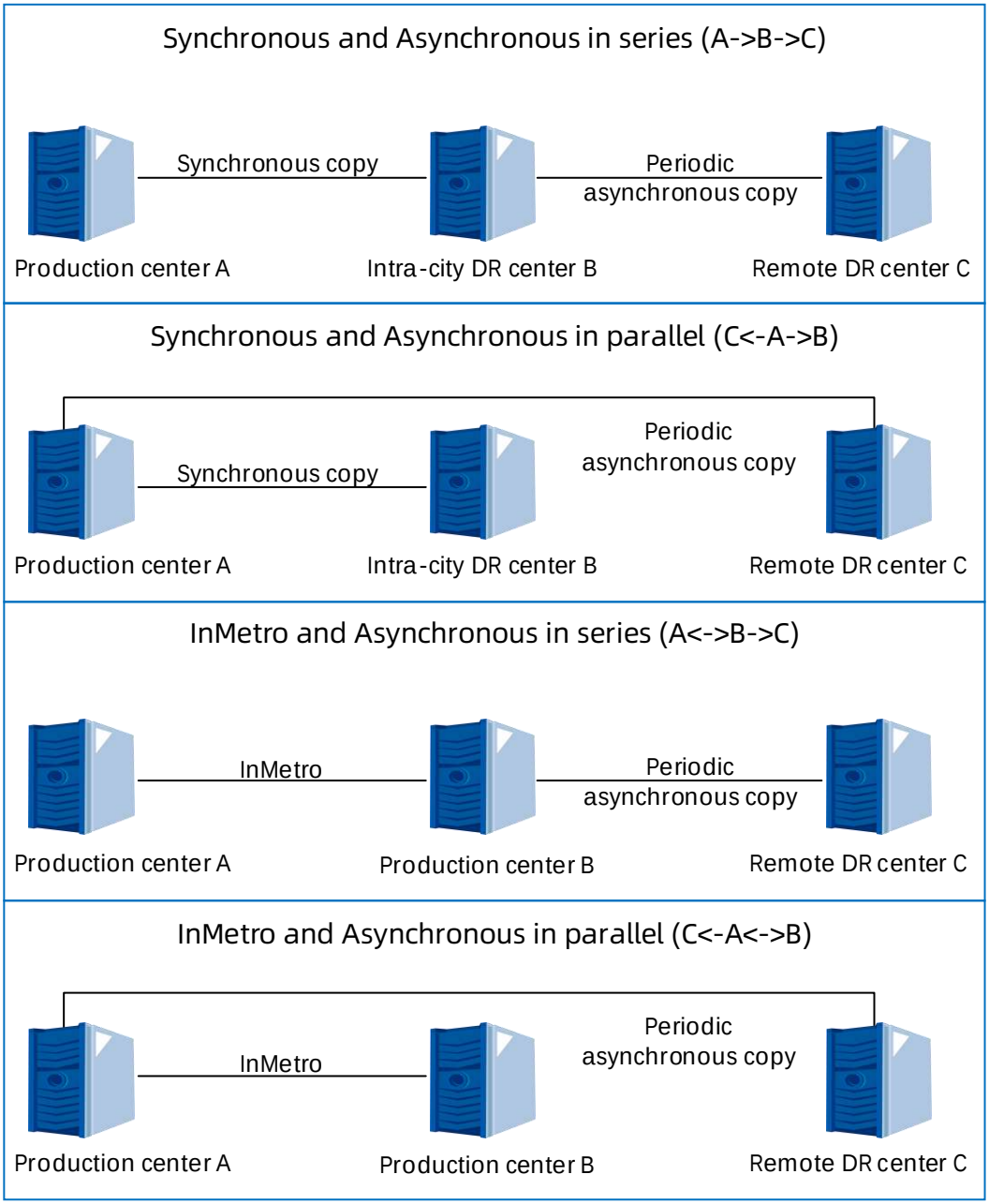


Table 4-9 Technical specifications of 3DC

Networking scheme	Maximum number of 3DC
Synchronous and asynchronous in series	2048
Synchronous and asynchronous in parallel	
InMetro and asynchronous in series	682
InMetro and asynchronous in parallel	

The highlights and advantages of 3DC solution are as shown below.

- Improve data transmission efficiency through WAN acceleration and data compression technology.

  The IP remote copy technology uses WAN acceleration technologies such as dynamically adjusting sliding windows and multi-virtualization connections, effectively solving the problems of high latency and high packet loss caused by long-distance WAN between intra-city disaster recovery center and remote disaster recovery center. Even in scenarios with poor network, it can make sure that the transmission speed is within the theoretical bandwidth range, while combining data compression technology to effectively decrease data transmission capacity and transmission cycle.

- Make sure of data consistency through consistency groups.

  Multiple 3DC strategies with the same attributes can be added to a consistency group for unified management, making sure of the consistency of collaborative data between multiple volumes with logical relationships on the host-end. A typical example is the data volume and log volume of a database, which are necessary for a transaction operation at the same time. If the backup time of the two volumes are different or only one volume is backed up, it will cause the loss of data consistency for the overall database will be inconsistent.

4.5 InErase

Feature description

InErase is used to erase data on a VDisk that are no longer necessary. The erase unit is VDisk.

InErase provides two data overwriting methods.

- DoD: The storage system overwrites the data stored on the selected VDisk for three times referring to the DoD standard.
- Custom: The storage system overwrites the data stored on the selected VDisk referring to the specified number of times. The number of overwriting times can be 3 to 100, with a default of 7.

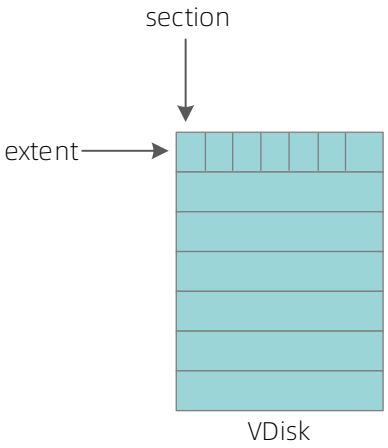
Feature principle

The specific rules for data overwriting are as shown below.

- For the first time, write all-zero data for overwriting.
- For the second time, write all-1 data for overwriting.
- For the third to the 100th time, write random data for overwriting.

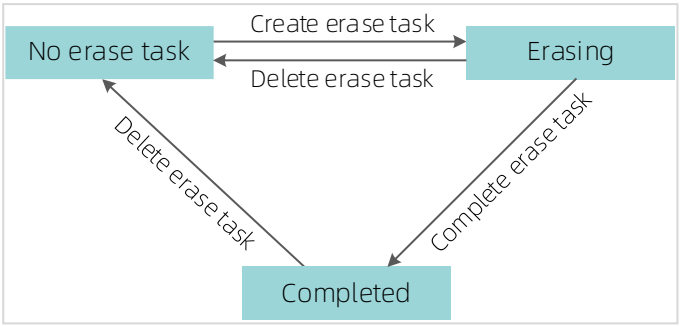
A VDisk is composed of one or several extents in a storage pool, and the data erase function divides each extent into several 256 KB sections, as shown in Figure 4-15.

Figure 4-15 VDisk



The minimum unit of data overwriting is section. To do an erase task, the overwriting operation starts from the first section of the first extent of the VDisk until the VDisk is completely overwritten. The state transition of an erase task is as shown in Figure 4 18.

Figure 4-16 State transition of erase task



Relationship with other features

InErase is only applicable to normal volumes in normal pools. Because the feature of the allocation on write and space reclamation, other volumes cannot complete the data erase of all space. There are other limits of InErase as shown below.

- Cannot do data erase on volumes that have been configured for value-added businesses, such as InLocalCopy, InRemoteCopy, and InMetro.
- Cannot do data erase on volumes that are in process of format, data migration, expansion, and compression.
- Make sure that the state of the VDisk is normal and cancel the mapping between the VDisk and the host before overwriting a VDisk.

Benefits

Make sure of the security of crucial data and sensitive data. Erase any obsolete data intelligently with one click, and no third-party tools can restore the erased data.

Application scenarios

- Scenarios where crucial data must be erased.

  For the crucial data stored in a VDisk, if they are no longer necessary, InErase can be used to completely clear them to prevent any leakage.

- Scenarios where the service life of drives is coming to an end.

  When the service life of a drive is coming to an end, it is necessary to quickly migrate the data stored on it to other drives. If the data on the initial drive are sensitive or important, InErase can be used to clear the data.

4.6 InEncryption

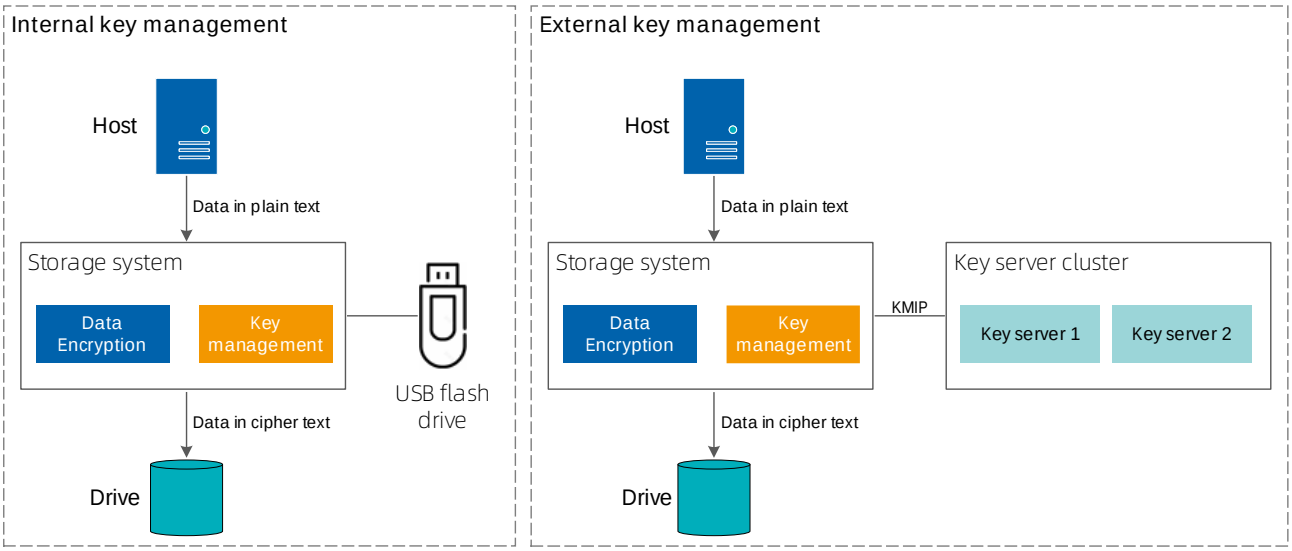
Feature description

InEncryption is used to encrypt data stored in the storage system, providing strong protection for sensitive data. InEncryption can effectively prevent illegal data access or data leakage during storage and transmission, making sure of data security and data integrity. InEncryption supports AES algorithm and SM4 algorithm for data encryption, and provides two methods for key management, including internal key management and external key management.

Feature principle

For internal key management, the master key is stored in the USB flash drives. For external key management, the master key is stored in the key servers.

Figure 4-17 Principle of key management



The process of data encryption is as shown below.

1. The host writes data in plain text to the storage system.
2. The storage system gets the master key from the USB flash drives or key servers.
3. The storage system gets the encapsulated data encryption key that should be used for the write operation.
4. The storage system uses the master key to de-capsulate the encapsulated data encryption key and get the data encryption key.
5. The storage system uses the data encryption key to encrypt the data.
6. The storage system writes data in cipher text to the drives.

The process of data decryption is as shown below.

1. The host sends a read data command to the storage system.
2. The storage system gets the master key from the USB flash drives or key servers.
3. The storage system gets the encapsulated data encryption key that should be used for the read operation.
4. The storage system uses the master key to de-capsulate the encapsulated data encryption key and get the data encryption key.
5. The storage system reads data in cipher text from the drives and uses the data encryption key to decrypt the data.
6. The storage system sends the data in plain text to the host.

Technical specifications

The technical specifications of InEncryption are as shown in Table 4-10.

Table 4-10 Technical specifications of InEncryption

Attribute	Value
Maximum number of key servers	1

Minimum number of USB flash drives	2

Relationship with other features

InEncryption has no conflict with other advanced features.

Benefits

- Data security

  InEncryption provides an additional protection for data in the storage system, which can prevent sensitive data leakage to meet the demands of enterprises for data confidentiality.

- Deployment flexibility

  Both USB flash drives and key servers can be used for key storage and key management, improving the flexibility of the deployment of InEncryption.

Application scenarios

InEncryption is used for enterprise data protection and is applicable to scenarios where there are a large quantity of sensitive data such as commercial secrets, customer information, and financial data. InEncryption can encrypt the data stored in the storage system though encryption algorithm, effectively preventing illegal access by internal personnel and attacks by external hackers. The encrypted data will not be decrypted and leaked without the correct key although the storage device is stolen or the data is intercepted during transmission, thus making sure of the confidentiality and integrity of enterprise data.

5 Efficient host plug-ins

KAYTUS storage provides rich interface software (plug-ins) for host applications, which can be used in different data centers to enhance storage availability and ease of use, effectively reducing management costs. Host plug-ins can be got and used free without license.

5.1 InPath plug-ins

Multipath I/Os refer to the multiple physical connections between the host-end and the storage-end, for high availability, high security, and high throughput. Through the multipath access mechanism provided by the host operating system, the storage devices can be get access to on multiple physical paths.

KAYTUS storage provides related InPath plug-in as multipath management and configuration tool for different platforms including AIX, Linux, VMware, and Windows. InPath plug-in provides functions such as failover, failback, fault isolation, path selection, and load balance, and supports online upgrade.

- InPath plug-in can sense the link status in real time and select the optimal path to process host I/Os each time to make sure of the continuity and stability of host business. When one of the physical links between host-end and storage-end is not stable, InPath plug-in will try to isolate all paths on that link and lower the priority of the isolated paths at the same time. The isolated paths will be selected only when there are no other available paths. InPath plug-in can automatically de-isolate related paths when it senses that the link has restored to be stable.
- InPath plug-in can be upgraded online, which makes sure that host business can be processed with no interruption during the upgrade. Online upgrade can decrease upgrade time and decrease the unwanted effect of upgrade on host business to a minimum.

InPath for AIX

The multipath framework of AIX platform includes a multipath disk driver and a path-control module. The multipath disk driver provides basic functions and the path-control module provides functions related to specified devices, such as finding devices and selecting paths. The multipath disk driver and path-control module work together to complete the multipath access to storage devices.

InPath for AIX is a multipath package, which is developed based on the multipath framework of AIX platform. The package includes a path-control module and a set of tools with functions such as path management, performance monitoring, and adaptation information query. The path-control module is a kernel extension that can configure related parameters through configuration files, load them into the kernel at runtime, and provide function calls related to multipath for AIX system to complete multipath access to this storage device. InPath for AIX can select the optimal path to send I/O requests, make KAYTUS

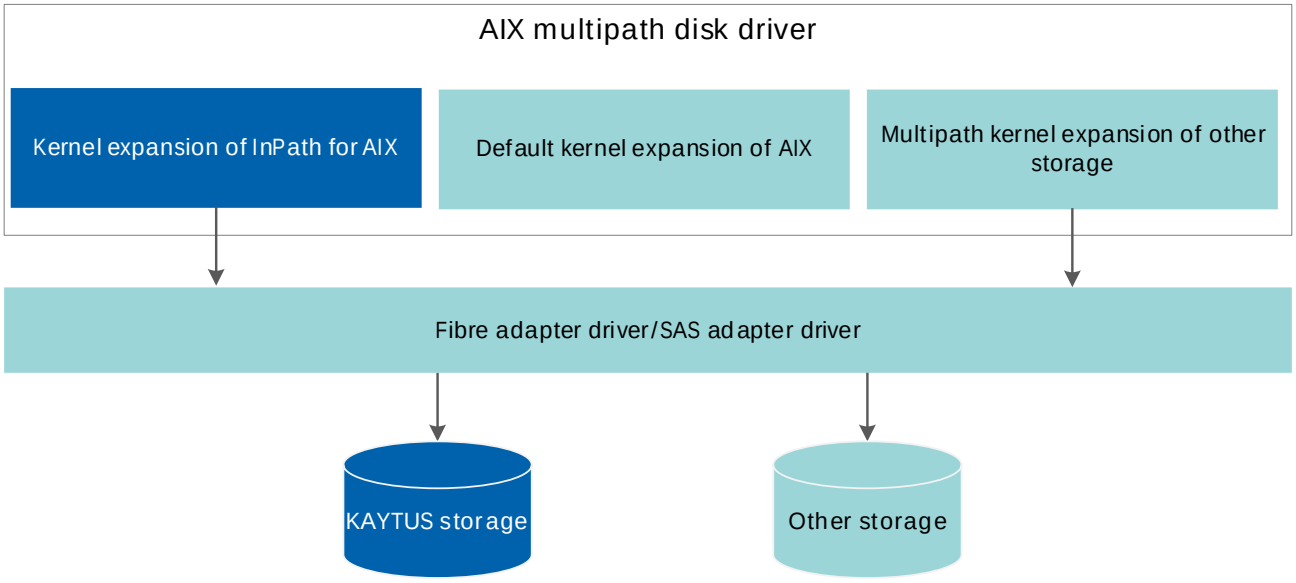
storage get high stability and high performance, and make storage device easy to manage and upgrade.

InPath for AIX works in collaboration with the multipath disk driver of AIX platform to provide multipath functions for KAYTUS storage. In addition to basic functions such as failover, failback, fault isolation, and load balance, it also provides unique functions such as I/O statistics, read/write performance monitoring, and log tracing.

- Failover: When an I/O failure occurs on one path, the I/O request will be sent to the other path to continue.
- Failback: When a failed path goes back to correct, it will be added to the available path set in real time.
- AIX HACMP: Multiple hosts can get access to logical units in storage devices in parallel, enabling higher availability for critical business applications.
- Load balance: The load balance algorithms can be set and take effect in real time. Get better access efficiency of storage devices by selecting the optimal path from available path set. Load balance algorithms of InPath for AIX are as shown below.
  - round robin
  - load balance
  - failover
  - load balance port
  - least queue depth
- Open/close path in real time: The path switch for each storage unit can be set and take effect in real time, using system management interface tool (SMIT) or toolkit of InPath for AIX.
- Monitor path performance: Through the InPath toolkit in the software package, the system administrator can get performance and access statistics of devices and paths in real time to make an estimate of performance and status of the storage system.
- Trace log in real time: Make an analysis of the failure and performance bottleneck related to multipath storage devices through the operation logs of InPath for AIX.
- One-stop System Information Collection: InPath for AIX has a script tool named inpathpcmgetdata that can be used to collect system information and pack to a file, which helps system administrators quickly collect data and save time. The system information includes information of host system and driver version related to multipath, HBA card, configuration information of storage devices, operation logs of InPath for AIX, and so on.

InPath for AIX and multipath software from other vendors can run in the system at the same time and there is no interference between them. Its location in the system is as shown in Figure 5-1.

Figure 5-1 Multipath module of AIX Platform



InPath for Linux

The multipath framework of Linux platform includes a multipath disk driver and a path-control module. The multipath disk driver provides basic functions and the path-control module provides functions related to specified devices, such as finding devices and selecting paths. The multipath disk driver and path-control module work together to complete the multipath access to storage devices.

InPath for Linux is a multipath package, which is developed based on the multipath framework of Linux platform. The package includes a path-control module and a set of tools that provide management functions. InPath for Linux can select the optimal path to send I/O requests, make KAYTUS storage get high stability and high performance, and make storage device easy to manage and upgrade.

InPath for Linux works in collaboration with the multipath disk driver of Linux platform to provide multipath functions for KAYTUS storage, such as failover, failback, fault isolation, load balance, and multipath configuration management.

- Failover: When an I/O failure occurs on one path, the I/O request will be sent to the other path to continue.
- Failback: When a failed path goes back to correct, it will be added to the available path set in real time.
- Load balance: The load balance algorithms can be set and take effect in real time. Get better access efficiency of storage devices by selecting the optimal path from available path set. Load balance algorithms of InPath for Linux are as shown below.
  - Round Robin
    This algorithm can make each path to serve as an I/O transmission path through assigning I/O requests to storage devices to all paths in turn. For a new I/O request, the system automatically selects a valid path that is not the same as the previous path to process it.
  - Queue Length
    This algorithm can count the number of pending I/O requests in each path queue in real time

and assign I/O requests to storage devices to the path with smallest number of pending I/O requests in the queue. For a new I/O request, the system automatically selects a valid path with smallest number of pending I/O requests in the queue to process it.

- Service Time
  This algorithm can count the time necessary for processing I/O in each path queue in real time and assign I/O requests to storage devices to the path with shortest service time of pending I/O in the queue.
- Least Bytes
  This algorithm can count the total number of bytes processed in each path queue in real time and assign I/O requests to storage devices to the path with minimum number of bytes in the queue.

- Path selection policy: The path selection policy for each storage unit can be set.
- Path switch: The path switch for each storage unit can be open and closed.

InPath for VMware

VMware multipath framework, called Pluggable Storage Architecture (PSA), is an open, modular framework that coordinates the simultaneous operation of multiple software modules. Storage vendors can develop three types of multipath plug-ins, including Multi-Pathing Plug-ins (MPPs), Storage Array Type Plug-ins (SATPs), and Path Selection Plug-ins (PSPs).

InPath for VMware is a PSPs multipath package, which is developed based on the multipath framework of VMware platform. The package includes path-control modules and an auxiliary tool. InPath for VMware provides two new path selection policies in addition to the three native path selection policies of VMware. InPath for VMware can work in coordination with the VMware multipathing module called Native Multipathing Plug-In (NMP) to provide multipath support for the connected storage devices in the system, effectively improving storage performance.

InPath for VMware provides functions such as failover, failback, load balance, and path configuration management.

- Failover: When an I/O failure occurs on one path, the I/O request will be sent to the other path to continue.
- Failback: When a failed path goes back to correct, it will be added to the available path set in real time.
- Load balance: The load balance algorithms can be set and take effect in real time. Get better access efficiency of storage devices by selecting the optimal path from available path set. Load balance algorithms of InPath for VMware are as shown below.
  - Least Queue Depth (INP_PSP_LQD)
    This algorithm can count the number of pending I/O requests in each path queue in real time and assign I/O requests to storage devices to the path with smallest number of pending I/O requests in the queue. For a new I/O request, the system automatically selects a valid path with smallest number of pending I/O requests in the queue to process it.
  - Min Task (INP_PSP_MT)
    This algorithm can count the total number of pending bytes in each path queue as queue task

in real time and assign I/O requests to storage devices to the path with smallest number of queue task.

- Path selection policy: The path selection policy for each storage unit can be set.
- Path switch: The path switch for each storage unit can be open and closed.

InPath for Windows

The multipath framework of Windows platform includes a multipath driver named mpio.sys and device support module. The multipath driver provides basic functions and the device support module provides functions related to specified devices, such as finding devices and selecting paths.

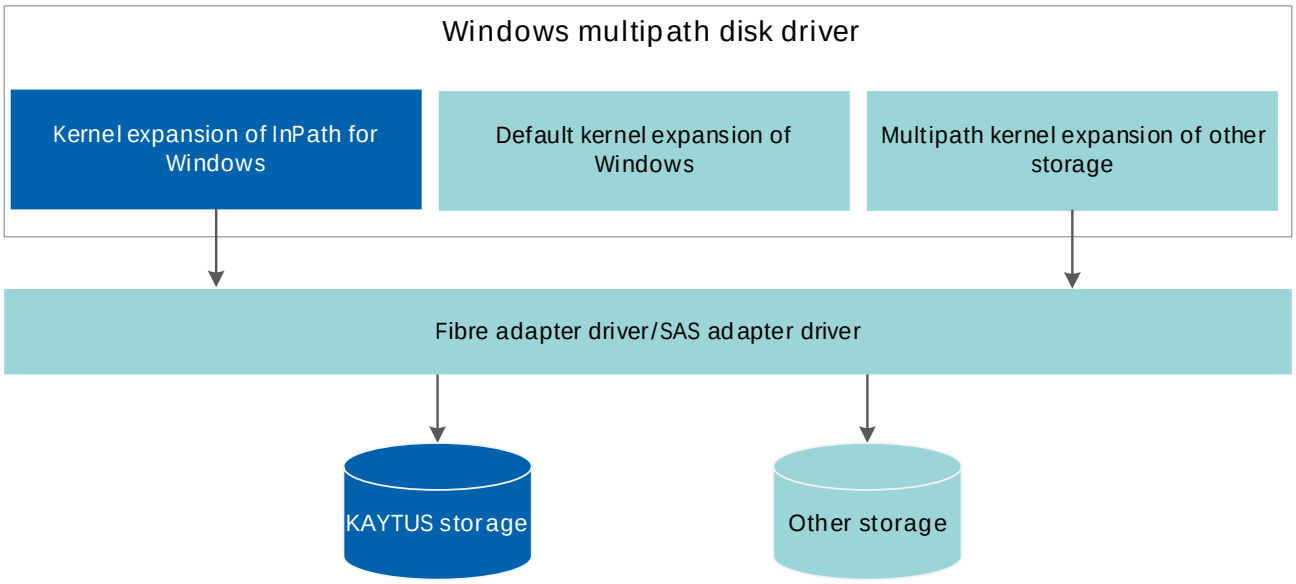
InPath for Windows is a multipath package, which is developed based on the multipath framework of Windows platform. The package includes a device support module named inpathdsm.sys and a set of tools that provide management functions. The device support module is developed based on the multipath software of Windows platform, working as a plugin to provide support for KAYTUS storage.

InPath for Windows works in collaboration with the multipath driver of Windows platform to provide multipath functions for KAYTUS storage. In addition to basic functions such as failover, failback, fault isolation, and load balance, it also provides unique functions such as I/O statistics, read/write performance monitoring, and log tracing.

- Failover: When an I/O failure occurs on one path, the I/O request will be sent to the other path to continue.
- Failback: When a failed path goes back to correct, it will be added to the available path set in real time.
- Host cluster: Multiple hosts can get access to logical units in storage devices in parallel.
- Load balance: The load balance algorithms can be set and take effect in real time. Get better access efficiency of storage devices by selecting the optimal path from available path set. Load balance algorithms of InPath for Windows are as shown below.
  - failover
  - load balance
  - load balance ext
  - shortest queue service time
  - least queue depth
  - least bytes
  - round robin
  - round robin with subset
- Open/close path in real time: The path switch for each storage unit can be set and take effect in real time.
- Monitor path performance: Monitor the I/O performance of paths through the performance monitor of Windows in real time, or get the I/O performance of paths through tools of InPath for Windows.
- Trace log in real time: Make an analysis of the failure and performance bottleneck related to multipath storage devices through the operation logs of InPath for Windows.
- Windows Event log: Give standard Windows Event logs for easily getting the status changes of storage system.

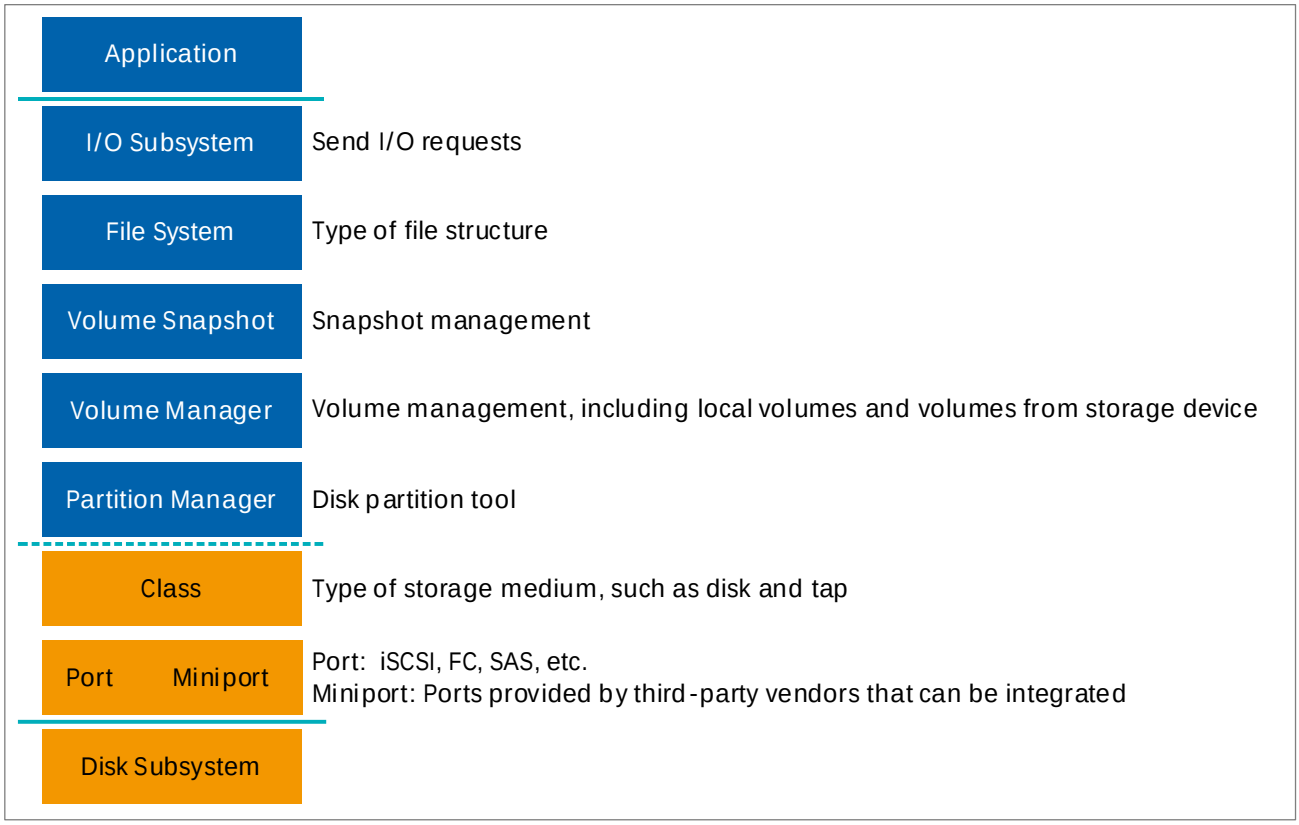
InPath for Windows can run with multipath software from other vendors at the same time and there is no interference between them. Its location in the system is as shown in Figure 5-2.

Figure 5-2 Multipath module of Windows Platform



The multipath module is located in the Class layer of the storage I/O stack, and the storage I/O stack of Windows platform is as shown in Figure 5-3.

Figure 5-3 Storage I/O stack of Windows platform



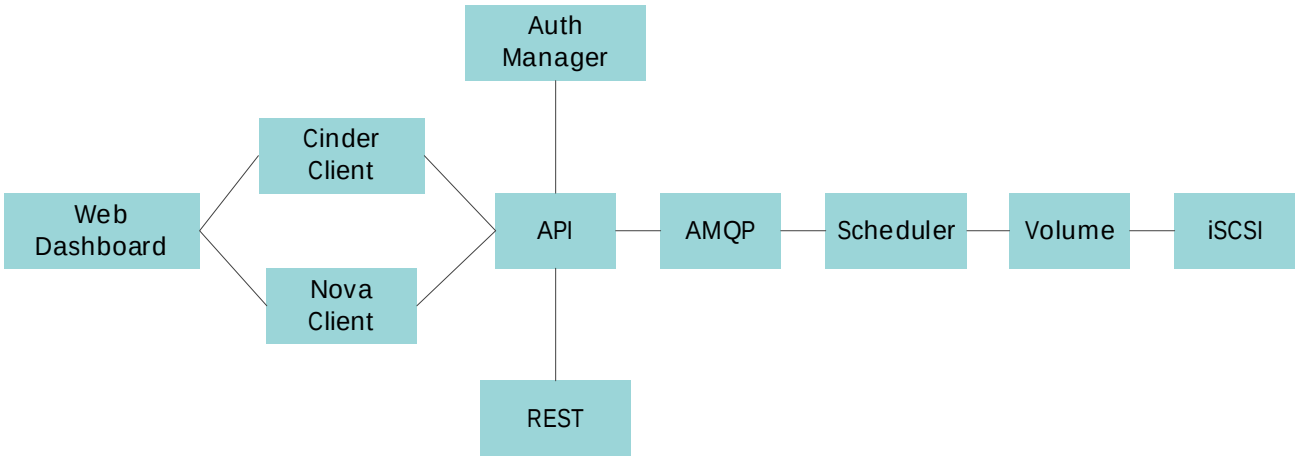
5.2 Cinder plug-in

OpenStack is an open-source cloud computing management platform project that combines several major components to complete specific tasks. Cinder is one of the three core components (Nova, Neutron, and Cinder) of OpenStack, mainly providing block storage services. OpenStack starts to replace the Nova-Volume service with Cinder from the Folsom version.

Cinder mainly includes three modules, including API, Scheduler, and Volume, its architecture is as shown in Figure 5-4. Cinder initiates request commands through CLI and dashboard, selects appropriate volume controllers with scheduler, and calls the drivers of storage manufacturers to operate and manage volumes, snapshots, backups, and consistency groups.

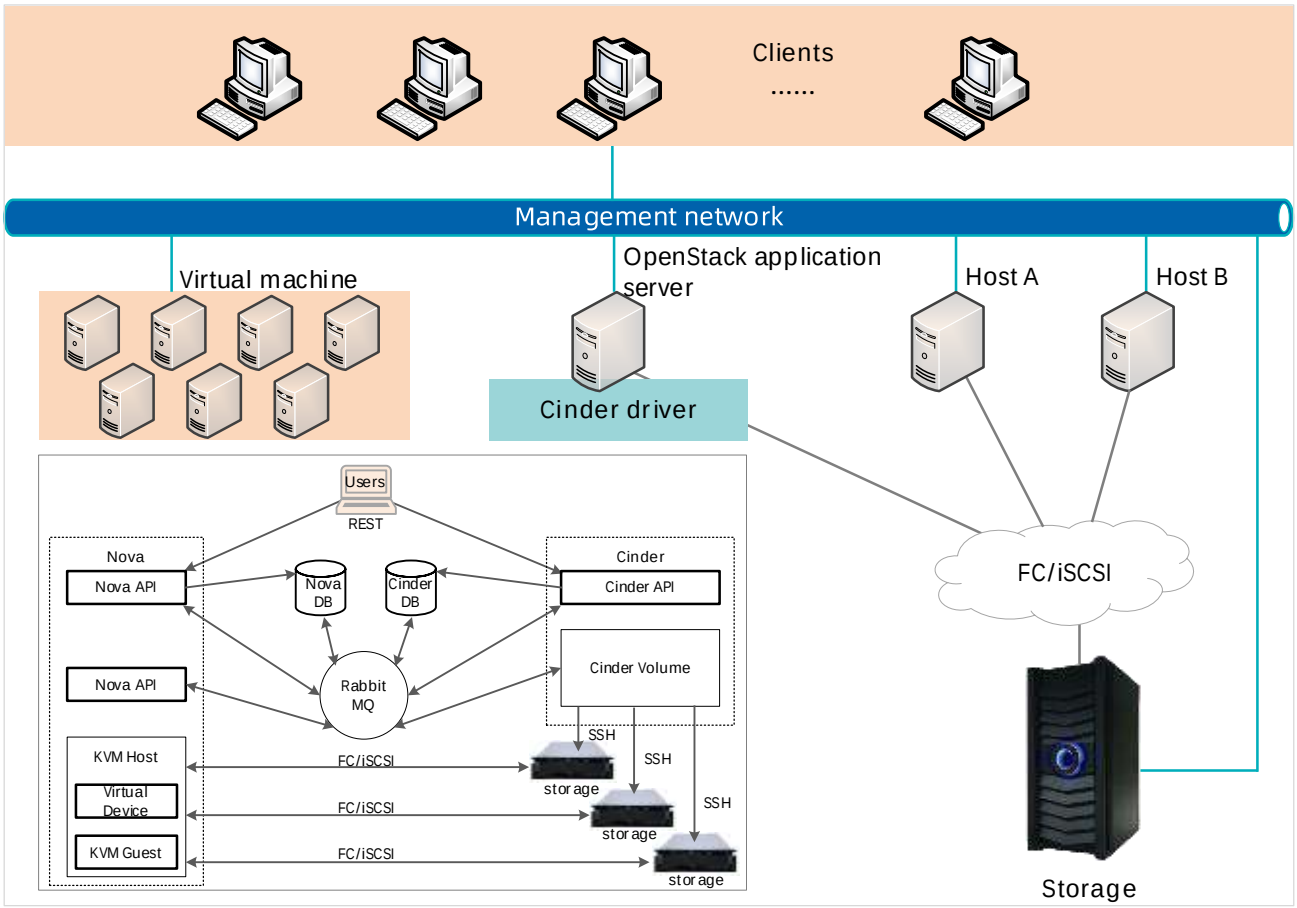
- API service: API service is the core module of Cinder, which is used to receive and process the REST request, and put them on the RabbitMQ queue.
- Scheduler service: Scheduler service is used to process tasks in the task queue and select appropriate volume service controllers to do tasks based on predetermined strategies. One end of the scheduler is connected to the MQ queue to do tasks in the MQ queue, and the other end is connected to the volume device to do tasks on the volume controller.
- Volume service: Volume service runs on the storage controller and is used to manage the storage space. Each storage controller has a volume service, and several storage controllers can be combined to form a storage resource pool.

Figure 5-4 Cinder architecture



Cinder relies on storage devices to work, which can be connected through Logical Volume Manager (LVM) or drivers of storage vendors. Cinder plug-in developed by KAYTUS storage is a driver for Cinder service in OpenStack and can manage the storage through OpenStack. The storage can be connected to OpenStack virtual machines through FC protocol, IP protocol, and RDMA protocol. Through Cinder driver, Cinder can connect and use the storage to provide block storage service for OpenStack users, with better performance. The application topology of Cinder driver is as shown in Figure 5-5.

Figure 5-5 Application topology of Cinder driver

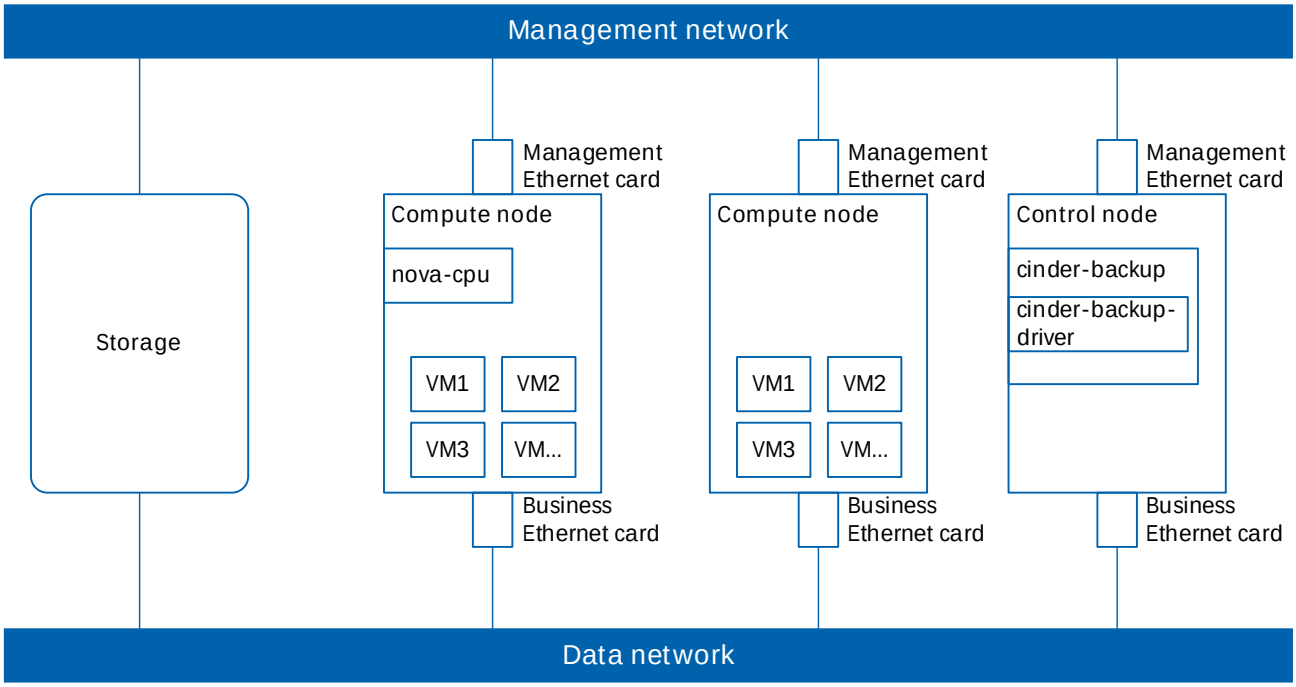


5.3 Cinder-backup plug-in

Cinder-backup plug-in developed by KAYTUS storage is a driver for backup in Cinder service of OpenStack. It can create backups for Cinder volumes and manage the backups. As an interface, Cinder-backup driver can interact with the front-end OpenStack platform and connect to the back-end storage to complete backup operations using storage snapshots through CLI commands. Cinder-backup driver can create backups on different storage devices, which means that the source volumes and backup volumes can be in different storage systems.

Through Cinder-backup driver, Cinder can use the snapshot function of KAYTUS storage to create, roll back, and delete backups. The application topology of Cinder-backup driver is as shown Figure 5-6.

Figure 5-6 Application topology of Cinder-backup driver

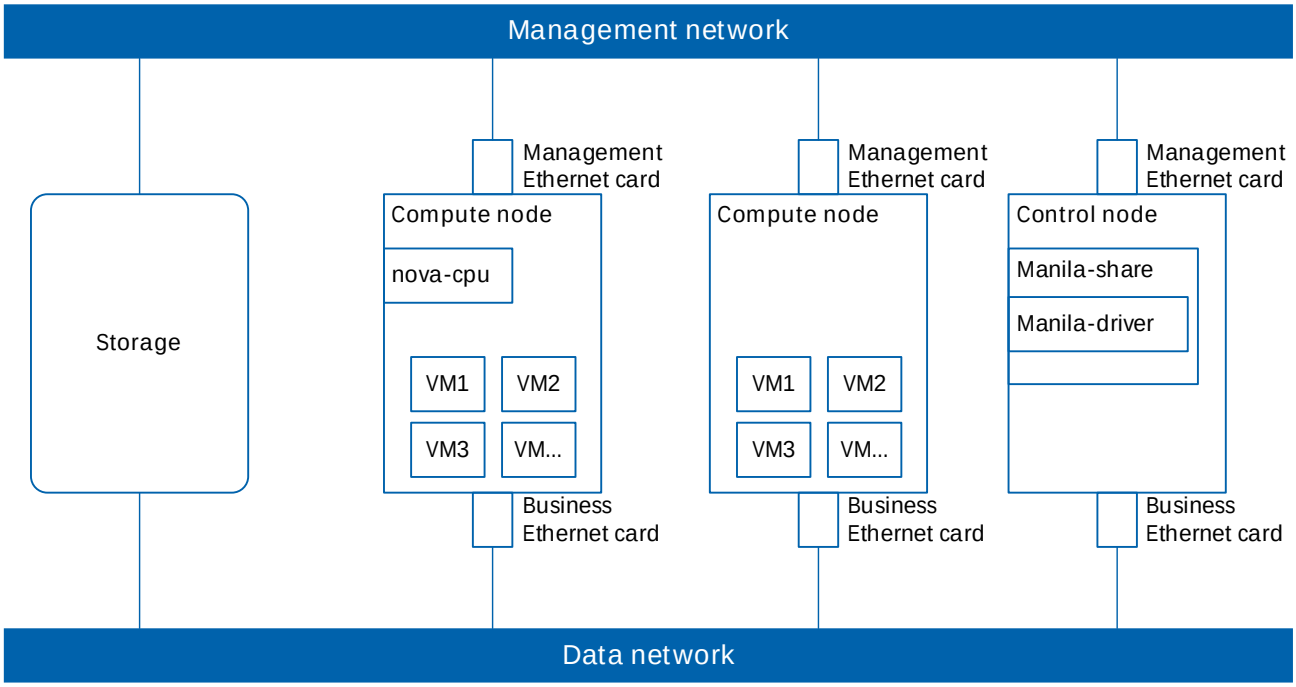


5.4 Manila plug-in

Manila is the OpenStack Shared File Systems service and provides a set of services for management of shared file systems in cloud environment as a component in OpenStack. Manila provides shared file systems for storage devices to virtual machines in the OpenStack environment. By calling the underlying drivers, the file share services of storage systems are integrated into OpenStack. Manila cooperates with Neutron services to configure the network so that virtual machines can get access to the file share services of storage systems.

Manila plug-in developed by KAYTUS storage is a driver for Manila service in OpenStack and can manage the storage through Manila service. Through Manila driver, Manila can use shared file systems provided by KAYTUS storage, enabling virtual machines in the OpenStack environment to use the NFS and CIFS share services and managing the share, share capacity, and share permissions on the storage. The application topology of Manila driver is as shown in Figure 5-7.

Figure 5-7 Application topology of Manila driver



5.5 K8sCSI plug-in

K8sCSI plug-in developed by KAYTUS storage can automatically provide persistent data storage for applications in the Kubernetes cluster. With K8sCSI plug-in, volumes, shares, and snapshots can be automatically created and managed on the storage system directly with the use of management commands of the Kubernetes cluster, and be provided to application services. K8sCSI plug-in supports FC, iSCSI, and RDMA protocols.

K8sCSI plug-in mainly includes three services, called Identity, Controller, and Node.

- Identity service is used to get the information, capabilities, and status of the plug-in.
- Controller service is used to create, delete, and clone volumes and shares, and create and delete snapshots.
- Node service is used to mount and unmount volumes and shares.

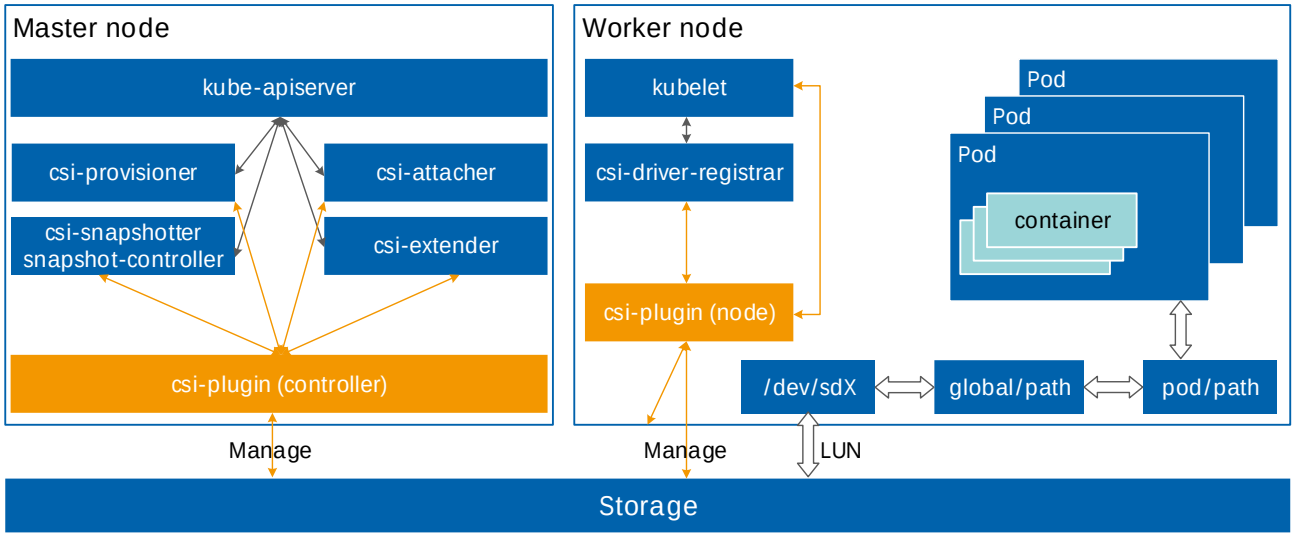
K8sCSI plug-in can work in three modes, including all-in-one mode, controller mode, and nodeworker mode.

- All-in-one mode includes the three services mentioned above, deployed on the master node of Kubernetes.
- Controller mode includes Identity service and Controller service, deployed on the master node of Kubernetes.
- Nodeworker mode includes Identity service and Node service, deployed on all nodes of Kubernetes.

Kubernetes cluster interfaces with K8sCSI plug-in through the sidecar container, which is a bridging service developed to facilitate the interfacing, and is essentially belonged to the Kubernetes CSI interface. Sidecar container interacts with Kubernetes referring to their implementation mechanism and calls K8sCSI

plug-in referring to the specifications defined by the CSI interface. The application topology of K8sCSI plug-in is as shown in Figure 5-8

Figure 5-8 Application Topology of K8sCSI plug-in



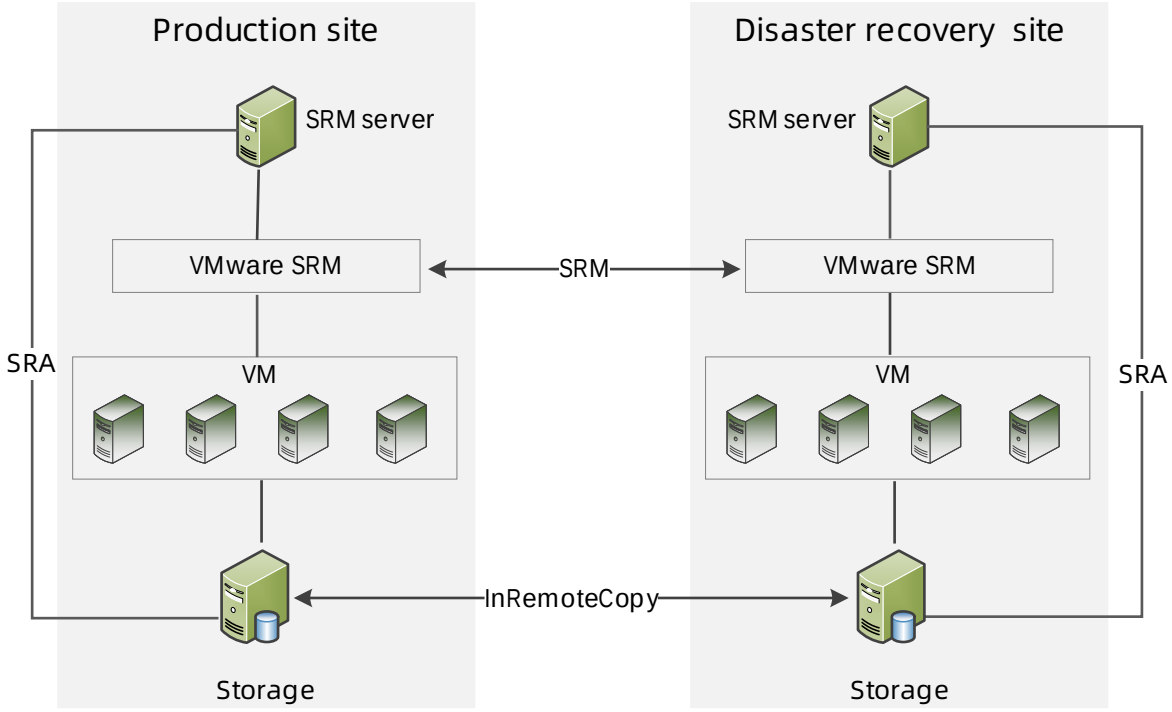
K8sCSI plug-in in controller mode receives tasks from the Kubernetes cluster through sidecar services such as csi-provisioner, csi-attacher, csi-snapshotter, snapshot-controller, and csi-extender, and interacts with the storage to complete management tasks. K8sCSI plug-in in nodeworker mode is registered with the kubelet service of the Kubernetes cluster work node through the sidecar service csi-driver-registrar. Subsequently, the kubelet service directly calls the K8sCSI plug-in in nodeworker mode to complete the mounting and unmounting of volumes on work node. To mount a device, K8sCSI plug-in in nodeworker mode will map volumes in the storage system based on the host information and scan the mounted device on the host. To unmount a device, K8sCSI plug-in in nodeworker mode will unmount and delete the device and delete the volume mapping from the storage system.

5.6 SRA plug-in

With the rise of big data, virtualization applications are becoming increasingly widespread, and the requirements for data security, business continuity, and high availability of virtualization applications are higher. KAYTUS storage has passed VMware certification and has engaged in deep cooperation with VMware. The storage devices can be connected seamlessly when customers deploy virtualization solutions.

SRA plug-in developed by KAYTUS storage is a Storage Replication Adapter (SRA) for VMware Site Copy Manager (SRM), can enable SRM to call its InRemoteCopy feature, achieving disaster recovery quickly and safely in case of a disaster and making sure of business continuity and data security. For example, install VMware SRM and SRA plug-in for the production site and the disaster recovery site to make sure of data consistency through data copy between the clustered systems, as shown in Figure 5-9.

Figure 5-9 SRA solution



In an environment with mature VMware SRM virtualization disaster recovery, integrating KAYTUS storage through SRA plug-in can make sure of data security, integrity, and business continuity. With SRA plug-in installed on a SRM server, SRM may call SRA plug-in to discover storage arrays and configure copy mappings between the production site and the disaster recovery site. In case of unavoidable disasters such as typhoon, earthquake, rainstorm, and power failure, SRM enables business to switch to the disaster recovery site to make sure of business continuity and data security.

SRA plug-in supports pre-configured mode and non pre-configured mode.

- Pre-configured: It indicates that the copy mapping must be configured in advance. When this configuration mode is selected through SRA Utility, it is necessary to configure a copy mapping in advance between the production site and the disaster recovery site.
- Non pre-configured: It is the default mode, indicating that configuring copy mapping in advance is not necessary. When this configuration mode is selected through SRA Utility, it is not necessary to configure a copy mapping in advance between the production site and the disaster recovery site. The copy mapping can be configured through SRA plug-in, related management permission is necessary.

SRA plug-in in combination with VMware SRM can complete the works as shown below.

- Test failover: Test the disaster recovery function to do a check of its effectiveness.
- Failover: Do disaster recovery. In case of a failure, switch to the disaster recovery site through the failover function of SRM and restore the business. There are two modes for disaster recovery, including planned and unplanned.
  - Planned: Do disaster recovery within the initial plan.
  - Unplanned: Do disaster recovery in case of unplanned situations such as natural disasters or sudden failures.

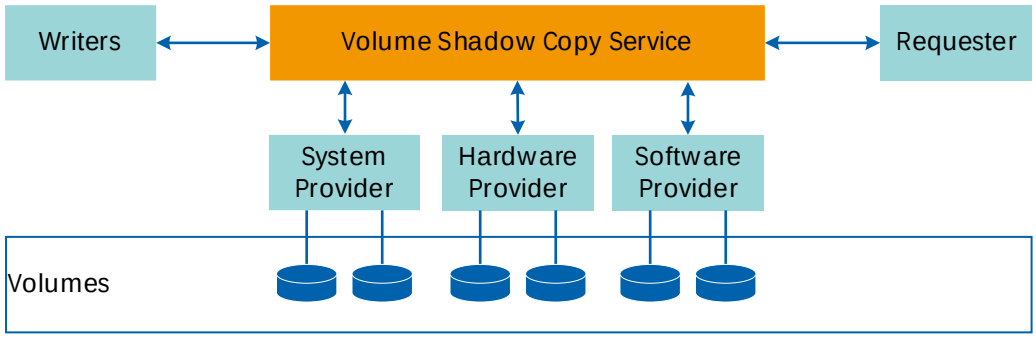
- Re-protection: Do re-protection of the production site. When the production site is restored after failure, protect the production site again and create a relationship between the production site and the disaster recovery site.

5.7 VSS plug-in

Microsoft Volume Shadow Copy Service (VSS) was initially introduced as a storage technology in Windows Server 2003. Through Microsoft VSS, point-in-time based images can be created and the data state at a certain point-in-time can be restored, enabling fast backup and recovery of data and preventing problems caused by accidental data deletion. When all the components support VSS services, online backup of application data can also be completed through VSS without taking the applications offline.

A complete VSS solution requires four components, including VSS service, VSS requester, VSS writer, and VSS provider, as shown in Figure 5-10.

Figure 5-10 VSS solution



- VSS service

  Part of the Windows operating system that makes sure that the other components can communicate with each other properly and work together.

- VSS requester

  The software that requests the actual creation of shadow copies. Typically, this is the backup application. The Windows Server Backup utility and the System Center Data Protection Manager application are VSS requesters. Other backup software that runs on Windows can also be requesters.

- VSS writer

  Typically provided as crucial business application that makes sure of a consistent data set to back up, such as SQL Server or Exchange Server. VSS writers for different Windows components, such as the registry, are included with the Windows operating system. Non-Microsoft VSS writers are included with many applications for Windows to make sure of data consistency during back up.

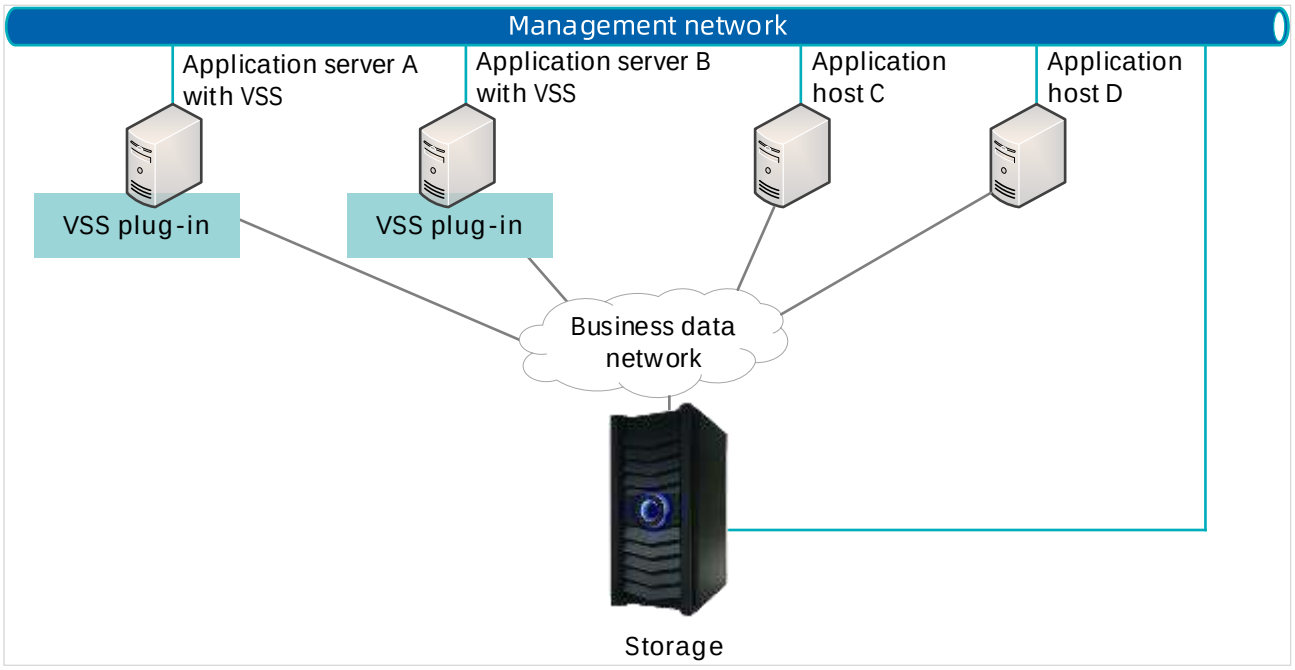
- VSS provider

  The component that creates and maintains the shadow copies, including VSS system provider, VSS software provider, and VSS hardware provider. The Windows operating system includes a VSS provider that uses copy-on-write. VSS hardware provider can be a storage array that provides

storage space in a storage area network (SAN).It is important that you install the VSS hardware provider for a SAN. A hardware provider can offloads the task of creating and maintaining a shadow copy from the host operating system. The storage system can provide InLocalCopy capabilities for VSS Service.

VSS plug-in (hardware provider) developed by KAYTUS storage is installed on the Windows host and runs as a service. The application topology of VSS plug-in is as shown in Figure 5-11. VSS plug-in uses Microsoft VSS framework as its management interface, calling the InLocalCopy feature on the storage-end to create and manage local copy mappings, and storing data in the storage system.

Figure 5-11 Application topology of VSS plug-in



6 Management and maintenance

6.1 Event monitoring and DMP

The unified graphical management interface provided by the storage system enables the configuration, monitoring, and management of storage arrays, and gives feedback on important health status information in time, mainly including complete failure information, user operation information, or process status changes, which makes sure that the storage system works correctly.

Directed Maintenance Procedures (DMP) is a recovery wizard for troubleshooting when software and hardware faults occur in the storage system, providing guidance and suggestions. When faults occur, related events will be reported by the storage system, presenting in InStorageManager by the priority of urgency. To report events and troubleshoot in time, the storage system also provides other notification methods.

- Email: Emails will be sent to one or more email addresses to report events to administrators, which can be seen on mobile, enabling timely response.
- SNMP: SNMP messages will be sent to report events to the data center management system. Administrators can integrate SNMP messages from multiple systems and monitor the data center from a single workstation.
- System logs: System logs will be sent to report events to the data center management system. Administrators can integrate the log messages from multiple systems and monitor the data center from a single workstation.

6.2 System security configuration

KAYTUS storage provides multiple security measures to make sure of data integrity, data confidentiality, and data availability, preventing unauthorized users from getting access to storage resources and data. The security architecture includes device security, network security, business security, and management security, obeying the principles of deep defense, security domain, and elastic design. This section mainly introduces configurations related to authentication, access control, log management, security policies, and security reinforcements.

Authentication

KAYTUS storage supports configuring and managing the storage system through InStorageManager and CLI.

- InStorageManager access authentication

InStorageManager supports two authentication methods to authenticate access permission to the storage system, including user name and password, and user name and email verification code. The client is connected to the storage system safely using the HTTPS protocol, providing a reliable communication channel for the storage system and its maintenance terminal. Additionally, InStorageManager also supports two-factor authentication, which adds a software certificate authentication based on user name and password to increase the system security.
Session principles such as verification code, user lock, and logout upon timeout are also used to make sure of system security.

- CLI access authentication

  CLI supports two authentication methods, including user name and password, and key pair. Data is encrypted and transmitted between the storage system and maintenance terminal using the SSH protocol, preventing domain name system (DNS) spoofing and IP address spoofing. Through data encryption, the SSH protocol can make sure of the security of data transmission between the storage system and maintenance terminal.
  To make sure of session security, the default timeout for SSH sessions is two minutes and cannot be changed. When a SSH session becomes timeout, the SSH connection is automatically closed.

- CHAP authentication

  For SAN business, the storage system supports CHAP authentication to restrict the access of application servers to it, making sure of its security. After CHAP is configured for the storage system and CHAP authentication is enabled for initiators, the application servers (initiators) will send an iSCSI connection request to the storage system (target). The storage system will authenticate the CHAP user name and password. Two-way CHAP authentication is recommended to get higher system security.

Access control

To make sure of the security of storage devices and business data, the storage system provide multiple methods to control the access to management resources and business resources.

- Black list and white list: The black list and white list are used to control the IP addresses that can get access to the storage system.
- Permission control: To prevent mis-operations from making unwanted effect on system stability and data security, the storage system supports role-based access control, which can control user permissions through roles. There are two types of roles in the storage system, including system roles and user-defined roles.
  - System roles: The storage system provides multiple system roles, which are built-in roles with specified permissions. The permissions of system roles cannot be changed.
  - User-defined roles: The storage system provides a user-defined role, whose permissions of related business objects can be specified.
- Lock user and unlock user: Security administrator can prevent an account from logging in to the storage system through locking it manually. To restore its permissions, security administrator can

unlock the account manually.

- Important operation authentication: To prevent mis-operations from making unwanted effect on system stability and data security, the storage system provides a secondary authentication for important operations. Before doing important operations such as powering off system, the password must be entered again for confirmation.

Log management

KAYTUS storage provides log download function and log dump function for tracking the overall information of operations, processes, and events related to security transactions and making an analysis.

Security policies

The storage system provide security policies such as password policy and login policy to improve system security.

- The password must include numbers, uppercase letters, lowercase letters, and special characters. The password complexity policy can be set for each account, including minimum length, and minimum number of uppercase letters, lowercase letters, and numbers, improving the security level of important accounts and preventing brute force cracking.
- Support setting the validity period, repeated times, and retry times of passwords. The password must be changed on the first login, and it must meet the complexity requirements.
- Support setting the lock duration and default logout time

Security reinforcements

The storage system supports the security reinforcements of device physical ports, firmware, operating systems, web servers, and user information, eliminating or decreasing potential security risks of the storage system.

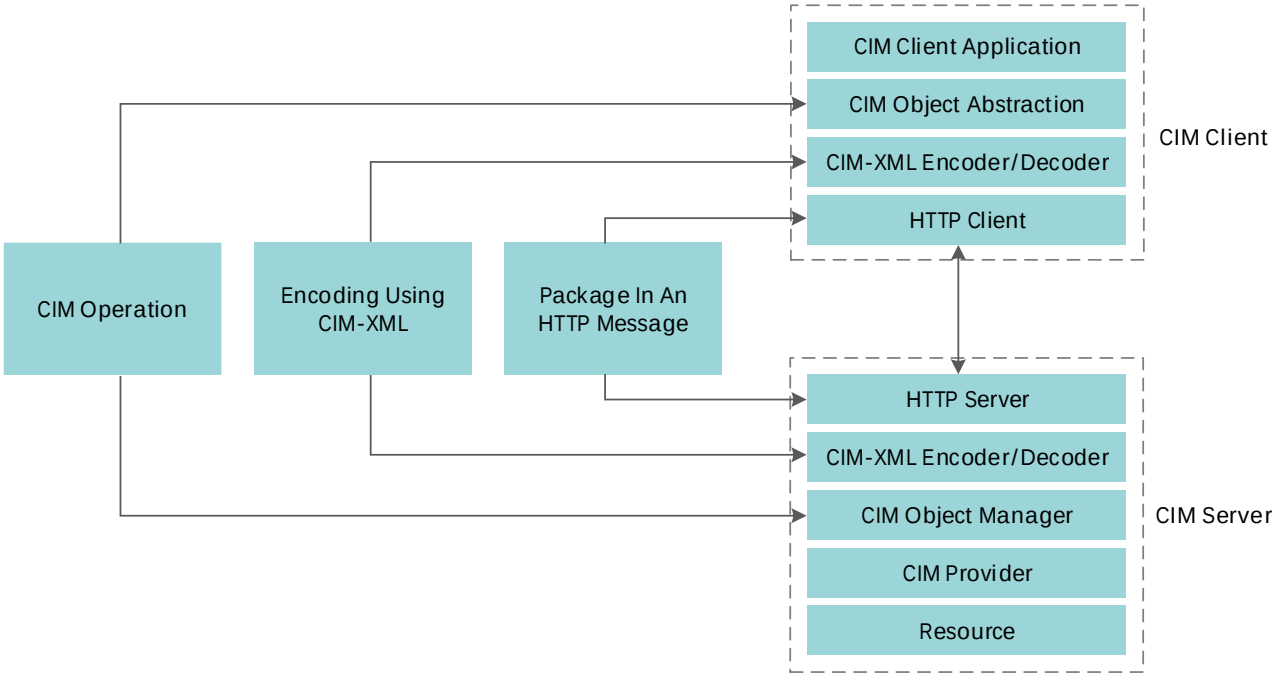
6.3 CIM interface

Common Information Model (CIM) is an open source standard proposed by DMTF. It defines how the managed objects and the relationships between the managed objects are represented in an object-oriented manner in the IT environment. Its purpose is to keep consistency in system management regardless of manufacturers and providers. Storage Management Initiative Specification (SMI-S) is a CIM standard management interface initiated and led by SNIA, with the participation of numerous storage manufacturers.

KAYTUS storage provides a CIM interface that complies with the SMI-S specifications and has been officially certified by SNIA. The functions of CIM interface include cluster management, host management, storage resource management (pools, volumes, and drives), performance monitoring, and data protection. CIM client monitors and manages the storage system by calling the CIM interface provided by

the storage system.

Figure 6-1 CIM architecture



CIM Client initiates a CIM operation request, and receives and processes the returned results of the CIM operation. CIM Server receives and processes the request of CIM operations, and returns the results for the CIM operations to the CIM Client.

There is a CIM Object Abstraction Layer between the CIM Client and CIM Server, which is responsible for implementing CIM objects and CIM operations without specific protocol dependencies. On the client-end, the layer is represented as CIM Object Abstraction, and on the server-end, it is represented as CIM Object Manager, also known as CIMOM. CIMOM is a database for all CIM-class instances, and all implemented classes are shown in it, which is the key to getting access to resources.

6.4 SNMP interface

Simple Network Management Protocol (SNMP) is a network management standard based on the TCP/IP protocol family, which is a standard protocol for managing network nodes (such as servers, workstations, routers, and switches) in an IP network.

KAYTUS storage provides a customized interface that complies with the SNMP specifications to be called by third-party systems or software to manage or monitor storage arrays. When the storage system generates an event, it will actively report the SNMP Trap message containing the event (including the alarm level and event content). SNMP GET/SET supports versions of v1, v2c, and v3, and Trap supports versions of v2c and v3.

The device information and status of storage system as shown below can be got through SNMP interface.

- System status such as CPU utilization rate.

- Capacity information about drives, MDisks, and storage pools.
- Component information about BBUs, fans, PSUs, and controllers.
- Port status of iSCSI ports, SAS ports, and FC ports.
- Drive information and utilization statistics.
- IP configuration, such as IP address, gateway, and mask code.

The IP, name, and time of the storage system can be set through SNMP Interface.

6.5 RESTful interface

Representational State Transfer (REST) refers to a set of architectural constraints and principles. An application or design that meets these constraints and principles is RESTful. RESTful has the features as shown below.

- Each URI represents one type of resource.
- The client-end mainly uses four methods to do operations on the resources on server-end, including GET, POST, PUT, and DELETE.
- The operation on resources is done through their representation.
- The representation of resources is XML or HTML.
- The interaction between the client-end and the server-end has no state between requests, and each request from the client-end to the server-end must contain the necessary information for understanding the request.

KAYTUS storage provides an interface that complies with REST constraints to be called by third-party systems or software to manage or monitor storage arrays. The storage resources can be operated through RESTful interface.

- Security: Create, change, query, and delete users.
- Device: Query the information about enclosures, controllers, drives, BBUs, PSUs, fans, and ports.
- Clustered system: Get the information about nodes and systems.
- Event: Get the information about system events.
- Block storage: Get the information about Mdisks, volumes, storage pools, and hosts.

7 Acronyms

B	
BBU	Battery Backup Unit
C	
CIFS	Common Internet File Systems
CLI	Command Line Interface
CPU	Central Processing Unit
D	
DMP	Directed Maintenance Procedures
DNS	Domain Name System
DRAID	Distributed RAID
F	
FC	Fibre Channel
FTP	File Transfer Protocol
G	
GUI	Graphical User Interface
H	
HDD	Hard Disk Drive
I	
I/O	Input/Output
iSCSI	Internet Small Computer System Interface
L	
LBA	Logical Block Address
M	

MPIO	Multipath I/O
N	
NFS	Network File System
NL-SAS	Near Line SAS
NQN	NVMe Qualified Name
P	
PBA	Physics Block Address
PSU	Power Supply Unit
R	
RACE	Random Access Compression Engine
RAID	Redundant Arrays of Independent Disks
S	
SAN	Storage Area Network
SAS	Serial Attached SCSI
SCSI	Small Computer System Interface
SMI-S	Storage Management Initiative specification
SNMP	Simple Network Management Protocol
SSD	Solid State Drive
SSH	Secure Shell
T	
TCO	Total Cost Ownership