

Huawei FusionCube HCI 3.2 Technical White Paper

Issue 01
Date 2019-04-15



Copyright © Huawei Technologies Co., Ltd. 2019. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions



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

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

Notice

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

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

Huawei Technologies Co., Ltd.

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

Website: <http://e.huawei.com>

About This Document

Overview

This document introduces FusionCube Hyper-converged Virtualization Infrastructure 3.2 (FusionCube HCI 3.2) from the following perspectives: benefits, architecture, performance, scalability, security, and reliability.

You can obtain comprehensive information about FusionCube by reading this document.





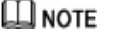
Intended Audience

This document is intended for:

- Marketing engineers
- Technical support engineers
- Maintenance engineers

Symbol Conventions

The following table lists the symbols that may be found in this document.

Symbol	Description
	Indicates an imminently hazardous situation which, if not avoided, will result in death or serious injury.
	Indicates a potentially hazardous situation which, if not avoided, could result in death or serious injury.
	Indicates a potentially hazardous situation which, if not avoided, could result in minor or moderate injury.
	Indicates a potentially hazardous situation which, if not avoided, could result in equipment damage, data loss, performance deterioration, or unanticipated results. NOTICE is used to address practices not related to personal injury.
	Calls attention to important information, best practices, and tips.

Symbol	Description
	NOTE is used to address information not related to personal injury, equipment damage, and environment deterioration.

Change History

Issue	Date	Description
01	2019-04-15	This issue is the first official release.

Contents

About This Document	ii
1 Product Overview	1
2 Benefits	2
3 Product Architecture	4
3.1 Architecture of the FusionSphere Scenario.....	5
3.1.1 Architecture.....	5
3.1.2 Typical Configuration	6
3.1.3 Networking	10
3.1.4 Working Principles.....	10
3.2 Architecture of the VMware Scenario	11
3.2.1 Architecture.....	12
3.2.2 Typical Configuration	12
3.2.3 Networking	16
3.2.4 Working Principles.....	16
4 Distributed Storage	19
4.1 Architecture Overview	20
4.2 Key Service Processes	22
4.2.1 Data Routing.....	22
4.2.2 I/O Paths	24
4.2.3 Cache Mechanisms	26
4.3 Storage Management	28
4.3.1 Storage Cluster Management.....	28
4.3.2 Storage as a Service	29
4.4 Data Redundancy.....	29
4.4.1 Multi-Copy	30
4.4.2 Erasure Code.....	30
4.5 Features.....	32
4.5.1 SCSI/iSCSI Block Interface.....	32
4.5.2 Thin Provisioning	34
4.5.3 Snapshot.....	34
4.5.4 Shared Volume Snapshot	35

4.5.5 Consistency Snapshot	36
4.5.6 Linked Clone	37
4.5.7 Multiple Resource Pools.....	38
4.5.8 QoS	39
4.5.9 Active-Active Storage.....	40
4.5.10 Asynchronous Storage Replication	40
5 Hardware Platform	43
5.1 Rack Servers	43
5.1.1 RH1288 V3	43
5.1.2 RH2288H V3	44
5.1.3 RH5885H V3	45
5.1.4 1288H V5.....	46
5.1.5 2288H V5.....	46
5.1.6 2488 V5	48
5.1.7 2488H V5.....	49
5.2 E9000 Blade Server	49
5.2.1 E9000 Chassis.....	49
5.2.2 E9000 Compute Nodes	50
5.2.3 High-Performance Switch Modules.....	56
5.3 X6800 and X6000 High-Density Servers	58
5.3.1 X6800 Chassis	58
5.3.2 X6800 Server Nodes.....	60
5.3.3 X6000 Chassis	64
5.3.4 X6000 Server Nodes.....	64
6 Installation, Deployment, and O&M	67
6.1 Automatic Deployment	67
6.1.1 FusionCube Builder	67
6.1.2 System Initialization	69
6.1.3 Automatic Device Discovery	70
6.2 Unified O&M	71
6.2.1 Service Provisioning and Management	73
6.2.2 One-Click O&M	74
6.2.3 Call Home.....	78
7 Performance and Scalability	80
7.1 High Performance	80
7.1.1 Distributed I/O Ring	80
7.1.2 Distributed SSD Cache Acceleration.....	81
7.1.2.1 Read/Write Cache	82
7.1.2.2 Pass-Through of Large Blocks.....	84
7.1.3 Hardware Acceleration	85
7.2 Linear Expansion.....	86

7.2.1 Smooth Storage Capacity Expansion.....	86
7.2.2 Linear Performance Expansion.....	87
7.2.3 One-Click Capacity Expansion.....	88
7.3 Advantages of FusionCube Distributed Storage over Conventional SAN.....	89
7.3.1 Higher performance.....	89
7.3.2 Linear Scale-up/Scale-out.....	90
7.3.3 Large Pool.....	92
7.3.4 SSD Cache vs SSD Tier.....	93
8 System Reliability.....	95
8.1 Data Reliability.....	95
8.1.1 Block Storage Cluster Reliability.....	95
8.1.2 Data Consistency.....	96
8.1.3 Data Redundancy.....	97
8.1.4 Rapid Data Rebuild.....	97
8.1.5 Multipathing for Data Storage.....	98
8.2 Hardware Reliability.....	98
8.3 Sub-Health Enhancement.....	98
8.4 Backup and Restoration.....	102
8.5 DR.....	104
8.5.1 Active-Active DR Solution.....	105
8.5.2 Asynchronous Replication Solution.....	106
9 System Security.....	107
9.1 System Security Threats.....	107
9.2 Overall Security Framework.....	108
9.2.1 Network Security.....	109
9.2.2 Application Security.....	110
9.2.2.1 Rights Management.....	110
9.2.2.2 Web Security.....	110
9.2.2.3 Database Hardening.....	111
9.2.2.4 Log Management.....	112
9.2.3 Host Security.....	112
9.2.3.1 OS Security Hardening.....	112
9.2.4 Data Security.....	112

1 Product Overview

With the rise of data and Internet services, new services are growing rapidly, resulting in an exponential increase of service data. The traditional server+storage architecture can no longer meet service development requirements. Distributed and cloud-based technologies emerge. An increasing number of enterprises use virtualization and cloud computing technologies to build their IT systems to improve the IT system resource utilization and shorten the time to market (TTM). However, the enterprises are facing the following challenges:

- Complex deployment and management of virtualization platforms and soaring operation and maintenance (O&M) costs
- High requirements for system planning, deployment, and optimization skills because hardware devices may be provided by different vendors
- Slow response from multiple vendors to troubleshooting and technical support requests
- Difficulty in maintenance of a huge system consisting of hardware devices and virtualization platforms from different vendors

In addition, customers now attach more importance to cost control, rapid service provisioning, and risk management. They want resource-scalable, reliable, and high-performance IT systems with low total cost of ownership (TCO) and short TTM.

To help customers address these concerns, Huawei rolls out the FusionCube HCI, which is an open, scalable system that boasts the following outstanding features:

- Convergence of compute, storage, and network resources
- Preintegration and automatic and quick service deployment
- High performance, reliability, and security
- Unified management, intelligent resource scaling, and easy O&M

The FusionCube HCI helps customers quickly deploy services and cloud applications while significantly simplifying maintenance and management.

2 Benefits

As a flagship product of the Huawei IT product line, Huawei FusionCube HCI complies with open architecture and standards. It integrates servers, distributed storage, and network switches in an out-of-the-box packaging without the need of external storage devices. It is preintegrated with distributed storage engines, virtualization platform, and management software and supports on-demand resource allocation and linear expansion. The FusionCube HCI provides the following benefits:

Convergence

FusionCube integrates computing, storage, and network resources.

- **Hardware convergence:** Computing, storage, and network resources are integrated and supports linear expansion.
- **Management convergence:** Centralized O&M significantly improves resource utilization and reduces operating expenses (OPEX).
- **Application convergence:** Hardware and software are optimized based on application service models to improve system performance.

Simplicity

FusionCube implements system preintegration, preinstallation, and preverification, automatic device discovery upon power-on, and unified O&M, greatly simplifying service delivery.

- **Simplified installation:** FusionCube has hardware and software preinstalled before delivery. It can be used on site with simple plug and play.
- **Rapid deployment:** Upon system power-on, FusionCube automatically discovers devices and configures parameters, making service rollout more efficient.
- **Easy O&M:** FusionCube provides a unified management GUI and supports automatic fault locating, making routine O&M simpler.

Optimization

FusionCube uses industry-leading hardware and distributed storage software to ensure optimal user experience.

- **Storage optimization:** FusionCube uses built-in distributed storage to provide storage services with high concurrency and throughput.

Open

FusionCube HCI is an open hyper-converged infrastructure platform independent of specific upper-layer applications. It provides computing, storage, and network resources for mainstream virtualization platforms and databases.

- FusionCube supports mainstream virtualization platforms, such as FusionSphere and VMware vSphere.
- A single system supports hybrid deployment of virtualization and database applications.

3 Product Architecture

The overall architecture of Huawei FusionCube HCI consists of the hardware platform, distributed storage software, installation and deployment and O&M management platforms, virtualization platforms, and backup and disaster recovery (DR) solutions. The virtualization platforms include the Huawei-developed FusionSphere virtualization platform and VMware virtualization platform. In addition, in the FusionSphere scenario, FusionCube HCI supports the hybrid deployment solution. In addition to the FusionSphere virtualization platform, FusionCube HCI supports physical node deployment and provides computing, storage, and network resources for the system database.

The following figure shows the overall architecture of Huawei FusionCube HCI.

Figure 3-1 Huawei FusionCube HCI architecture

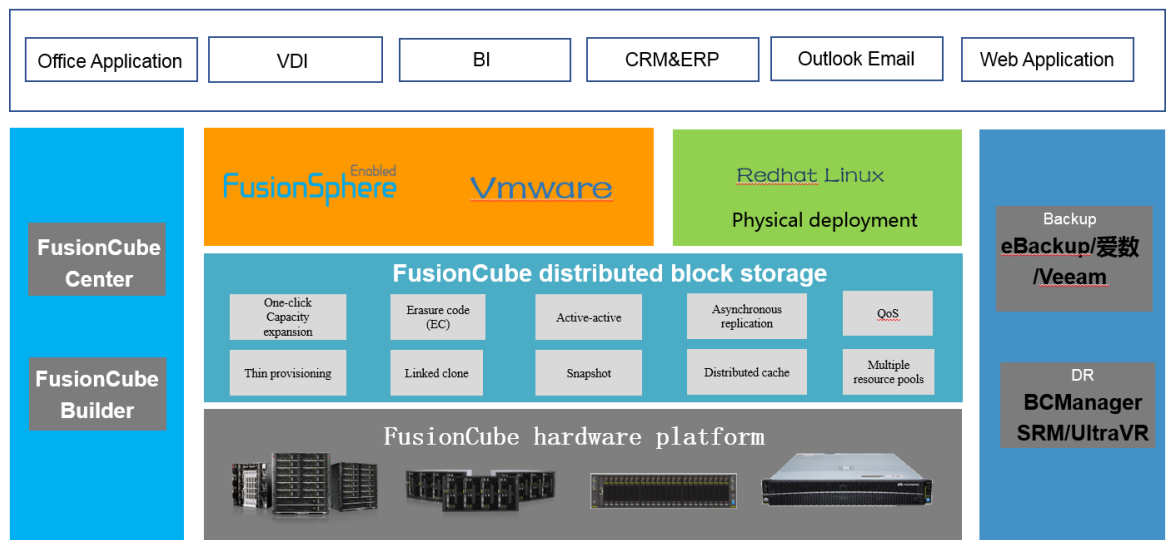


Table 3-1 Huawei FusionCube HCI components

Name	Description
FusionCube Center	Manages FusionCube virtualization and hardware resources, and implements system monitoring and O&M.
FusionCube Builder	Provides quick installation and deployment of FusionCube system software. It can be used to replace or update the

Name	Description
	virtualization platform software.
FusionStorage	Provides high-performance and high-reliability block storage services by using distributed storage technologies to schedule local hard disks on servers in an optimized manner.
Virtualization platform	Implements system virtualization management. The Huawei FusionSphere virtualization platform and VMware virtualization platform are supported.
Backup	Provides the system service virtualization backup function, including the Huawei-developed backup software eBackup and mainstream third-party backup software, such as Veeam, Commvault, and EISOO.
DR	Provides DR solutions based on active-active storage and asynchronous storage replication. The DR software includes Huawei-developed BCManager and UltraVR.
Hardware platform	E9000, X6800, X6000, and rack servers are used. The servers support computing, storage, switch, and power modules and allow on-demand configuration of compute and storage nodes. FusionCube supports I/O expansion for components, such as graphics processing units (GPUs) and peripheral component interconnect express (PCIe) solid-state drives (SSDs), and various switch modules (including 10GE and RoCE) to meet different configuration requirements.

As a flagship product of the Huawei IT product line, Huawei FusionCube HCI complies with open architecture and standards. It integrates servers, distributed storage, and network switches in an out-of-the-box packaging. FusionCube is pre-integrated with distributed storage engines, virtualization platforms, and management software to implement on-demand resource allocation and linear expansion.

3.1 Architecture of the FusionSphere Scenario

3.2 Architecture of the VMware Scenario

3.1 Architecture of the FusionSphere Scenario

In FusionCube HCI 3.2, the FusionSphere virtualization scenario uses the KVM virtualization architecture. The system consists of the Huawei-developed hardware platform, FusionCube distributed storage system, FusionSphere virtualization platform, and FusionCube Center management platform. FusionCube Builder provides corresponding software installation operations. Software including eBackup, Veeam, and UltraVR can provide advanced features such as backup and DR for the system.

3.1.1 Architecture

In the FusionSphere virtualization deployment, the FusionCube distributed storage software is directly deployed in the hypervisor kernel. The HDD and SSD cache storage media of nodes use the FusionCube distributed storage software to construct shared storage pool resources. In

In addition, the FusionSphere virtualization platform virtualizes the computing resources of nodes and then provides the resources to service VMs on the nodes. Based on the functions and features provided by nodes, the nodes are classified into converged management nodes, converged storage nodes, compute nodes, and physical database nodes. The following figure shows the node architecture.

Figure 3-2 Node architecture in the FusionSphere scenario

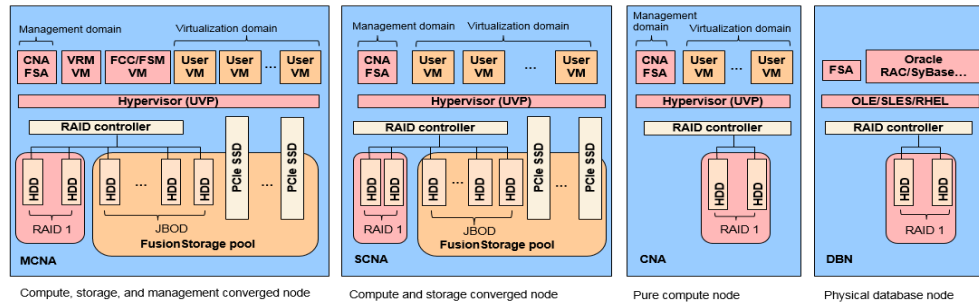


Table 3-2 Description of nodes in the FusionSphere scenario

Name	Description	Deployment Principle
MCNA (management node)	A node that provides management functions. The VRM, FusionCube Center, and FusionStorage Manager management VMs are deployed on the node. It also provides storage and computing functions.	Two MCNAs must be deployed.
SCNA (storage and compute node)	A node that provides storage and computing functions. It provides FusionCube distributed storage HDDs, SSD cache storage resources, and virtualized computing resources.	One or more SCNAs can be deployed based on service requirements.
CNA (compute node)	A node that provides computing functions. It provides only virtualized computing resources.	Zero or multiple CNAs can be deployed based on service requirements.
DBN (database node)	A physical deployment node, which provides computing resources for the system database.	Zero or multiple DBNs can be deployed based on service requirements.

3.1.2 Typical Configuration

In FusionCube HCI 3.2, the FusionSphere virtualization scenario supports the large-capacity HDD+SSD cache hybrid deployment and high-performance all-SSD deployment scenarios. The specific scenario configuration is as follows:

Typical configuration of nodes in the hybrid deployment scenario:

Configuration Item	Typical Configuration	Description
Server type	V5 rack server/E9000 V5 blade/X6800 V5 and X6000 V5 high-density servers	<p>A proper server type needs to be selected based on the customer's cabinet space, disk size, density, and number of PCIe NICs.</p> <p>Rack server: Various types of hard disks are supported. Multiple PCIe slots are reserved. GPUs are supported. However, a lot of space is required.</p> <p>E9000 blade: The computing, storage, and network devices can be integrated in an E9000 cabinet, but only 2.5-inch HDDs and NVMe SSDs are supported. The capacity of a single node is small and the NIC configuration is fixed.</p> <p>X6800 V5 high-density server: The storage and computing density is high. Four nodes are supported in 4U space. Each node supports two system disks and ten 3.5-inch disks. However, the rear PCIe slots are insufficient (two x8 slots), and GPUs are not supported.</p> <p>X6000 V5 high-density server: The computing density is high. Four nodes are supported in 2U space. However, each node supports only six 2.5-inch disks (including system disks). The disk capacity of a single node is small. Only LOMs with two GE ports and two 10GE ports are supported. Only one NVMe SSD card is used as the cache. GPUs are not supported.</p>
CPU and memory configuration	2 Intel® Xeon® Gold 5120 processors 8 x 32 GB DDR4 RDIMMs (2666 MHz)	The CPU and memory configuration can be dynamically adjusted based on the customer's service specifications and configuration to provide more computing resources.
Disk	2 TB/4 TB/6 TB/8 TB SATA disks and 1.2 TB/1.8 TB/2.4 TB SAS disks OS disks: 2 x 600 GB SAS disks (default)	<p>The FusionCube distributed storage requires that SATA disks must use three copies or a redundancy policy whose EC ratio is greater than N+2. SAS disks can use two copies, three copies, or a redundancy policy whose EC ratio is greater than N+2.</p> <p>During the deployment of the FusionCube distributed storage, one disk on three nodes is used to store ZK</p>

Configuration Item	Typical Configuration	Description
		metadata.
Cache	Huawei-developed NVMe SSD V5 disks or cards and Huawei-developed SAS SSD V5 disks	<p>The cache size of the system can be flexibly configured based on the customer's service pressure. Generally, 800 GB NVMe SSD V5 disks or cards are configured by default.</p> <p>In addition to the NVMe SSDs and SAS SSDs developed by Huawei, the cache also supports SAS/SATA SSD disks that support compatibility verification from other vendors, such as Intel, Samsung, and Micron.</p>
NIC	2 x 10GE + 2 x 10GE	By default, it is recommended that the management and service planes share two 10GE network ports and the storage network plane exclusively use two 10GE network ports. If DR is configured, you are advised to add two 10GE network ports for the replication network plane.

Typical configuration of nodes in the all-flash deployment scenario:

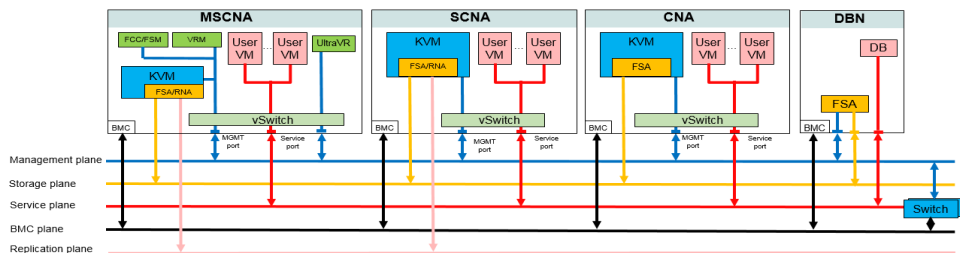
Configuration Item	Typical Configuration	Description
Server type	V5 rack server/E9000 V5 blade/X6800 V5 and X6000 V5 high-density servers	<p>A proper server type needs to be selected based on the customer's cabinet space, disk size, density, and number of PCIe NICs.</p> <p>Rack server: Various types of hard disks are supported. Multiple PCIe slots are reserved. GPUs are supported. However, a lot of space is required.</p> <p>E9000 blade: The computing, storage, and network devices can be integrated in an E9000 cabinet, but only 2.5-inch HDDs and NVMe SSDs are supported. The capacity of a single node is small and the NIC configuration is fixed.</p> <p>X6800 V5 high-density server: The storage and computing density is high. Four nodes are supported in 4U space. Each node supports two system disks and ten 3.5-inch disks. However, the rear PCIe slots are insufficient (two x8</p>

Configuration Item	Typical Configuration	Description
		<p>slots), and GPUs are not supported.</p> <p>X6000 V5 high-density server: The computing density is high. Four nodes are supported in 2U space. However, each node supports only six 2.5-inch disks (including system disks). The disk capacity of a single node is small. Only LOMs with two GE ports and two 10GE ports are supported. Only one NVMe SSD card is used as the cache. GPUs are not supported.</p>
CPU and memory configuration	2 Intel® Xeon® Gold 5120 processors 8 x 32 GB DDR4 RDIMMs (2666 MHz)	The CPU and memory configuration can be dynamically adjusted based on the customer's service specifications and configuration to provide more computing resources.
Disk	ES3000 NVMe SSD V5 disks and ES3000 SAS SSD V5 disks OS disks: 2 x 480 GB SATA SSD disks or 2 x 600 GB SAS disks	<p>By default, the FusionCube distributed storage uses two copies or a redundancy policy whose EC ratio is greater than N+2 in the all-flash scenario. If the customer requires higher reliability, three copies (with higher costs) can be used.</p> <p>By default, Huawei-developed disks are used as all-flash disks. The 3 DWPD disks are recommended. If the volume of data to be written is small, 1 DWPD disks can be used.</p> <p>In the all-flash scenario, ZK metadata of the distributed storage is stored in the OS disk space. To meet ZK metadata performance requirements, it is recommended that 480 GB SATA SSD disks be used as OS disks. If more than five nodes are deployed, 600 GB SAS disks can be used (ZK metadata and management VMs are not on the same node).</p>
NIC	2 x 10GE + 2 x 10GE	By default, it is recommended that the management and service planes share two 10GE network ports and the storage network plane exclusively use two 10GE network ports. If DR is configured, you are advised to add two 10GE network ports for the replication network plane.

3.1.3 Networking

The system networking of Huawei FusionCube HCI 3.2 in the FusionSphere scenario consists of the management plane, storage plane, service plane, BMC plane, and replication and arbitration planes involved in DR. The following figure shows the detailed networking.

Figure 3-3 Networking diagram (FusionSphere)



The communication plane types are described as follows:

Management plane: The management network plane of the FusionCube system is used for service operations and O&M management of the system. It supports the TCP/IP protocol and GE/10GE networking. It can share NICs with the service plane and be isolated by VLANs.

Storage plane: The network plane for data read and write between FusionCube distributed storage nodes supports the TCP/IP protocol, RDMA protocol, and 10GE/IB networking. It is recommended that NICs be exclusively used by the storage plane and network ports in active-backup mode be configured.

Service plane: The customer service communication network plane supports the TCP/IP protocol and GE/10GE networking. It can share NICs with the management plane and be isolated by VLANs.

BMC plane: The server device management IP plane is used to access the O&M management of server devices in the FusionCube system.

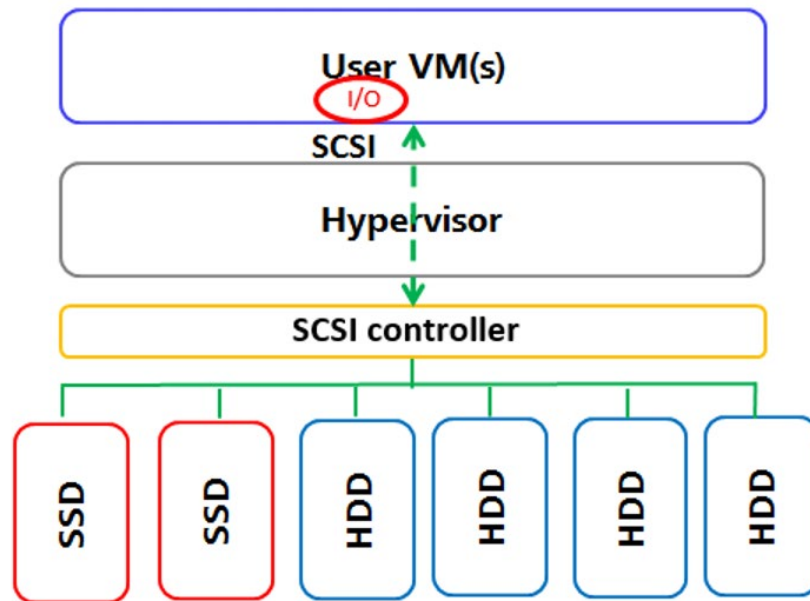
Replication plane: In the DR solution, the data synchronization network plane between the active and standby sites supports the TCP/IP protocol and GE/10GE networking. It is recommended that the system exclusively use NICs to prevent resources from being preempted by other network planes.

3.1.4 Working Principles

Service I/O Process

In the FusionSphere scenario, service VMs are deployed on the host KVM virtualization platform. The storage system uses the FusionCube distributed storage resources. VM I/Os interact with the distributed storage software running in the hypervisor kernel through the SCSI protocol. For details about the service I/O process, see Figure 3-4.

Figure 3-4 Service I/O process in the FusionSphere scenario



In the FusionSphere scenario, the node architecture has a short I/O path and higher efficiency.

Service Management and O&M

In the FusionSphere scenario, the FusionCube HCI system supports system preinstallation and integration. Before delivery, the system hardware BIOS, disk RAID configuration, virtualization platform, management software, and distributed storage software are preinstalled on nodes. After the product is powered on, the customer can complete the system configuration and initialization within dozens of minutes. After the system is initialized, service VMs can be provisioned in the system.

On the FusionCube Center management page, you can create service VMs, manage VM lifecycles, monitor and manage system hardware, storage, computing resources, and VMs, report and manage system component alarms in a unified manner, and implement one-click O&M, including one-click capacity expansion, upgrade, log collection, and health check. Some advanced features or configurations of the virtualization platform, such as computing cluster creation and configuration, DVS creation and configuration, and advanced VM feature configuration, can be switched to the FusionCompute virtualization management platform in SSO mode.

3.2 Architecture of the VMware Scenario

In FusionCube HCI 3.2, the VMware virtualization scenario uses the ESXi virtualization architecture. The system consists of the Huawei-developed hardware platform, FusionCube distributed storage system, ESXi virtualization platform, and FusionCube Center management platform. FusionCube Builder provides corresponding software installation operations. Software including eBackup, Veeam, and BCManager can provide advanced features such as backup and DR for the system.

3.2.1 Architecture

In the VMware virtualization deployment, the FusionCube distributed storage software is deployed on the controller VM (CVM). The HDD and SSD cache storage media of nodes are directly connected to the CVM through the VMDirectPath I/O. Then, the FusionCube distributed storage software is used to construct the HDD and SSD cache storage media as shared storage pool resources. In addition, the ESXi virtualization platform virtualizes computing resources of nodes and provides them for service VMs on the nodes. Based on the functions and features provided by nodes, the nodes are classified into converged management nodes, converged storage nodes, and compute nodes. The following figure shows the node architecture.

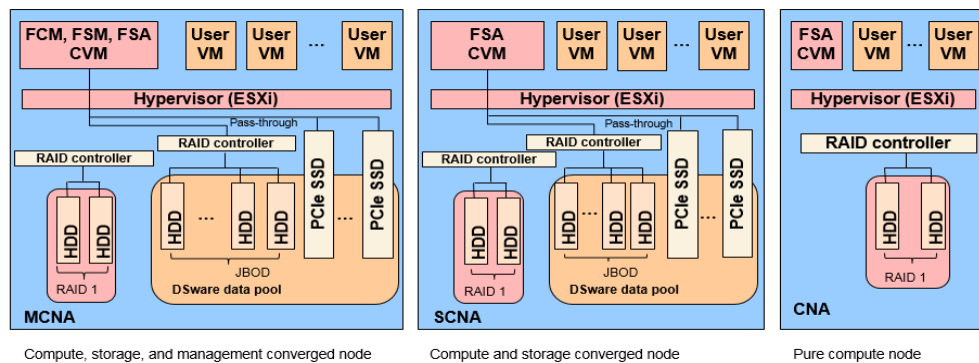


Table 3-3 Description of nodes in the VMware scenario

Name	Description	Deployment Principle
MCNA (management node)	A node that provides management functions. The FusionCube Center and FusionStorage Manager management processes are deployed on the CVM on the node. It also provides storage and computing functions.	Two MCNAs must be deployed.
SCNA (storage and compute node)	A node that provides storage and computing functions. It provides FusionCube distributed storage HDDs, SSD cache storage resources, and virtualized computing resources.	One or more SCNAs can be deployed based on service requirements.
CNA (compute node)	A node that provides computing functions. It provides only virtualized computing resources.	Zero or multiple CNAs can be deployed based on service requirements.

3.2.2 Typical Configuration

In FusionCube HCI 3.2, the VMware virtualization scenario supports the large-capacity HDD+SSD cache hybrid deployment and high-performance all-SSD deployment scenarios. The specific scenario configuration is as follows:

Typical configuration of nodes in the hybrid deployment scenario:

Configuration Item	Typical Configuration	Description
Server type	V5 rack server/X6800 V5 high-density server	<p>A proper server type needs to be selected based on the customer's cabinet space, disk size, density, and number of PCIe NICs.</p> <p>Rack server: Various types of hard disks are supported. Multiple PCIe slots are reserved. GPUs are supported. Two RAID controller cards are required for the 2288H V5, and M.2 cards are required for RAID group configuration on the 1288H V5. However, a lot of space is required.</p> <p>X6800 V5 high-density server: The storage and computing density is high. Four nodes are supported in 4U space. Each node supports two system disks and ten 3.5-inch disks. The front PCIe slots are configured with SSD cache and M.2 cards. However, the rear PCIe slots are insufficient (two x8 slots), and GPUs are not supported.</p>
CPU and memory configuration	2 Intel® Xeon® Gold 5120 processors 8 x 32 GB DDR4 RDIMMs (2666 MHz)	The CPU and memory configuration can be dynamically adjusted based on the customer's service specifications and configuration to provide more computing resources.
Disk	2 TB/4 TB/6 TB/8 TB SATA disks and 1.2 TB/1.8 TB/2.4 TB SAS disks OS disks: 2 x 600 GB SAS disks (default)	<p>The FusionCube distributed storage requires that SATA disks must use three copies or a redundancy policy whose EC ratio is greater than N+2. SAS disks can use two copies, three copies, or a redundancy policy whose EC ratio is greater than N+2.</p> <p>During the deployment of the FusionCube distributed storage, one disk on three nodes is used to store ZK metadata.</p> <p>Hard M.2 RAID groups need to be configured for storage models other than the 2288H V5 to install and deploy ESXi.</p>
Cache	Huawei-developed NVMe SSD V5 disks or cards and Huawei-developed SAS SSD V5 disks	The cache size of the system can be flexibly configured based on the customer's service pressure. Generally, 800 GB NVMe SSD V5 disks or cards are configured by default.

Configuration Item	Typical Configuration	Description
		In addition to the NVMe SSDs and SAS SSDs developed by Huawei, the cache also supports SAS/SATA SSD disks that support compatibility verification from other vendors, such as Intel, Samsung, and Micron.
NIC	2 x 10GE + 2 x 10GE	By default, it is recommended that the management and service planes share two 10GE network ports and the storage network plane exclusively use two 10GE network ports. If DR is configured, you are advised to add two 10GE network ports for the replication network plane.

Typical configuration of nodes in the all-flash deployment scenario:

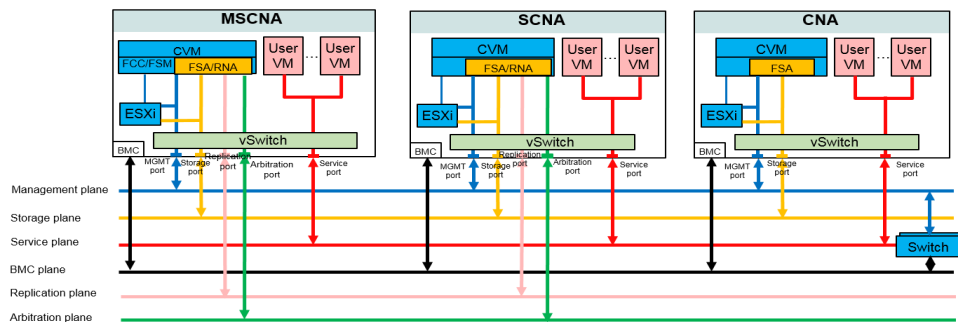
Configuration Item	Typical Configuration	Description
Server type	V5 rack server/E9000 V5 blade/X6800 V5 and X6000 V5 high-density servers	<p>A proper server type needs to be selected based on the customer's cabinet space, disk size, density, and number of PCIe NICs.</p> <p>Rack server: Various types of hard disks are supported. Multiple PCIe slots are reserved. GPUs are supported. Two RAID controller cards are required for the 2288H V5, and M.2 cards are required for RAID group configuration on the 1288H V5. However, a lot of space is required.</p> <p>E9000 blade: The computing, storage, and network devices can be integrated in an E9000 cabinet, but only NVMe SSDs are supported. The NIC configuration is fixed.</p> <p>X6800 V5 high-density server: The storage and computing density is high. Four nodes are supported in 4U space. Each node supports two system disks and ten 2.5-inch SAS/SATA SSD disks. The front PCIe slots are configured with SSD cache and M.2 cards. However, the rear PCIe slots are insufficient (two x8 slots), and GPUs are not supported.</p> <p>X6000 V5 high-density server: The computing density is high. Four nodes</p>

Configuration Item	Typical Configuration	Description
		<p>are supported in 2U space. However, the primary storage of each node supports only NVMe SSD disks. The disk capacity of a single node is small. Only LOMs with two GE ports and two 10GE ports are supported. Only one NVMe SSD card is used as the cache. GPUs are not supported.</p>
CPU and memory configuration	2 Intel® Xeon® Gold 5120 processors 8 x 32 GB DDR4 RDIMMs (2666 MHz)	The CPU and memory configuration can be dynamically adjusted based on the customer's service specifications and configuration to provide more computing resources.
Disk	ES3000 NVMe SSD V5 disks and ES3000 SAS SSD V5 disks OS disks: 2 x 480 GB SATA SSD disks or 2 x 600 GB SAS disks	<p>By default, the FusionCube distributed storage uses two copies or a redundancy policy whose EC ratio is greater than N+2 in the all-flash scenario. If the customer requires higher reliability, three copies (with higher costs) can be used.</p> <p>By default, Huawei-developed disks are used as all-flash disks. The 3 DWPD disks are recommended. If the volume of data to be written is small, 1 DWPD disks can be used.</p> <p>In the all-flash scenario, ZK metadata of the distributed storage is stored in the OS disk space. To meet ZK metadata performance requirements, it is recommended that 480 GB SATA SSD disks be used as OS disks. If more than five nodes are deployed, 600 GB SAS disks can be used (ZK metadata and management VMs are not on the same node).</p>
NIC	2 x 10GE + 2 x 10GE	By default, it is recommended that the management and service planes share two 10GE network ports and the storage network plane exclusively use two 10GE network ports. If DR is configured, you are advised to add two 10GE network ports for the replication network plane.

3.2.3 Networking

The system networking of Huawei FusionCube HCI 3.2 in the VMware scenario consists of the management plane, storage plane, service plane, BMC plane, and replication and arbitration planes involved in DR. The following figure shows the detailed networking.

Figure 3-5 Networking diagram (VMware)



The communication plane types are described as follows:

Management plane: The management network plane of the FusionCube system is used for service operations and O&M management of the system. It supports the TCP/IP protocol and GE/10GE networking. It can share NICs with the service plane and be isolated by VLANs.

Storage plane: The network plane for data read and write between FusionCube distributed storage nodes supports the TCP/IP protocol, RDMA protocol, and 10GE/IB networking. It is recommended that NICs be exclusively used by the storage plane and network ports in active-backup mode be configured.

Service plane: The customer service communication network plane supports the TCP/IP protocol and GE/10GE networking. It can share NICs with the management plane and be isolated by VLANs.

BMC plane: The server device management IP plane is used to access the O&M management of server devices in the FusionCube system.

Replication plane: In the DR solution, the data synchronization network plane between the active and standby sites supports the TCP/IP protocol and GE/10GE networking. It is recommended that the system exclusively use NICs to prevent resources from being preempted by other network planes.

Arbitration plane: In the DR active-active solution, the network communication plane between the service site and the arbitration site supports the TCP/IP protocol and GE/10GE networking.

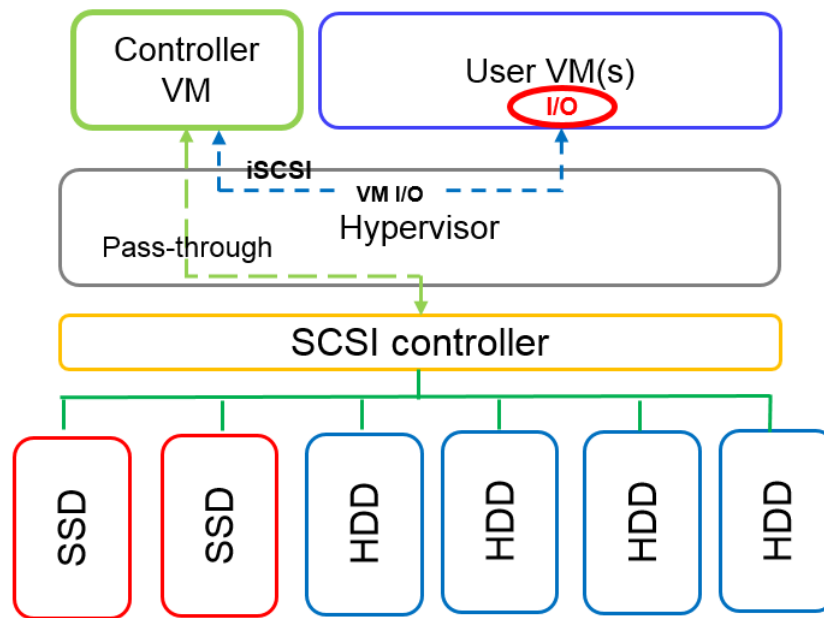
3.2.4 Working Principles

Service I/O Process

In the VMware scenario, the FusionCube distributed storage software is deployed on the CVM. The HDD and SSD cache storage media of nodes are directly connected to the CVM through the VMDirectPath I/O. The FusionCube distributed storage software is used to

construct the HDD and SSD cache storage media as shared storage pool resources. The storage mounts volume devices to hosts using the iSCSI protocol for service VMs to use. The following figure shows the service I/O process.

Figure 3-6 Service I/O process in the VMware scenario



In the VMware scenario, the distributed storage software runs on the CVM, and the I/O path is longer.

Service Management and O&M

In the VMware scenario, the system hardware BIOS and disk RAID have been configured before delivery of the FusionCube HCI system. After receiving and powering on the product, the customer can use the FusionCube Builder installation tool to quickly complete installation and deployment parameter settings and install software such as the ESXi, CVM OS, FusionCube Center management platform, and distributed storage software. After the system is installed and deployed, the customer configures the system on the FusionCube Center management platform and initializes the system. The system needs to connect to vCenter for management. If the customer already has vCenter, the customer can directly connect the system to vCenter. If the customer does not have vCenter, the customer can deploy the vCenter management VM on the system and connect the system to vCenter.

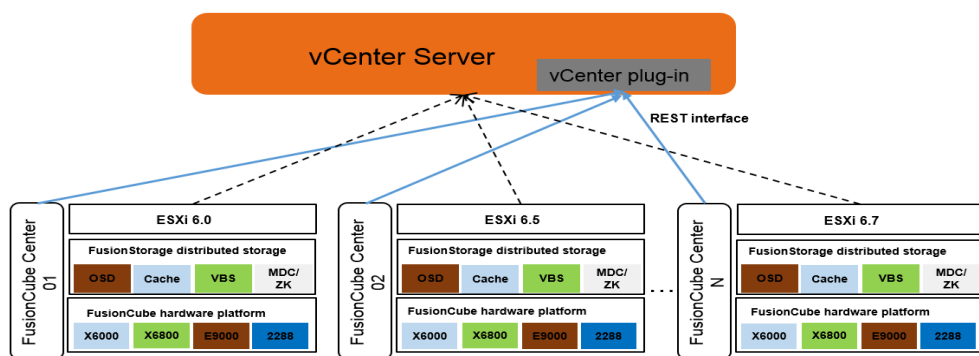
During routine service management and O&M, customer service management and provisioning are performed on vCenter. The customer performs operations related to the virtualization platform on vCenter Server, such as VM and host cluster creation and management.

The O&M and management of the system are performed on FusionCube Center. The system summarizes and reports alarms of hardware devices and storage devices for management. These alarms do not include error alarms of the ESXi virtualization platform. The alarms of the ESXi virtualization platform are managed by vCenter Server. FusionCube Center provides

one-click O&M capabilities, including one-click capacity expansion, upgrade, log collection, and health check.

vCenter Plug-in

To implement management and operation of the FusionCube system on vCenter, FusionCube provides the vCenter plug-in feature. After the plug-in component of FusionCube is deployed on vCenter, FusionCube can be connected to the vCenter plug-in by using the management IP address, user name, and password of FusionCube Center. In this way, vCenter Server can monitor and manage the hardware and storage resources of the FusionCube system, report alarms, create management volume devices, and create VM snapshots. In addition, the vCenter plug-in of FusionCube can connect to multiple sets of FusionCube at the same time. It supports vCenter Server of different versions, including 6.0, 6.5, and 6.7. The following figure shows the detailed framework.



4 Distributed Storage

The built-in distributed storage provides storage services for FusionCube. The FusionCube distributed storage provides block storage devices and uses an innovative cache algorithm and adaptive data distribution algorithm based on a unique parallel architecture, which eliminates high data concentration and improves system performance. It also allows rapid automatic self-recovery and ensures high system availability and reliability.

- **Linear scalability and elasticity**

The FusionCube distributed storage uses the distributed hash table (DHT) to distribute all metadata among multiple nodes. This mode prevents performance bottlenecks and allows linear expansion. The FusionCube distributed storage leverages innovative data slicing technology and DHT-based data routing algorithm to evenly distribute volume data to fault domains of large resource pools. This allows load balancing on hardware devices and higher input/output operations per second (IOPS) and megabit per second (MBPS) performance of each volume.
- **High performance**

The FusionCube distributed storage uses a lock-free scheduled I/O software subsystem to prevent conflicts of distributed locks. I/O paths are shortened and the latency is reduced as there is no lock operation or metadata query on I/O paths. Distributed stateless engines make hardware nodes to be fully utilized, greatly increasing the concurrent IOPS and MBPS of the system. In addition, the distributed SSD cache technology of the FusionCube distributed storage can be used together with large-capacity SAS or SATA disks (serving as the primary storage) to ensure high performance and large storage capacity.
- **High reliability**

The FusionCube distributed storage supports multi-copy (two-copy and three-copy) data backup and erasure code (EC) to protect data. Users can configure flexible data storage policies to ensure data reliability. For example, data copies can be stored on different servers. Data will not be lost and can still be accessed even if a server is faulty. The FusionCube distributed storage also protects valid data slices against loss. If a hard disk or server is faulty, valid data can be rebuilt concurrently. It takes less than 30 minutes to rebuild data of 1 TB. All these measures improve system reliability.
- **Rich advanced storage functions**
 - The thin provisioning function provides users with more virtual storage resources than physical storage resources. Physical storage space is allocated to a volume only when data is written into the volume.

- The volume snapshot function saves the state of the data on a logical volume at a certain time point. The number of snapshots is not limited, and performance does not deteriorate.
- The linked clone function is implemented based on incremental snapshots. A snapshot can be used to create multiple cloned volumes. When a cloned volume is created, the data on the volume is the same as the snapshot. Subsequent modifications on the cloned volume do not affect the original snapshot and other cloned volumes.

- 4.1 Architecture Overview
- 4.2 Key Service Processes
- 4.3 Storage Management
- 4.4 Data Redundancy
- 4.5 Features

4.1 Architecture Overview

The FusionCube distributed storage uses the distributed cluster control and DHT technologies to provide distributed storage functions. Figure 4-1 shows the function architecture of the FusionCube distributed storage.

Figure 4-1 Function architecture of the FusionCube distributed storage

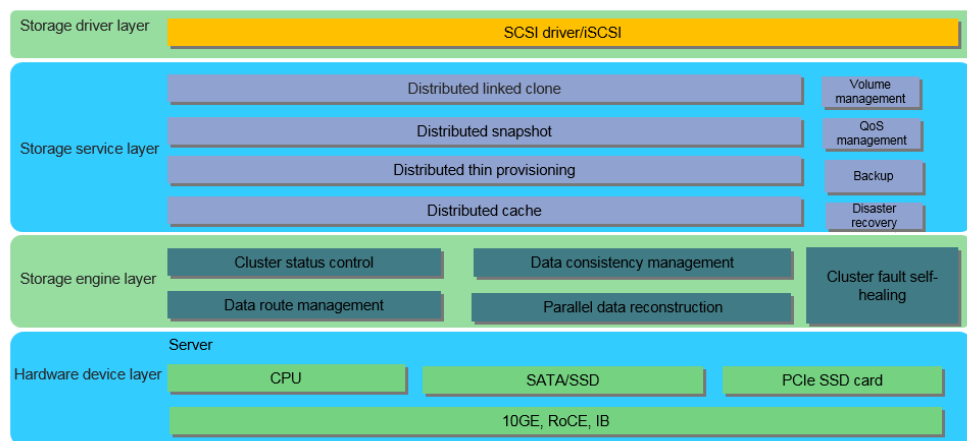


Table 4-1 Components of the FusionCube distributed storage

Name	Description
Storage driver layer	Provides volumes for OSs and databases over Small Computer System Interface (SCSI) or Internet Small Computer Systems Interface (iSCSI).
Storage service layer	Provides various advanced storage features, such as snapshot, linked clone, thin provisioning, distributed cache, and backup

Name	Description
	and DR.
Storage engine layer	Provides basic FusionCube distributed storage functions, including management status control, distributed data routing, strong-consistency replication, cluster self-recovery, and parallel data rebuild.

Figure 4-2 shows the logical architecture of the FusionCube distributed storage.

Figure 4-2 Logical architecture of the FusionCube distributed storage

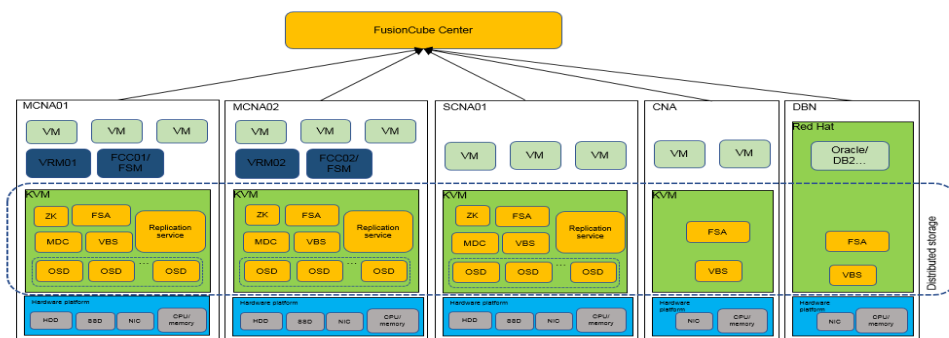


Table 4-2 Logical components of the FusionCube distributed storage

Name	Description
FSM	FusionStorage Manager (FSM) is the management module of FusionStorage. It provides O&M functions including alarm management, service monitoring, operation logging, and data configuration. It is deployed together with FusionCube Center and works in active/standby mode.
FSA	FusionStorage Agent (FSA) is the agent process deployed on each node, for communicating with FusionStorage Manager, collecting monitoring and alarm information of the node, and receiving upgrade packages and performing upgrades when software components on the node need to be upgraded.
ZK	A ZooKeeper (ZK) process. An odd number of ZooKeepers (example, three, five, or seven) need to be deployed in the system. The ZooKeeper cluster provides primary arbitration for the MDC cluster. At least three ZooKeepers need to be deployed and more than half of all ZooKeepers need to be active and accessible.
MDC	The MetaData Controller (MDC) is the metadata control component for controlling the distributed cluster status, data distribution rules, and data rebuild rules. At least three MDCs need to be deployed in the system to form an MDC cluster.

Name	Description
	<p>When the system is started, the ZooKeeper cluster selects the primary MDC from multiple MDCs. The primary MDC monitors other MDCs. When the primary MDC is faulty, a new primary MDC is generated. Each resource pool has a home MDC. When the home MDC of the pool is faulty, the primary MDC assigns another MDC to host the resource pool. One MDC can manage a maximum of two resource pools. The MDC can be started on each storage node as a process. When a resource pool is added, the MDC is automatically started. A maximum of 96 MDCs can be started in a system.</p>
VBS	<p>The Virtual Block System (VBS) is the virtual block storage management component for managing volume metadata. The VBS provides the distributed storage access point service over SCSI or iSCSI, enabling computing resources to access distributed storage resources. The VBS can communicate with all OSDs in the resource pool that can be accessed by the VBS so that the VBS can concurrently access all disks in the resource pool. A VBS process is deployed on each node by default. VBS processes on multiple nodes form a VBS cluster. When the VBS starts, it connects to the primary MDC to determine the primary VBS. Multiple VBSs can be deployed on one node to improve the I/O performance.</p>
OSD	<p>The Object Storage Device (OSD) is the key-value (KV) device service for performing specific I/O operations. Multiple OSD processes are deployed on each node. One OSD process is deployed on one disk by default. When SSD cards are used as the primary storage, multiple OSD processes can be deployed on one SSD card to maximize the SSD card performance. For example, six OSD processes can be deployed on one 2.4 TB SSD card, and each OSD process manages the I/O operations for 400 GB space.</p>

4.2 Key Service Processes

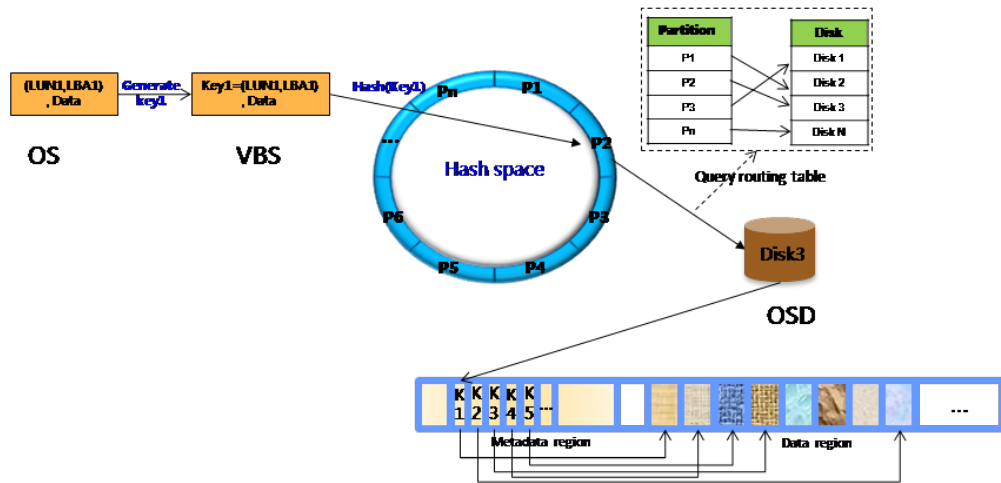
4.2.1 Data Routing

Data routing in the FusionCube distributed storage is implemented in a hierarchical manner:

- The VBS identifies, by calculation, the server and the hard disk where data is stored.
- The OSD identifies, by calculation, the specific location on the hard disk.

The following figure shows the detailed process.

Figure 4-3 Data routing diagram of the FusionCube distributed storage



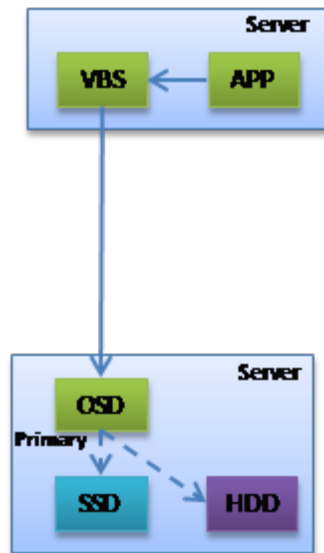
1. When the system is initialized, the FusionCube distributed storage divides the hash space (0 to 2^{32} , available space of the resource pool in the unit of MB) into N equal portions. Each portion is a partition, and all these partitions are evenly allocated to hard disks in the system. For example, in a two-copy scenario, the system has 3600 partitions by default. If the system is equipped with 36 hard disks, each hard disk is allocated with 100 partitions. The partition-hard disk mapping is configured during system initialization and may be adjusted subsequently based on the change of the hard disk quantity. The mapping table requires only small space, and FusionStorage nodes store the mapping table in the memory for rapid routing. In addition, the relationship among the partition, primary disk, secondary disk 1, and secondary disk 2 (no secondary disk 2 when two copies are available) is determined based on the number of copies in the resource pool and other reliability configurations. The primary disk and secondary disks are deployed on different servers and even on different cabinets when a cabinet security plan is developed. The partitioning mechanism used in an EC scenario is the same as that used in a copy scenario. The partition-hard disk mapping is still used to manage hard disks. The difference is that the EC reliability is implemented through data disks and redundant disks.
2. The FusionCube distributed storage logically divides a LUN by every 1 MB of space. For example, a LUN of 1 GB space is divided into 1024 slices of 1 MB space. When an upper-layer application accesses FusionStorage, the SCSI command carries the LUN ID, logical block addressing (LBA) ID, and read/write data content. The OS forwards the message to the VBS of the local node. The VBS generates a key based on the LUN ID and LBA ID. The key contains rounding information of the LBA ID based on the unit of 1 MB. An integer is obtained by hash calculation based on the DHT. The integer ranges from 0 to 2^{32} and falls within the specified partition. The specific hard disk is identified based on the partition-hard disk mapping recorded in the memory. The VBS forwards the I/O operation to the OSD to which the hard disk belongs.
3. Each OSD manages a hard disk. During system initialization, the OSD divides the hard disk into slices of 1 MB and records the slice allocation information in the metadata management area of the hard disk. After receiving the I/O operation sent by the VBS, the OSD searches the hard disk by key for data slice information, obtains the data, and returns the data to the VBS. The entire data routing process is complete. For a write request on a copy, the OSD instructs each secondary OSD to perform the write operation based on the partition-primary disk-secondary disk 1-secondary disk 2 mapping table. Data is returned to the VBS after the primary and secondary OSDs complete the write operation.

4.2.2 I/O Paths

Read I/O Process

Figure 4-4 shows the read I/O process in the FusionCube distributed storage system.

Figure 4-4 Read I/O process of the FusionCube distributed storage



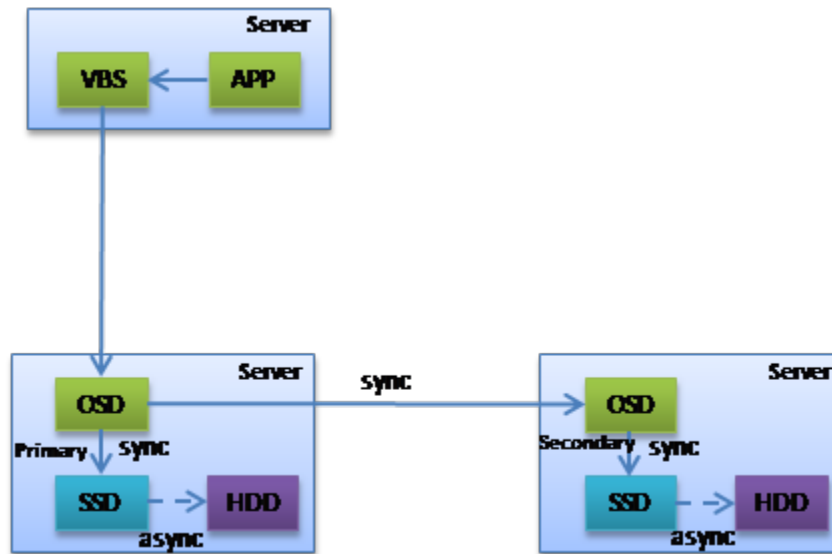
An application delivers a read I/O request to the OS. The OS forwards the read I/O request to the VBS of the local server. The VBS uses the data routing mechanism (for details, see 4.2.1 Data Routing) to identify the primary OSD where the data is located based on the LUN and LBA information in the read I/O request. If the primary OSD is faulty, the VBS reads the data from the secondary OSD.

After receiving the read I/O request, the primary OSD obtains the required data based on the read cache mechanism (for details, see 4.2.3 Cache Mechanisms) and returns a read I/O success message to the VBS.

Write I/O Process

Figure 4-5 shows the write I/O (two copies) process in the FusionCube distributed storage system.

Figure 4-5 Write I/O (two copies) process of the FusionCube distributed storage



An application delivers a write I/O request to the OS. The OS forwards the write I/O request to the VBS of the local server. The VBS uses the data routing mechanism (for details, see 4.2.1 Data Routing) to identify the primary OSD where the data is located based on the LUN and LBA information in the write I/O request.

After receiving the write I/O request, the primary OSD synchronously writes the data in the SSD cache of the local server and in the secondary OSD of another server where the data copy is located. The secondary OSD also synchronously writes the data in the SSD cache of the server where the secondary OSD is located. After the two write operations are successful, the primary OSD returns a write I/O success message to the VBS. In addition, the data in the SSD cache is asynchronously moved to the hard disk.

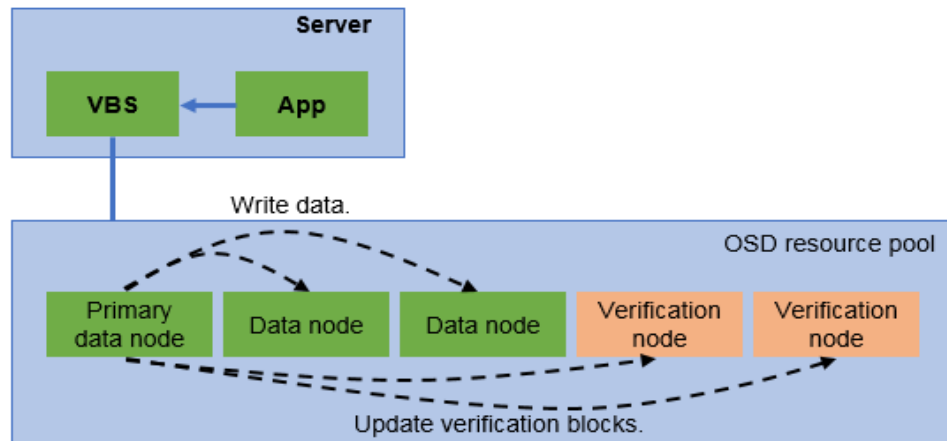
The VBS returns a write I/O success message.

 **NOTE**

If three copies are available, the primary OSD synchronizes the write I/O operation to both the secondary OSD and the third OSD.

The following figure shows the write I/O (EC) process in the FusionCube distributed storage system.

Figure 4-6 Write I/O (EC) process of the FusionCube distributed storage



An application delivers a write I/O request to the OS. The OS forwards the I/O request to the VBS of the local server. The VBS uses the data routing mechanism to determine the primary OSD where the data is located based on the LUN and LBA information in the write I/O request.

After receiving the write I/O request, the primary OSD converts the data to EC data blocks based on the data range of the volume, determines the nodes where the data blocks are located, writes the data blocks to the SSD cache of the corresponding nodes, calculates the verification node data, and writes the verification node data to the SSD cache of the corresponding verification node. In addition, the data in the SSD cache is asynchronously moved to the hard disk.

The VBS returns a write I/O success message.

4.2.3 Cache Mechanisms

The FusionCube distributed storage uses hierarchical cache mechanisms to improve the storage I/O performance. Write and read cache mechanisms follow different processes.

Write Cache

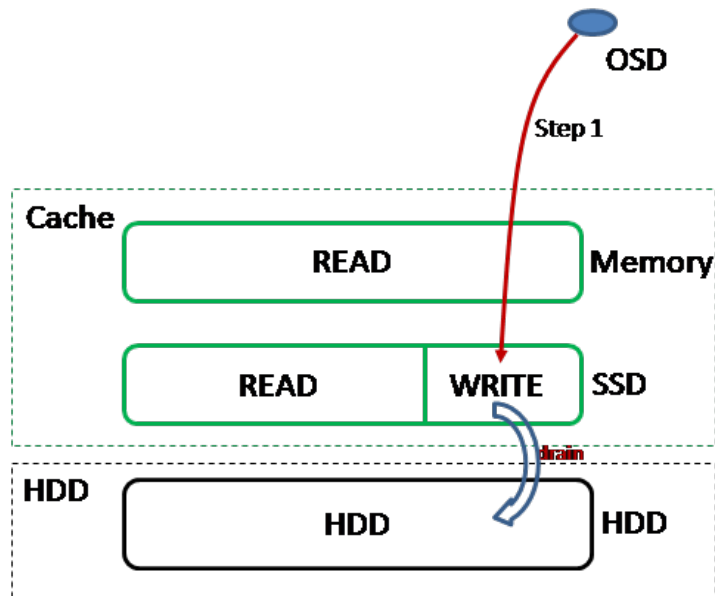
When receiving a write I/O operation sent by the VBS, the OSD temporarily stores the write I/O in the SSD cache to complete the write operation on the local node. In a copy scenario, the content in the SSD cache is complete I/O data. In an EC scenario, the content in the SSD cache is data or redundant strips. In addition, the OSD periodically writes the write I/O data from the SSD cache to the hard disk in batches. A threshold is set for the write cache. Even if the disk refreshing period is not due, the data is written from the cache to the hard disk when the threshold is reached.



NOTE

FusionStorage supports pass-through of large blocks. By default, data blocks greater than 256 KB will be written directly to hard disks rather than being cached. This configuration can be modified.

Figure 4-7 Write cache mechanism of the FusionCube distributed storage



Read Cache

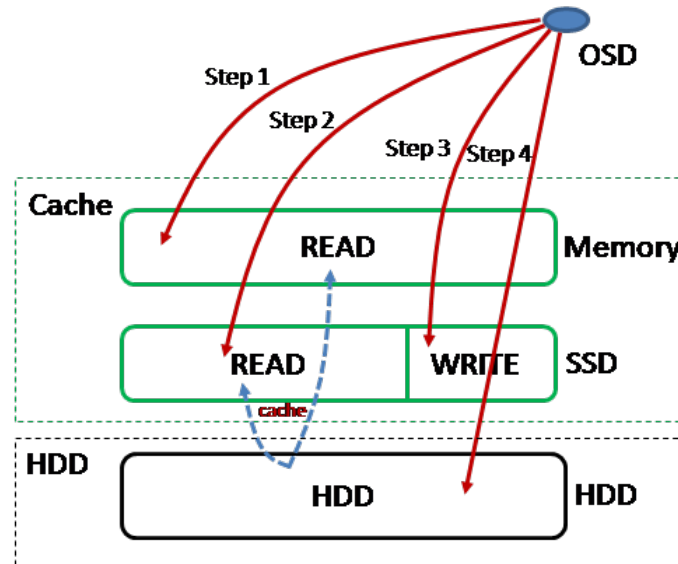
FusionStorage uses SSDs as read cache media to speed up storage access. FusionStorage adopts a hierarchical mechanism for read cache. The first layer is the memory cache, which caches data using the least recently used (LRU) mechanism. The second layer is the SSD cache, which functions based on the hotspot read mechanism. The system collects statistics on each piece of read data and the hotspot access factor. When the threshold is reached, the system automatically caches data to the SSD and removes the data that has not been accessed for a long time. In addition, FusionStorage supports the prefetch mechanism. It collects statistics on the correlation of read data and automatically reads the highly correlated blocks and caches them to the SSD when reading specific data.

When receiving a read I/O operation sent by the VBS, the OSD performs the following operations:

1. Search the read cache of the memory for the required I/O data.
 - If the I/O data is found, return it directly in a copy scenario. In EC is used, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and move the I/O data to the LRU queue head of the read cache.
 - If the I/O data is not found, perform 2.
2. Search the read cache of the SSD for the required I/O data.
 - If the I/O data is found, return it directly in a copy scenario. In an EC scenario, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and add the hotspot access factor of the I/O data.
 - If the I/O data is not found, perform 3.
3. Search the write cache of the SSD for the required I/O data.
 - If the I/O data is found, return it directly in a copy scenario. In an EC scenario, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and add the hotspot access factor of the I/O data. If the hotspot access factor reaches the threshold, the I/O data is recorded to the read cache of the SSD.

- If the I/O data is not found, perform 4.
- 4. Search the hard disk for the required I/O data. Return the data directly in a copy scenario. In an EC scenario, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and add the hotspot access factor of the I/O data. If the hotspot access factor reaches the threshold, the I/O data is recorded to the read cache of the SSD.

Figure 4-8 Read cache mechanism of the FusionCube distributed storage



4.3 Storage Management

4.3.1 Storage Cluster Management

The FusionCube distributed storage uses the cluster management software to manage clusters. The cluster management software performs basic cluster information monitoring, performance monitoring, alarm management, user management, license management, and hardware management.

- Basic cluster information monitoring: allows users to view basic cluster information, including the cluster name, health status, running status, node information, and node process information.
- Performance monitoring: allows users to view the CPU usage, memory usage, bandwidth, IOPS, latency, disk usage, and storage pool usage statistics.
- Alarm management: provides the functions of viewing alarm information, clearing alarms, and shielding alarms.
- User management: enables the system administrator to create new administrators and grant them management permission for managing the system or resources. The administrator can query, delete, create, unlock, and freeze user accounts. Password policies can also be configured to enhance system security.
- License management: allows users to view activated licenses and import new licenses.
- Hardware management

Hardware management includes server management and disk management. Server management allows users to view server software installation status, software version, cluster information, and status and topology of storage pools created on servers, set the maintenance mode to facilitate fault handling, and monitor CPU and memory performance of servers. Disk management allows users to view the hard disk status, slot number, SN, disk usage, and type, and collect statistics on the IOPS, latency, bandwidth, and utilization of disks.

4.3.2 Storage as a Service

Users of the FusionCube distributed storage management portal can be classified into system administrators, system operators, and system viewers by their roles. The functions of the management portal can be divided into the following categories: resource access and configuration, resource management and maintenance, and system management and maintenance. Resource management and maintenance include system overview summary, storage pool management, block storage client management, volume management, virtual file system management, and hardware management.

- **Storage pool management**
You can view the statistics and hard disk topology of the selected storage pool, expand or reduce the capacity of the selected storage pool, and delete the storage pool. You can also create a storage pool.
- **Block storage client management**
You can create and delete clients. You can also view the mounting information of the block storage client and the CPU and memory monitoring statistics to mount or unmount volumes on the block storage client.
- **Volume management**
You can create and delete volumes. When creating a volume, you need to set a resource pool, volume name, and volume size. If the created volume is used based on the SCSI protocol, the volume needs to be mounted. If the volume is used based on the iSCSI protocol, the iSCSI mapping is required. On the iSCSI volume mapping page, you can create hosts or host groups, configure initiators, configure CHAP authentication, and map or unmap volumes to hosts or host groups.

Note: By default, the iSCSI function is disabled. To use the iSCSI function, enable the iSCSI function and add the IP address and port for iSCSI listening.
- **QoS policy management**
You can create or delete QoS policies, and check the QoS policy information on multiple pages.
- **Snapshot management**

You can create linked clone volumes, set QoS policies, delete snapshots, and view the snapshot list on multiple pages. The list information includes the snapshot name, capacity, owning storage pool, and creation time.

4.4 Data Redundancy

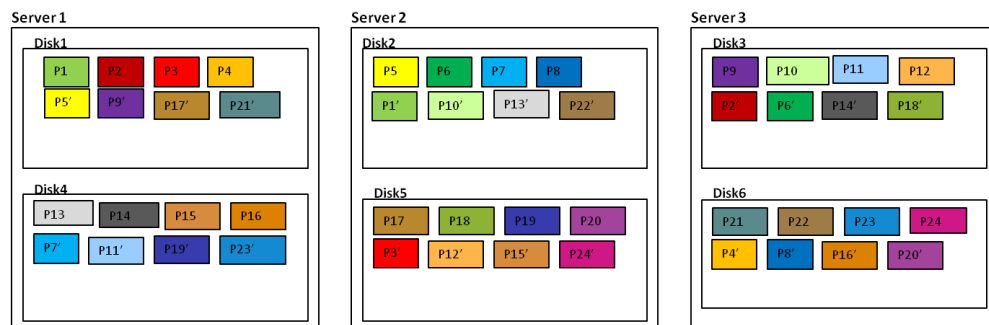
The FusionCube distributed storage supports two data redundancy protection mechanisms: One is the multi-copy mode and the other is the erasure code (EC) mode.

4.4.1 Multi-Copy

The FusionCube distributed storage uses the multi-copy backup mechanism to ensure data reliability. That is, one piece of data can be replicated and saved as 2 or 3 copies. Each volume in the system is fragmented based on 1 MB by default. The fragmented data is stored on cluster nodes based on the DHT algorithm.

Figure 4-9 shows the multiple data copies. For data block P1 on disk 1 of server 1, its data backup is P1' on disk 2 of server 2, and P1 and P1' constitute two copies of the same data block. If disk 1 becomes faulty, P1' can take the place of P1 to provide storage services.

Figure 4-9 Multi-copy diagram of the FusionCube distributed storage

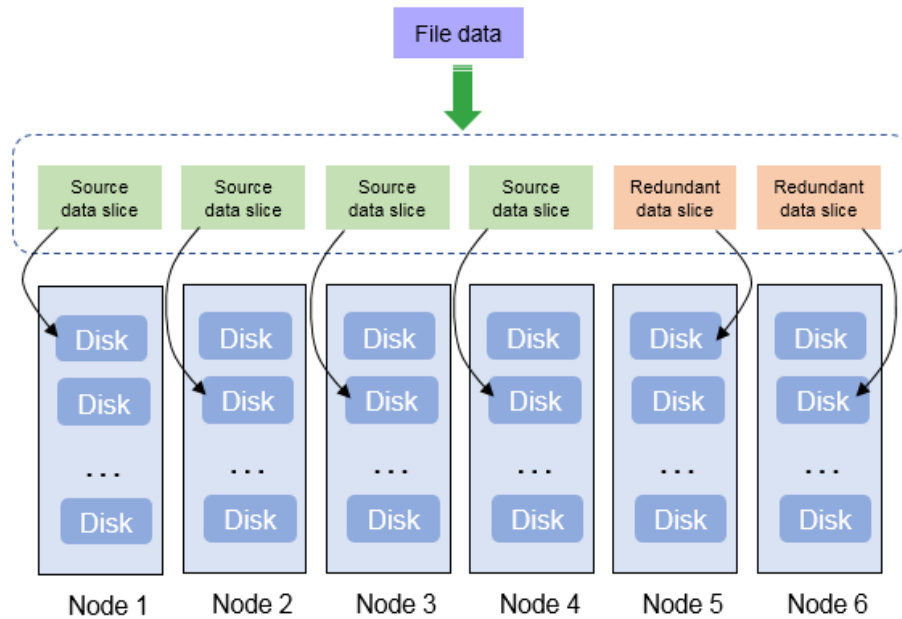


4.4.2 Erasure Code

The FusionCube distributed storage can also use the erasure code (EC) mode to ensure data reliability. Compared with the three-copy mode, the EC data redundancy protection mechanism provides high reliability and high disk utilization.

The EC-based data protection technology used by the FusionCube distributed storage is based on distributed and inter-node redundancy. The FusionCube distributed storage uses the Huawei-developed Low Density Erasure Code (LDEC) algorithm. It is an MDS array code based on the XOR and Galois field multiplication. The minimum granularity is 512 B. It supports Intel instruction acceleration and various mainstream ratios. Data written into the system is divided into N data strips, and then M redundant data strips are generated (both N and M are integers). These data strips are stored on N+M nodes.

Figure 4-10 EC diagram of the FusionCube distributed storage



Example: Four source data slices and two redundant data slices are stored on six nodes.

Data in the same strip is stored on different nodes. Therefore, data in the FusionCube distributed storage system not only supports disk-level faults, but also supports node-level faults to ensure data integrity. As long as the number of concurrently failed nodes is smaller than M , the system can continue to provide services properly. Through data reconstruction, the system is able to restore damaged data to protect data reliability.

The EC data protection mode provided by the FusionCube distributed storage achieves high reliability similar to that provided by traditional RAID based on data replication among multiple nodes. Furthermore, the data protection mode maintains a high disk utilization rate of up to $N/(N + M)$. Different from traditional RAID that requires hot spare disks to be allocated in advance, the FusionCube system allows any available space to serve as hot spare space, further improving storage utilization.

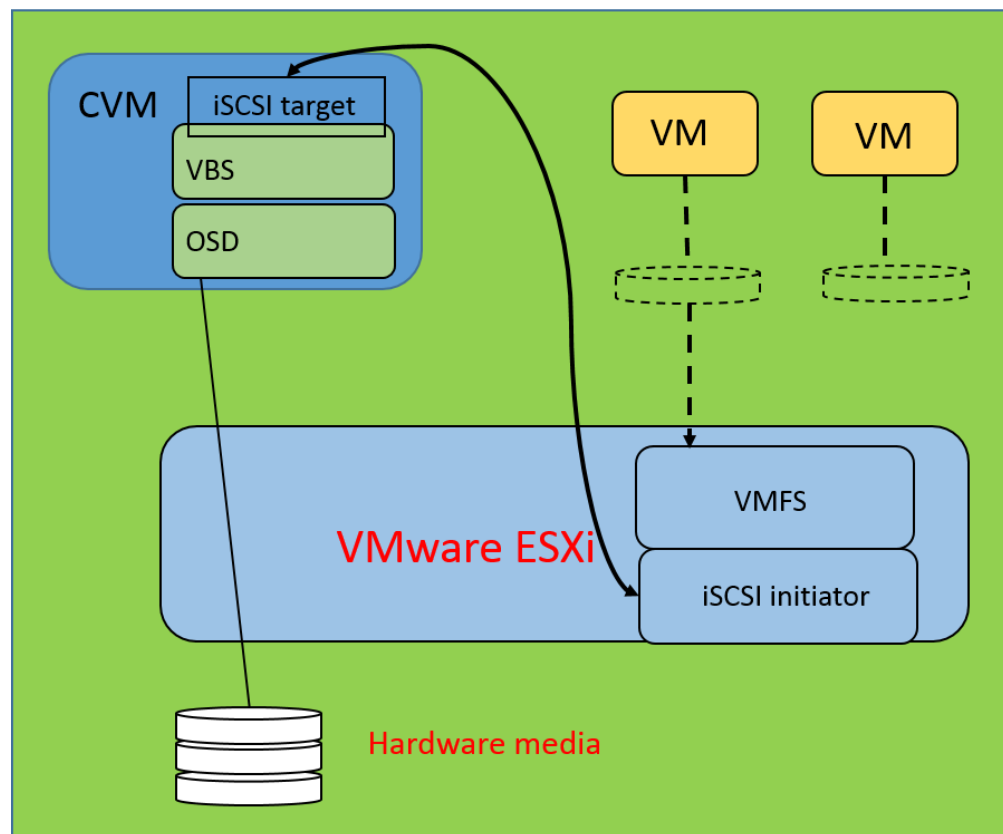
The FusionCube distributed storage provides multiple $N+M$ redundancy ratios. Users can set redundancy ratios based on service requirements. Currently, the following ratios are supported: 3+1, 3+2, 4+2, 8+2, and 12+3. In addition, the FusionCube distributed storage provides the following data slice sizes: 8 KB, 16 KB, 32 KB, and 64 KB. In this way, users can flexibly configure data redundancy ratios and data slice sizes based on actual service requirements to obtain desired reliability levels.

4.5 Features

4.5.1 SCSI/iSCSI Block Interface

The FusionCube distributed storage uses the VBS to provide block interfaces in SCSI or iSCSI mode. The SCSI mode can provide storage access for the local host where the VBS is installed. The physical deployment, FusionSphere, or KVM uses the SCSI mode. The iSCSI mode provides storage access for the VMs or hosts without the VBS. The VMware and Microsoft SQL Server clusters use the iSCSI mode.

Figure 4-11 iSCSI application diagram of the FusionCube distributed storage



The SCSI protocol supports SCSI-3 persistent reservation locks and non-persistent reservation locks.

- Persistent reservation locks can be used for HANA clusters.
- Non-persistent reservation locks can be used for MSCS clusters.

Users access storage devices by connecting the local initiator to the iSCSI target provided by the VBS.

The FusionCube distributed storage supports the following standards of secure access over iSCSI:

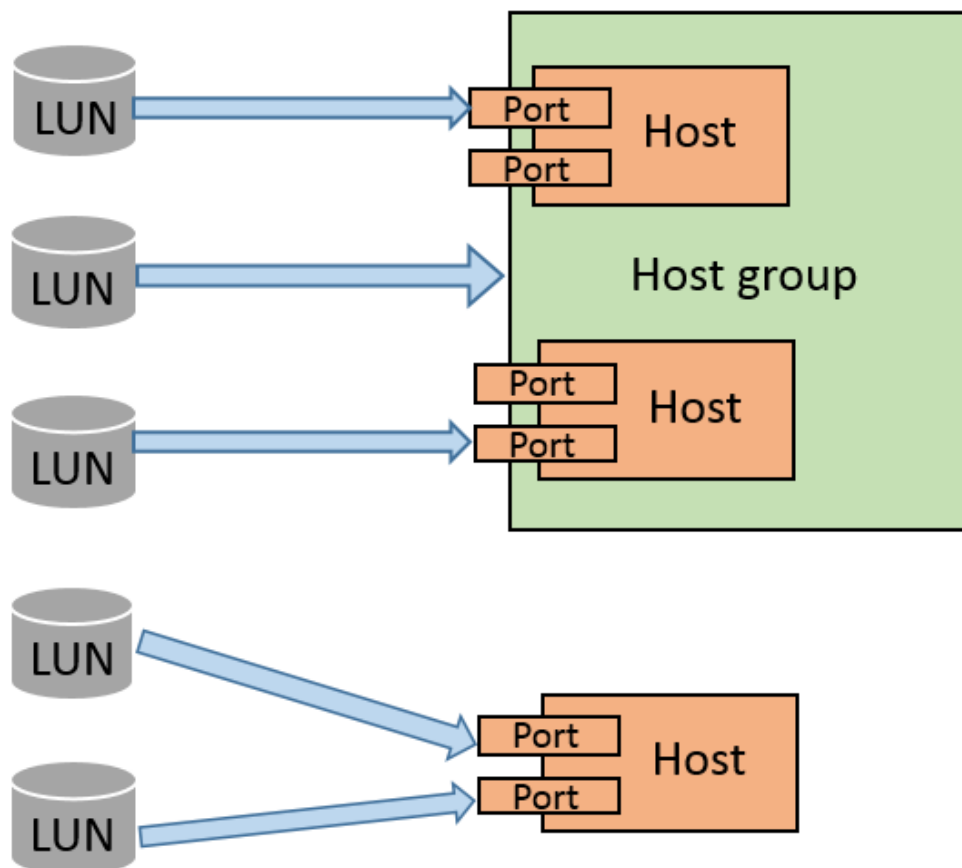
- The FusionCube distributed storage supports CHAP identity verification to ensure that the client access is reliable and secure. CHAP is short for Challenge Handshake Authentication Protocol. The protocol can periodically check the identity of the peer end

through three-way handshakes. The identity authentication can be performed when the initial link is being established or be performed repeatedly after the link is established. By incrementally changing identifiers and variable queries, the protocol provides protection against replay attacks from endpoints, limiting the time for being exposed to a single attack.

- The FusionCube distributed storage authorizes hosts to access LUNs using LUN Masking. LUNs are local devices for SAN storage hosts. To maintain host data, users need to isolate LUNs for each host, to prevent host data from being damaged. LUNs and the world wide name (WWN) addresses of host bus adapters (HBAs) are bound using LUN Masking, to ensure that the LUNs can be accessed only by specified hosts or host clusters. The relationship between hosts and LUNs can be multiple-to-one or one-to-multiple. The one-to-multiple mapping meets storage requirements of small LUNs in virtualization scenarios, and the multiple-to-one mapping meets requirements of cluster systems such as Oracle RAC for shared volumes.

LUN Masking is implemented through mappings among ports, hosts, host groups, and LUNs.

Figure 4-12 LUN Masking component relationship diagram



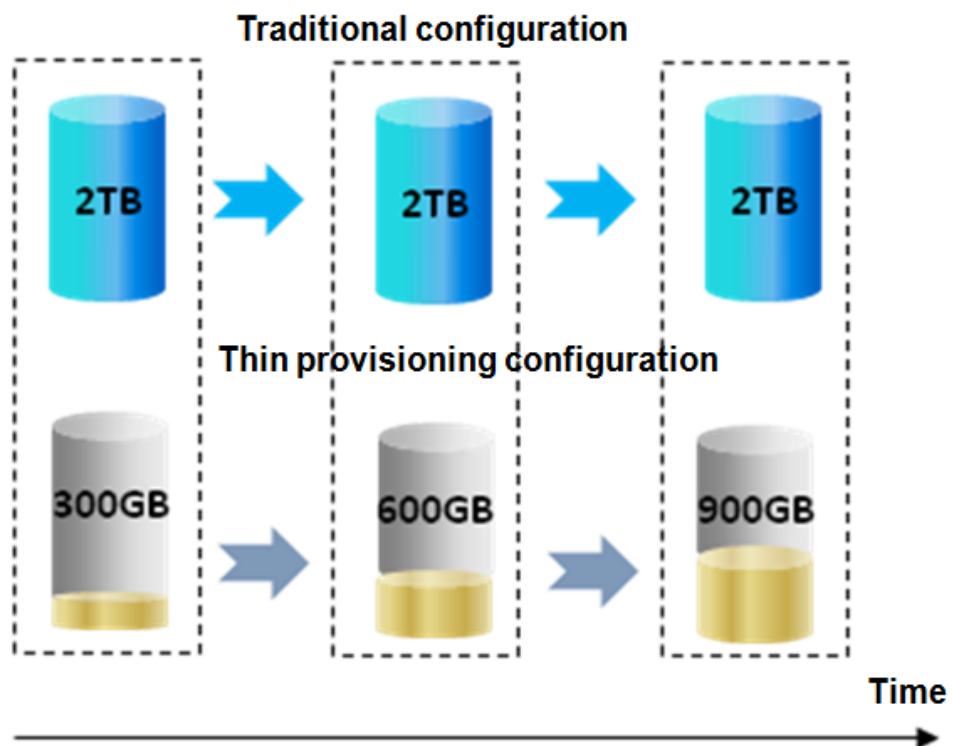
LUNs and storage device ports are bound using LUN Mapping. Hosts are connected to different ports using different LUNs. LUN Mapping can be used when a storage system provides data storage services for multiple application systems and hosts of different application systems are located in different geographical locations.

4.5.2 Thin Provisioning

The FusionCube distributed storage provides the thin provisioning function, which allows users to use much more storage space than that actually available on physical storage devices. The feature remarkably improves storage utilization compared with the traditional method of directly allocating physical storage resources.

The FusionCube distributed storage uses the DHT mechanism. No centralized metadata is required for recording thin provisioning information. Compared with traditional SAN storage devices, the FusionCube distributed storage does not cause system performance deterioration.

Figure 4-13 Thin provisioning of the FusionCube distributed storage

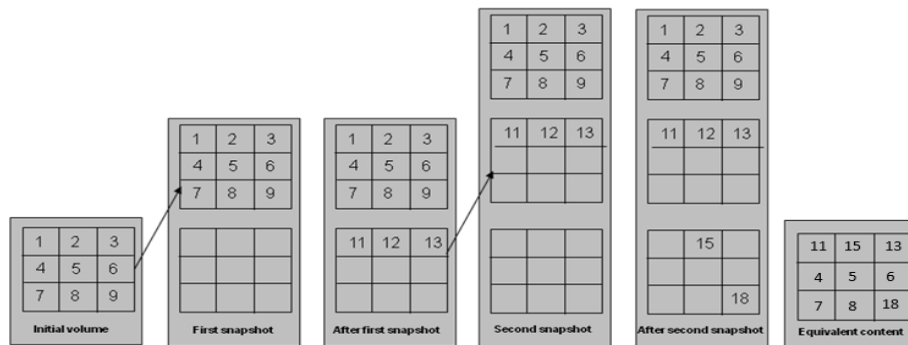


4.5.3 Snapshot

The FusionCube distributed storage provides the snapshot mechanism, which allows the system to capture the status of the data written into a logical volume at a particular time point. The data snapshot can then be exported and used for restoring the volume data when required.

The FusionCube distributed storage uses the redirect-on-write (ROW) mechanism when storing snapshot data. Snapshot creation does not deteriorate performance of the original volume.

Figure 4-14 Snapshot diagram of the FusionCube distributed storage



4.5.4 Shared Volume Snapshot

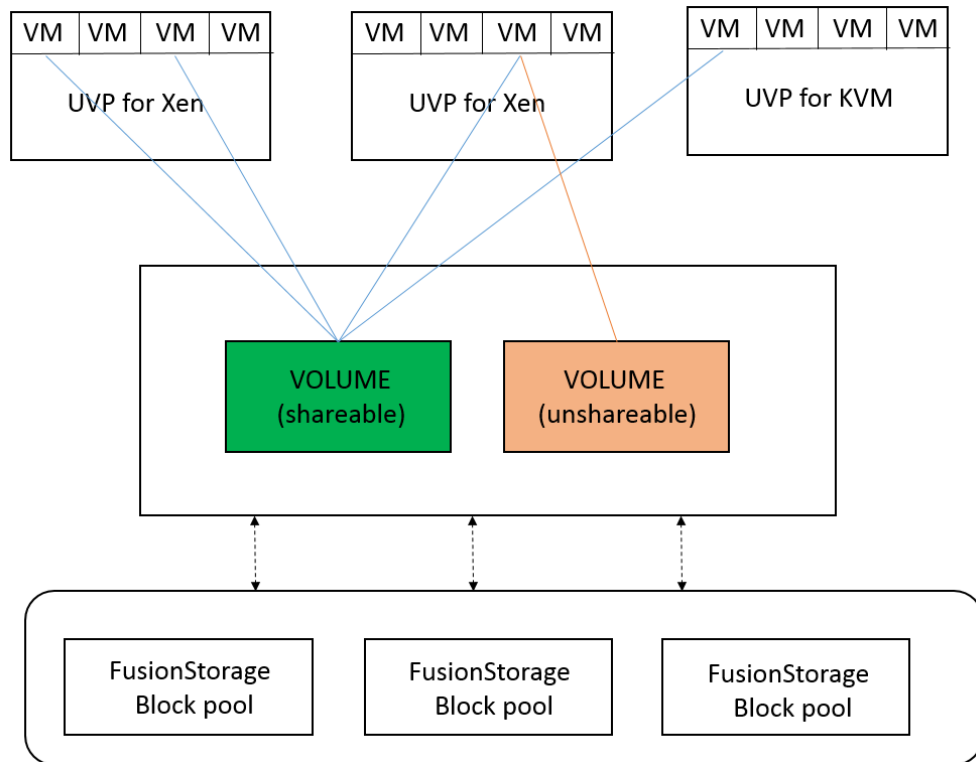
The FusionCube distributed storage also supports snapshot of shared volumes and provides the shared volume backup capability.

SCSI volumes with multiple mount points are called shared volumes. All iSCSI volumes are shared volumes.

Different from the snapshot process of a common volume, a shared volume has multiple mount points and VBSs of all these mount points are likely to send I/O operations. Inter-node cooperation is required to implement I/O suspension in the snapshot process of a shared volume. The FusionCube distributed storage adopts two phases to implement the function.

1. The primary VBS sends a prepare message to all mount point VBSs. The mount point VBSs perform I/O suspension after receiving the prepare message and return an OK message to the primary VBS.
2. The primary VBS sends a commit message to all mount point VBSs and carries the volume metadata to be updated. After receiving the message, the participants update the local metadata, unlock the I/O suspension, and return an OK message to the primary VBS. The transaction is complete.

Figure 4-15 Shared volume snapshot diagram of the FusionCube distributed storage



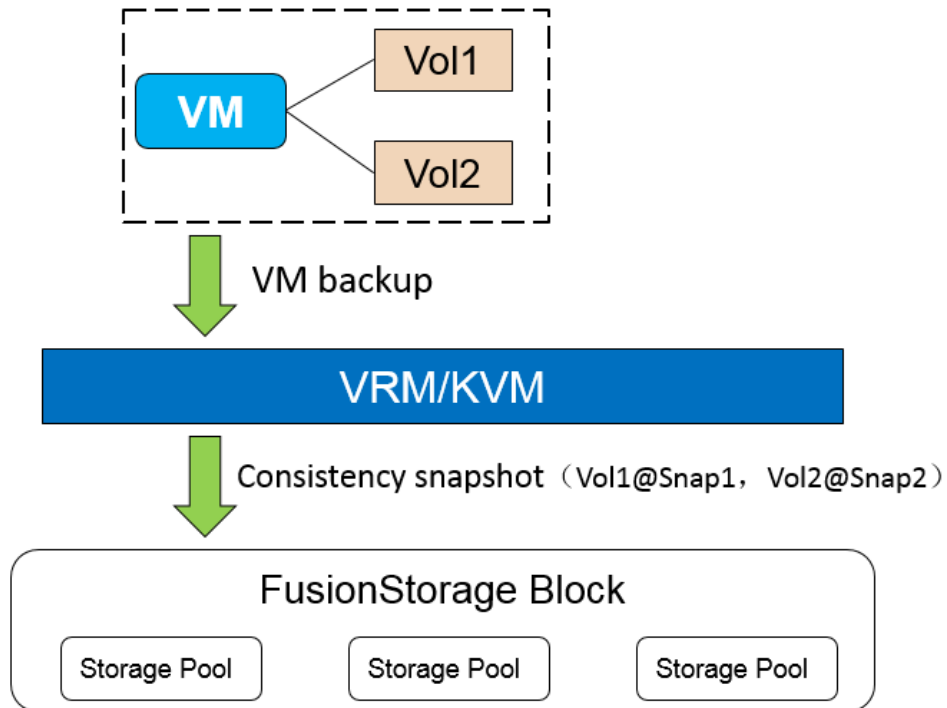
4.5.5 Consistency Snapshot

The consistency snapshot function is used for VM backup. A VM is usually mounted with multiple volumes. When a VM is backed up, all volume snapshots must be at the same time point to ensure data restoration reliability.

The FusionCube distributed storage supports the consistency snapshot capability. Specifically, the FusionCube distributed storage ensures that the snapshots of multiple volumes are at the same time point if an upper-layer application delivers a consistency snapshot request.

To ensure time consistency of snapshots of multiple volumes, the FusionCube distributed storage implements I/O suspension for volumes and then updates the snapshot information operation.

Figure 4-16 Consistency snapshot diagram of the FusionCube distributed storage



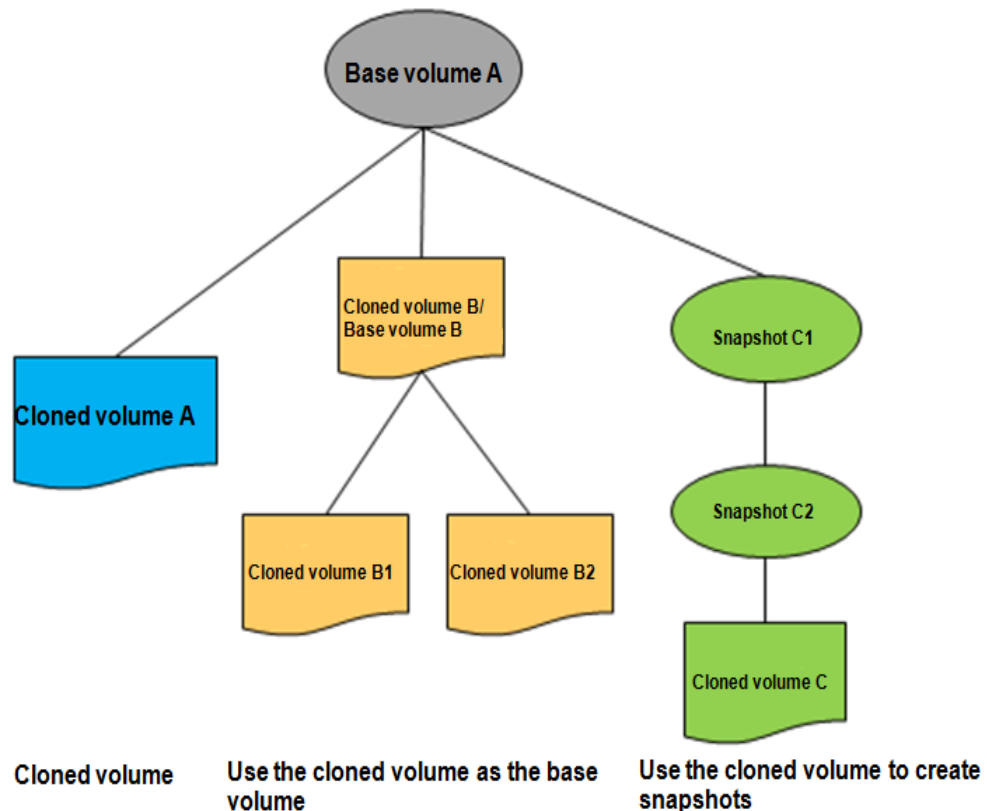
4.5.6 Linked Clone

The FusionCube distributed storage system provides the linked clone mechanism. Multiple cloned volumes can be created based on a volume snapshot. The data content of each cloned volume is the same as that of the volume snapshot. The modification of a cloned volume does not affect the original snapshot and other cloned volumes.

A linked clone ratio of 1:256 is supported to significantly improve storage space utilization.

A cloned volume inherits all the functions of a common volume. You can create snapshots for a cloned volume, use the snapshots to restore the data in the cloned volume, and clone the data in the cloned volume.

Figure 4-17 Linked clone diagram of the FusionCube distributed storage



4.5.7 Multiple Resource Pools

To meet the requirements for storage media of different performance and for fault isolation, the FusionCube distributed storage supports multiple resource pools. A set of FusionCube distributed storage manager manages multiple resource pools. Multiple resource pools share a FusionCube distributed storage cluster. In this cluster, ZooKeepers and the primary MDC are shared. Each resource pool has a home MDC. When a resource pool is created, an MDC automatically starts as the home MDC of the resource pool. The maximum quantity of resource pools is 128 and that of MDCs is 96. If there are more than 96 resource pools, the existing MDCs will be appointed as the home MDCs for the excessive resource pools. Each MDC manages a maximum of 2 resource pools. The home MDC of a resource pool is responsible for the initialization of the resource pool. At the initialization stage, the storage resources are partitioned and the views of the partitions and OSDs are stored in a ZK disk. If the home MDC of a resource pool is faulty, the primary MDC appoints a hosting MDC for the resource pool.

The FusionCube distributed storage supports offline volume migration between multiple resource pools.

Multiple resource pools are planned according to the following rules:

- A resource pool can be planned in copy mode (two or three copies) or EC mode (supporting the 3+1, 3+2, 4+2, 8+2, and 12+3 redundancy ratios).
- A resource pool in two-copy mode supports a maximum of 288 hard disks, and a resource pool in three-copy mode supports a maximum of 2048 hard disks. For more hard disks, a new resource pool must be planned.

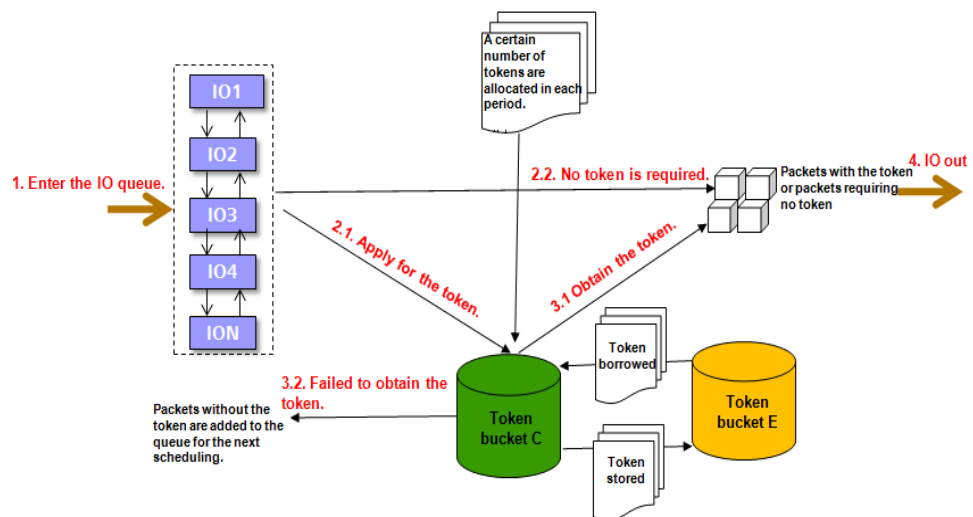
- If the EC mode is used, a resource pool supports a maximum of 2400 hard disks when the redundancy ratio is 3+2, 4+2, 8+2, or 12+3. If hard disks are 8 TB SATA disks, a resource pool with the redundancy ratio of 8+2 supports a maximum of 2136 hard disks. If hard disks are 10 TB SATA disks, a resource pool with the redundancy ratio of 8+2 supports a maximum of 336 hard disks. If the redundancy ratio is 3+1, a resource pool supports a maximum of 96 hard disks.
- The hard disks in a resource pool are of the same type. Hard disks of different types are assigned to different resource pools. The hard disks in a resource pool are of the same capacity; where their capacities are different, they are used as if their capacities are of the smallest.
- The cache media in a resource pool are of the same type. Cache media of different types are assigned to different resource pools.
- It is recommended that each storage node in a resource pool have the same number of hard disks. The gap between hard disk quantities on different nodes cannot exceed 2, and the proportion of the gap to the maximum number of hard disks on a node cannot be greater than 33%.
- Hard disks on a server can be of different types. Resource pools can be created in a way that the hard disks of different types on a server are assigned to different resource pools.

4.5.8 QoS

The FusionCube distributed storage QoS function provides refined I/O control for volumes and provides the burst function. The burst function means that when the volume requirement exceeds the baseline IOPS (bandwidth), the quota that exceeds the benchmark performance can be used within a certain period.

The FusionCube distributed storage QoS function uses the dual-token bucket algorithm to implement I/O control, bandwidth control, and burst function for volumes.

Figure 4-18 QoS mechanism of the FusionCube distributed storage



The token bucket C is used for I/O control. The token bucket E is used to store the remaining tokens. The two buckets work together to implement the burst function.

4.5.9 Active-Active Storage

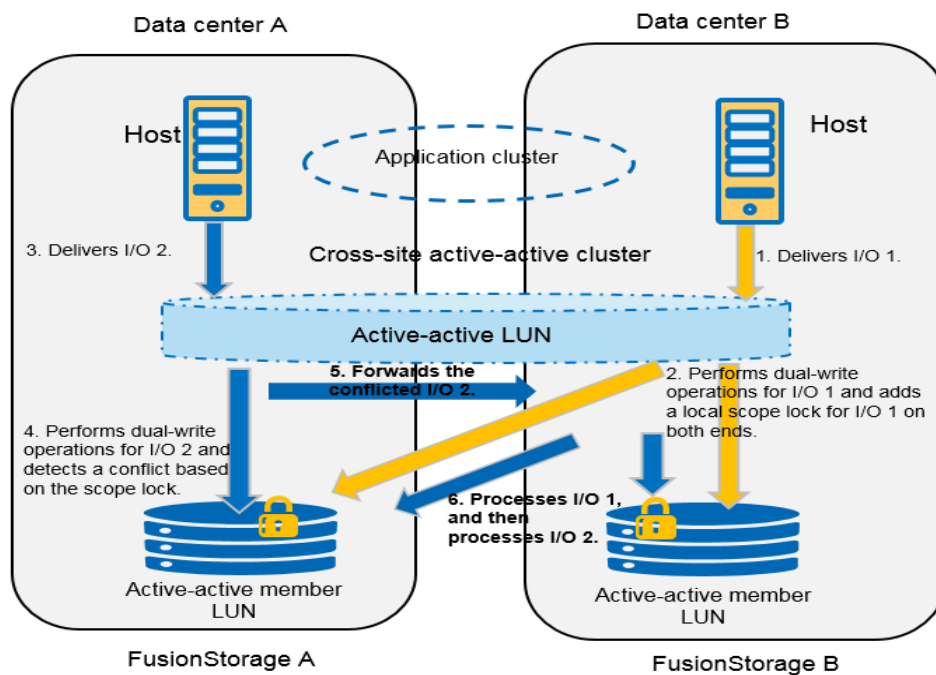
Active-active DR is provided based on two sets of FusionCube distributed storage clusters in data centers A and B. An active-active volume is virtualized based on the volumes of the two sets of FusionCube distributed storage. Hosts in the two data centers can perform read and write services at the same time. If any data center is faulty, no data is lost, and services can be quickly switched to the other site to ensure service continuity.

Replication clusters are added to the original basic services to provide service-based active-active services. Replication clusters are deployed on physical servers, and can be independently installed, upgraded, and expanded to provide active-active services by volume or VM.

Preferred-site arbitration and third-party arbitration modes are supported. When a data center fails, the other one automatically takes over services without manual intervention.

The active-active function interworks with upper-layer applications, such as VMware applications, to form an end-to-end active-active DR solution.

Figure 4-19 Active-active principle of the FusionCube distributed storage



4.5.10 Asynchronous Storage Replication

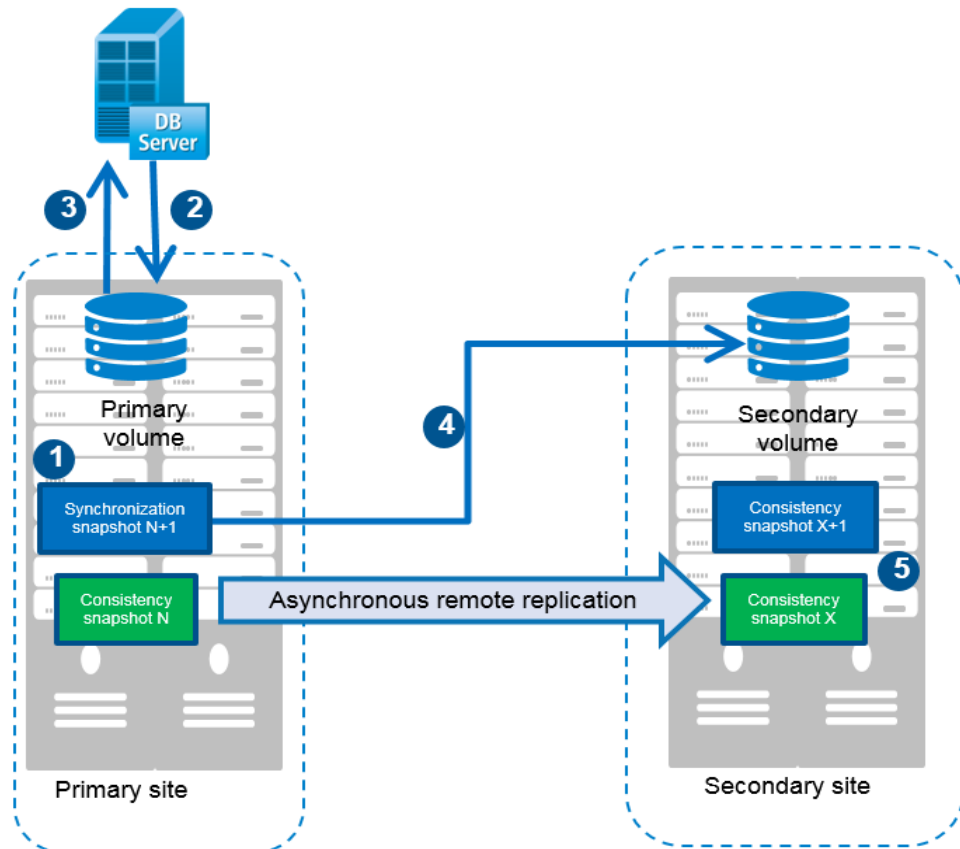
The FusionCube system obtains the data difference by comparing snapshots. The implementation mechanism is as follows:

- When an asynchronous remote replication relationship is set up between the primary volume at the primary site and the secondary volume at the remote secondary site, a full backup is performed for the initial synchronization. The primary site creates a

synchronous snapshot for the primary volume and copies all data from the primary volume to the secondary volume.

- After the initial synchronization is complete, the synchronization snapshot of the primary site is changed to a consistency snapshot and the data status of the secondary volume changes to "consistency". A consistency snapshot is created (the data of the secondary volume is the consistency copy from the primary volume when the primary volume starts the full synchronization). Then, the I/O processing is performed as follows.

Figure 4-20 Asynchronous replication process of the FusionCube distributed storage



Description:

1. A new synchronization snapshot is created at the primary volume at the start of each replication period. At the primary site, the synchronization snapshot newly created is N+1, and the consistency snapshot after the synchronization period is N. The consistency snapshot generated on the secondary site after the synchronization period is X.
2. New host data is written to the primary volume.
3. A write completion response is returned to the host.
4. The primary site compares the consistency snapshot N and the synchronization snapshot N+1 to obtain the data difference, reads data from the snapshot N+1, and writes the differential data to the secondary volume.
5. After the synchronization is complete, the primary site deletes the old consistency snapshot N, converts the new synchronization snapshot N+1 to a consistency snapshot.

The secondary site creates a consistency snapshot X+1, and deletes the old consistency snapshot X.

3. The system automatically synchronizes the incremental data from the primary site to the secondary site at intervals (specified by the user, and the value range is 150 seconds to 1440 minutes). If the synchronization mode is manual, manual intervention is required to trigger the synchronization.

5 Hardware Platform

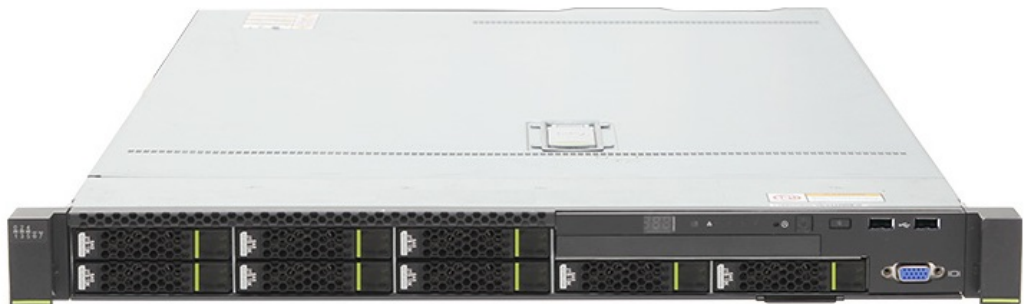
- 5.1 Rack Servers
- 5.2 E9000 Blade Server
- 5.3 X6800 and X6000 High-Density Servers

5.1 Rack Servers

FusionCube HCI 3.2 supports V3 and V5 rack servers. By default, V5 servers are recommended. The supported rack servers include 1-socket, 2-socket, and 4-socket servers. The hardware devices required by the customer can be flexibly configured based on customer's requirements.

5.1.1 RH1288 V3

Figure 5-1 RH1288 V3 compute and storage server



Form factor	2-socket rack server
Processors	1 or 2 Intel® Xeon® E5-2600 v3/v4 processors
Memory	16 DDR4 DIMMs
Hard disks	8 x 2.5-inch SAS/SATA HDDs or SSDs

RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards
PCIe expansion	3 PCIe slots

5.1.2 RH2288H V3

Figure 5-2 RH2288H V3 compute and storage server



Form factor	2-socket rack server
Processors	1 or 2 Intel® Xeon® E5-2600 v3/v4 processors
Memory	24 DDR4 DIMMs
Hard disks	8 x 2.5-inch SAS/SATA HDDs or SSDs 12 x 3.5-inch SAS/SATA HDDs or SSDs + 2 x 2.5-inch SAS HDDs 12 x NVMe SSDs
RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards
PCIe expansion	<ul style="list-style-type: none"> 1 LOM, with 2 x GE ports, 4 x GE ports, or 2 x 10GE ports 4 standard PCIe slots (supporting 2 standard FHHL cards and 2 standard HHL cards)

5.1.3 RH5885H V3

Figure 5-3 RH5885H V3 compute server



Form factor	4-socket rack server
Processors	2 or 4 Intel® Xeon® E7 v3/v4 processors
Memory	96 DDR4 DIMMs
Hard disks	8 x 2.5-inch SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
LOM	4 x GE ports, 2 x GE ports, or 2 x 10GE ports
PCIe expansion	17 PCIe slots (one for the RAID controller card)

5.1.4 1288H V5

Figure 5-4 1288H V5 compute and storage server



Form factor	2-socket rack server
Processors	1 or 2 Intel® Xeon® Scalable processors
Memory	24 DDR4 DIMMs, with a maximum memory speed of 2666 MT/s
Hard disks	8 x 2.5-inch SAS/SATA HDDs or SSDs
RAID support	RAID 0 and RAID 1
LOM	Up to 3 standard PCIe slots
PCIe expansion	1 LOM, with 2 x GE ports and 2 x 10GE ports 1 flexible LOM, with 2 x GE ports, 4 x GE ports, or 2 x 10GE ports 2 standard HHHL PCIe x16 slots and 1 standard FHHL PCIe x8 slot

5.1.5 2288H V5

Figure 5-5 2288H V5 compute and storage server



Form factor	2-socket rack server
Processors	1 or 2 Intel® Xeon® Scalable processors (81/61/51/41 series)
Memory	24 DDR4 DIMMs, with a maximum memory speed of 2666 MT/s
Hard disks	8 x 2.5-inch SAS/SATA/NL-SAS HDDs or SSDs 12 or 16 x 3.5-inch SAS/SATA/NL-SAS HDDs or SSDs 20 x 2.5-inch SAS/SATA HDDs or SSDs 12 x NVMe SSDs
RAID support	RAID 0 and RAID 1
LOM	2 x GE ports + 2 x 10GE ports
PCIe expansion	1 flexible LOM, with 2 x GE ports, 4 x GE ports, 2 x 10GE ports, or 1 or 2 x 56 Gbit/s FDR IB ports Up to 8 standard PCIe slots: 4 standard FHFL PCIe 3.0 x16 cards (bandwidth: x8), 3 standard FHHL PCIe 3.0 x16 cards (bandwidth: x8), and 1 standard FHHL PCIe 3.0 x8 standard card (bandwidth: x8), 1 RAID controller card, and 1 flexible LOM

5.1.6 2488 V5

Figure 5-6 2488 V5 compute server



Form factor	4-socket rack server
Processors	2 or 4 Intel® Xeon® Scalable processors (81/61/51 series)
Memory	32 DDR4 DIMMs
Hard disks	8 x 2.5-inch SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
LOM	2 × GE + 2 x 10GE optical or electrical ports
PCIe expansion	9 standard PCIe slots

5.1.7 2488H V5

Figure 5-7 2488H V5 compute server



Form factor	4-socket rack server
Processors	2 or 4 Intel® Xeon® Scalable processors
Memory	48 DDR4 DIMMs
Hard disks	8 x 2.5-inch SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
LOM	2 × GE + 2 x 10GE optical or electrical ports
PCIe expansion	3 standard HHHL PCIe 3.0 x16 cards and 7 standard HHHL PCIe 3.0 x8 cards

5.2 E9000 Blade Server

5.2.1 E9000 Chassis

Huawei E9000 is a 12U blade server that integrates compute nodes, switch modules, and management modules in flexible configuration.

The E9000 has the following functions and features:

- The chassis can house up to 8 full-width or 16 half-width compute nodes in flexible configuration.
- The cooling capacity for a half-width slot is 850 W.

- The cooling capacity for a full-width slot is 1700 W.
- A half-width compute node supports up to 2 processors and 24 DIMMs.
- A full-width compute node supports up to 4 processors and 48 DIMMs.
- An E9000 server supports a maximum of 32 processors and 24 TB of memory.
- The midplane provides a maximum of 5.76 Tbit/s switch capacity.
- The server provides two pairs of slots for pass-through or switch modules supporting a variety of switching protocols, such as Ethernet and IB, and provide direct I/O ports.

Figure 5-8 E9000 appearance



 **NOTE**

FusionCube supports a maximum of three E9000 chassis in one cabinet.

5.2.2 E9000 Compute Nodes

FusionCube supports the following blades:

- 2-socket CH121 V3 compute node
- 2-socket CH222 V3 compute and storage node
- 2-socket CH220 V3 compute and I/O expansion node
- 2-socket CH225 V3 compute and storage node
- 4-socket CH242 V3 compute node
- 2-socket CH121 V5 compute node
- 2-socket CH225 V5 compute and storage node
- 4-socket CH242 V5 compute node

Figure 5-9 CH121 V3 compute node



Form factor	2-socket half-width blade
Processors	1 or 2 Intel Xeon E5-2600 v3/v4 processors
Memory	24 DDR4 DIMMs
Hard disks	2 x 2.5-inch SAS/SATA disks, or 2 PCIe SSDs
RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards, 2 SATADOMs, and 1 USB flash drive (USB 3.0)
PCIe expansion	<ul style="list-style-type: none"> • 2 PCIe x16 mezzanine cards • 1 standard FHHL PCIe x16 card

Figure 5-10 CH220 V3 compute and I/O expansion node



Form factor	2-socket full-width blade
Processors	1 or 2 Intel Xeon E5-2600 v3/v4 processors
Memory	16 DDR4 DIMMs
Hard disks	2 x 2.5-inch SAS/SATA drives, or 2 x PCIe SSDs

RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards, 2 SATADOMs, and 1 USB flash drive (USB 3.0)
PCIe expansion	<ul style="list-style-type: none"> • 4 mezzanine cards (two x16 and two x8) • 6 standard PCIe 3.0 x16 cards in any of the following combinations: <ul style="list-style-type: none"> – 6 FHHL PCIe cards – 1 full height full length (FHFL) PCIe card (occupying 2 slots) and 4 FHHL PCIe cards – 2 FHFL PCIe cards (each occupying 2 slots)

Figure 5-11 CH225 V3 compute and storage node



Form factor	2-socket full-width blade
Processors	1 or 2 Intel® Xeon® E5-2600 v3/v4 processors
Memory	24 DDR4 DIMMs
Hard disks	12 x 2.5-inch NVMe SSDs and 2 x 2.5-inch SSDs or SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards, 2 SATADOMs, and 1 USB flash drive (USB 3.0)
PCIe expansion	4 PCIe x16 mezzanine cards

Figure 5-12 CH222 V3 compute and storage node



Form factor	2-socket full-width blade
Processors	1 or 2 Intel Xeon E5-2600 v3/v4 processors
Memory	24 DDR4 DIMMs
Hard disks	15 x 2.5-inch SSDs or SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards, 2 SATADOMs, and 1 USB flash drive (USB 3.0)
PCIe expansion	<ul style="list-style-type: none"> • 2 PCIe x16 mezzanine cards • 1 standard FHHL PCIe x16 card

Figure 5-13 CH242 V3 compute node



Form factor	4-socket full-width blade
Processors	2 or 4 Intel® Xeon® E7 v2 or v3 series processors, with up to 18 cores and 165 W output power

Memory	<p>When equipped with Intel® Xeon® E7 v2 series processors, the compute node supports up to 32 DDR3 DIMMs and 1600 MHz bandwidth.</p> <p>When equipped with Intel® Xeon® E7 v3 series processors, the compute node supports up to 32 DDR4 DIMMs and 1866 MHz bandwidth.</p>
Hard disks	8 SSDs or SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards and 1 USB flash drive (USB 3.0)
PCIe expansion	<ul style="list-style-type: none"> • 4 PCIe x16 mezzanine cards • 2 standard FHHL PCIe x16 cards

Figure 5-14 CH121 V5 compute node



Form factor	2-socket half-width blade
Processors	1 or 2 Intel® Xeon® Scalable processors
Memory	24 DDR4 DIMMs
Hard disks	2 x 2.5-inch SAS/SATA disks, or 2 PCIe SSDs
RAID support	RAID 0 and RAID 1
Built-in flash	Up to 4 M.2 SSDs (SATA ports)
PCIe expansion	<ul style="list-style-type: none"> • 2 PCIe x16 mezzanine cards • 1 standard FHHL PCIe x16 card

Figure 5-15 CH225 V5 compute and storage node



Form factor	2-socket full-width blade
Processors	1 or 2 Intel® Xeon® Scalable processors
Memory	24 DDR4 DIMMs
Hard disks	12 x 2.5-inch SATA/SAS/NVMe disks, mixed configuration of HDDs and SSDs supported 2 x 2.5-inch SSDs or SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
Built-in flash	Up to 6 M.2 SSDs (including 2 built-in SSDs)
PCIe expansion	4 PCIe x16 mezzanine cards

Figure 5-16 CH242 V5 compute node



Form factor	4-socket full-width blade
Processors	2 or 4 Intel® Xeon® Scalable processors

Memory	48 DDR4 DIMMs, with a maximum memory speed of 2666 MT/s
Hard disks	4 x 2.5-inch SSDs or SAS/SATA HDDs, or 4 NVMe SSDs, or a maximum of 8 M.2 SSDs (SATA ports) Hot-swap of a single disk is supported.
RAID support	RAID 0 and RAID 1
Built-in flash	2 microSD cards and 1 USB flash drive (USB 3.0)
PCIe expansion	<ul style="list-style-type: none"> • 4 PCIe x16 mezzanine cards • 1 standard HHHL PCIe x16 card

5.2.3 High-Performance Switch Modules

FusionCube uses CX310 switch modules, which support 10GE network. Each chassis can be configured with two CX310s.

Figure 5-17 CX310 10GE switch module



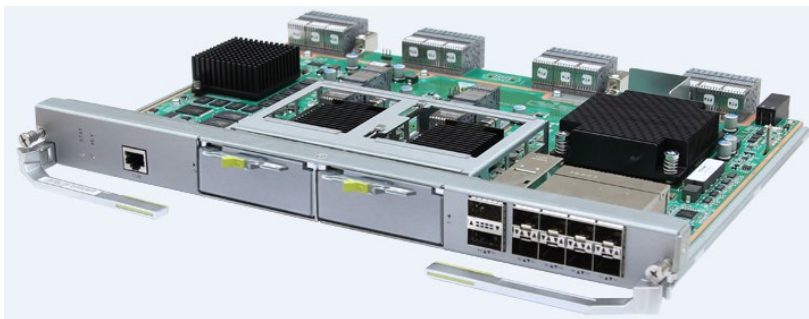
Model	CX310 10GE switch module
Network ports	16 x 10GE uplink ports 32 x 10GE downlink ports
Network feature	L2: VLAN/MSTP/LACP/TRILL/Stack/IGMP L3: RIP/OSPF/ISIS/BGP/VRRP/BFD/PIM QoS: DCBX/PFC/ETS/ACL/CAR/DiffServ Security: IPSG/MFF/DAI/FSB/DHCP Snooping
Management ports	2 RS232 management serial ports (one for service management and one for device management)

Figure 5-18 CX611 InfiniBand switch module



Model	CX611 IB switch module
Network ports	18 QDR QSFP or FDR QSFP+ uplink ports 16 QDR or FDR 4X downlink ports (one QDR or FDR 4X downlink port for each half-width slot)
Network feature	QDR/FDR auto-negotiation. Ideal for applications demanding low latency and high bandwidth.
Management ports	2 RS232 management serial ports (one for service management and one for device management)

Figure 5-19 CX320 10GE switch module



Model	CX320 10GE switch module
Network ports	8 x 10GE + 2 x 40GE uplink ports 32 x 10GE downlink ports
Network feature	L2: VLAN/MSTP/LACP/TRILL/Stack/IGMP L3: RIP/OSPF/ISIS/BGP/VRRP/BFD/PIM QoS: DCBX/PFC/ETS/ACL/CAR/DiffServ Security: IPSG/MFF/DAI/FSB/DHCP Snooping

Management ports	2 RS232 management serial ports (one for service management and one for device management)
------------------	--

Figure 5-20 CX620/CX621 IB switch module



Model	CX620/CX621 IB switch module
Network ports	18 FDR or EDR uplink ports 16 FDR/EDR downlink ports
Network feature	FDR/EDR autonegotiation Ideal for applications demanding low latency and high bandwidth.
Management port	1 RS232 management serial port

5.3 X6800 and X6000 High-Density Servers

FusionCube HCI 3.2 supports two high-density server platforms:

- **X6000**
 The X6000 server provides high computing density. It comes with four nodes in a 2U chassis. Each node supports six 2.5-inch disks (including the system disk), an LOM supporting two GE and two 10GE ports, and one NVMe SSD card serving as the cache.
- **X6800**
 The X6800 provides high computing and storage density. It comes with four nodes in a 4U chassis. Each node supports two system disks, ten 3.5-inch disks, and two rear PCIe x8 slots.

5.3.1 X6800 Chassis

The X6800 server uses a new-generation server architecture developed by Huawei. It features high density and reliability, flexible expansion, easy maintenance and management, and high

energy efficiency. The X6800 supports a wide range of server nodes to meet different service requirements.

The X6800 has the following features:

- **Flexible configuration**

The X6800 delivers excellent flexibility and scalability.

- The 4U architecture integrates advantages of blade servers, allowing users to configure single-slot, dual-slot, four-slot, or eight-slot server nodes as required.
- The X6800 provides a wide variety of server nodes, such as pure compute nodes, GPU acceleration nodes, and nodes with different configurations of compute and storage resources, to meet different service requirements.

- **High computing density**

The X6800 provides higher density than a conventional rack server, saving the footprint in equipment rooms.

- It provides compute density twice that of a conventional 1U rack server and four times that of a conventional 2U rack server in the same cabinet space, which greatly improves space utilization in the equipment room.
- The X6800 provides storage density twice that of a conventional 1U rack server. When fully configured with four server nodes, the cabinet supports up to 480 x 3.5-inch hard disks.
- A fully configured cabinet supports up to 80 4U8 server nodes (each node occupies one slot in a chassis), with up to 160 processors and 80 TB memory capacity.

- **Unified management and easy maintenance**

The X6800 leverages the blade server architecture to provide unified management and easy maintenance.

- The X6800 uses the intelligent Baseboard Management Controller (iBMC) and Hyper Management Module (HMM) to implement unified server management. By incorporating advantages of rack and blade servers, the X6800 supports front maintenance and front and rear cabling. This feature meets deployment requirements of traditional equipment rooms (requiring rear cabling) and new equipment rooms (requiring front cabling) to support maintenance in the cold air area.
- It adopts a modular design with hot-swappable key components, greatly improving O&M efficiency.

- **Shared architecture and high energy efficiency**

All server nodes in an X6800 chassis share the power supplies and heat dissipation system.

- The server nodes share four PSUs and five fan modules in an X6800 chassis, which simplifies deployment and increases PSU and fan module utilization.
- It uses Huawei Dynamic Energy Management Technology (DEMT) to control system energy consumption, maximizing energy efficiency.

- **Redundancy design and high reliability**

The X6800 uses reliable system architecture to ensure stable and long-term service running.

- The passive backplane prevents single point of failures (SPOFs), delivering higher reliability and security than an active backplane.
- Redundant fan modules and PSUs and RAID configuration prevent data loss and service interruption.

- Carrier-class components and manufacturing processes provide higher stability and longer lifecycle.

Figure 5-21 X6800 front view



Figure 5-22 X6800 rear view



5.3.2 X6800 Server Nodes

XH628 V3 Server Node

The XH628 V3 server node (XH628 V3) is a 2-socket dual-slot storage node used in Huawei X6800 servers. It uses an innovative design to deliver high storage density and performance while breaking through power limits. It is also easy to manage and maintain.

Figure 5-23 XH628 V3



Form factor	2-socket dual-slot server node
Number of processors	2
Processor model	Intel® Xeon® E5-2600 v3/v4
Memory	16 DDR4 DIMMs
Hard disks	12 x 3.5-inch or 2.5-inch hot-swappable SAS/SATA HDDs or SSDs
RAID support	RAID 0 and RAID 1
LOM	2 or 4 GE ports or 2 x 10GE ports
PCIe expansion	5 PCIe slots (2 HHHL PCIe3.0 x8 cards + 2 rear HHHL PCIe3.0 x8 cards +1 RAID controller card) If the two front PCIe cards are configured, the two 2.5-inch hard disks cannot be configured.
Onboard storage medium	2 miniSSDs (SATADOMs) + 1 USB flash drive
Front ports	1 universal connector port (UCP) (supports 1 VGA port, 3 USB 2.0 ports, and 1 serial port) + 1 GE management port

XH622 V3 Server Node

The XH622 V3 server node (XH622 V3) is a 2-socket GPU server node. An X6800 chassis can hold a maximum of four XH622 V3 nodes. It is designed to break through energy restrictions and improve system computing density. It features high performance and computing density, and easy management and maintenance.

Figure 5-24 XH622 V3



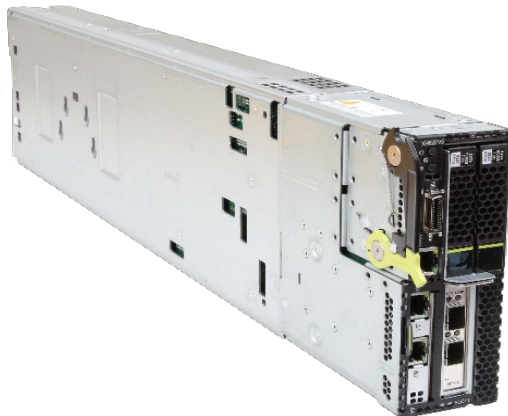
Form factor	2-socket dual-slot server node
Number of processors	2

Processor model	Intel® Xeon® E5-2600 v3/v4
Memory	16 DDR4 DIMMs
Hard disks	4 x 2.5-inch SSDs or SAS/SATA HDDs
RAID support	RAID 0 and RAID 1
LOM	2 or 4 GE ports or 2 x 10GE ports
PCIe expansion	Up to 5 PCIe slots
Onboard storage medium	2 miniSSDs (SATADOMs) + 1 USB flash drive
Front ports	1 UCP (supports 1 VGA port, 3 USB 2.0 ports, and 1 serial port) + 1 GE management port

XH620 V3 Server Node

The XH620 V3 server node (XH620 V3) is 2-socket server node of Huawei X6800 servers. An X6800 chassis can hold a maximum of eight XH620 V3 nodes. It is designed to make breakthroughs in space and improve computing density. It features high performance and computing density, and easy management and maintenance.

Figure 5-25 XH620 V3



Form factor	2-socket dual-slot server node
Number of processors	2
Processor model	Intel® Xeon® E5-2600 v3/v4
Memory	16 DDR4 DIMMs
Hard disks	4 x 2.5-inch SSDs or SAS/SATA HDDs
RAID support	RAID 0 and RAID 1

LOM	2 or 4 GE ports or 2 x 10GE ports
PCIe expansion	Up to three PCIe slots
Onboard storage medium	2 miniSSDs (SATADOMs) + 1 USB flash drive
Front ports	1 UCP (supports 1 VGA port, 3 USB 2.0 ports, and 1 serial port) + 1 GE management port

XH628 V5 Server Node

The XH628 V5 server node (XH628 V5) is a 2-socket dual-slot storage node used in Huawei X6800 servers. It uses an innovative design to deliver high storage density and supreme performance while breaking through power limits. It is also easy to manage and maintain.

Figure 5-26 XH628 V5



Form factor	2-socket dual-slot server node
Number of processors	2
Processor model	Intel® Xeon® Scalable processor
Memory	16 DDR4 DIMMs
Hard disks	12 x 3.5-inch or 2.5-inch hot-swappable SAS/SATA HDDs or SSDs
RAID support	RAID 0 and RAID 1
LOM	2 x GE ports + 2 x 10GE ports
PCIe expansion	5 PCIe slots (2 front HHHL PCIe3.0 x8 cards + 2 rear HHHL PCIe3.0 x8 cards + 1 RAID controller card)
Front ports	1 UCP (supports 1 VGA port, 3 USB 2.0 ports, and 1 serial port) + 1 GE management

	port
--	------

5.3.3 X6000 Chassis

The X6000 is a new-generation 2U multi-node high-density server developed by Huawei. It features flexible configuration, high density and reliability, easy O&M, and high energy efficiency. The X6000 has the following features:

- **Ultra-high density and small footprint**

The X6000 provides higher density than a conventional rack server, saving the footprint in equipment rooms.

 - It provides compute density twice that of a conventional 1U rack server and four times that of a conventional 2U rack server in the same cabinet space, which greatly improves space utilization in the equipment room.
 - Each node supports six 2.5-inch hard disks.
- **Unified management and easy maintenance**

The X6000 leverages the blade server architecture to provide unified management and easy maintenance.

 - It uses the iBMC and HMM to implement unified server management. By incorporating advantages of rack and blade servers, the X6000 allows nodes to be installed at the rear and supports rear cabling.
 - It adopts a modular design with hot-swappable key components, greatly improving O&M efficiency.
- **Shared architecture and high energy efficiency**

All server nodes in an X6000 chassis share the power supplies and heat dissipation system.

 - The server nodes share the two PSUs and four fan modules, which simplifies deployment and improves PSU and fan module utilization.
 - It uses Huawei Dynamic Energy Management Technology (DEMT) to control system energy consumption, maximizing energy efficiency.
- **Redundancy design and high reliability**

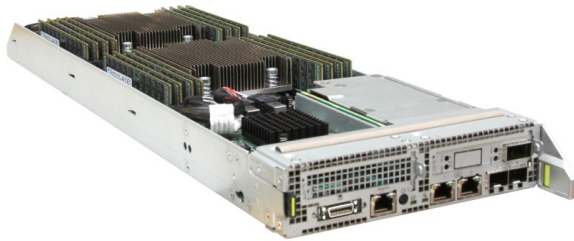
The X6000 uses reliable system architecture to ensure stable and long-term service running.

 - Redundant fan modules and PSUs and RAID configuration prevent data loss and service interruption.
 - Carrier-class components and manufacturing processes provide higher stability and longer lifecycle.

5.3.4 X6000 Server Nodes

The XH321 V3 is a 2-socket server node designed for Huawei X6000 servers. An X6000 server can hold a maximum of four server nodes in a 2U chassis. The XH321 V3 adopts an innovative design to provide high compute density and break through space limits and it is easy to manage and maintain.

Figure 5-27 XH321 V3



Number of processors	2
Processor model	Intel® Xeon® E5-2600 v3/v4
Memory	16 DDR4 DIMMs
Hard disks	6 x 2.5-inch SAS/SATA HDDs or SSDs or NVMe SSDs
RAID support	RAID 0 and RAID 1
LOM	2 x GE ports + 2 x 10GE ports
PCIe expansion	2 PCIe slots
Onboard storage medium	1 SATADOM and 1 M.2 SSD
Front ports	1 UCP (supports 1 VGA port, 3 USB 2.0 ports, and 1 serial port) + 1 GE management port

Figure 5-28 XH321 V5



Number of processors	2
Processor model	Intel® Xeon® Scalable processor
Memory	16 DDR4 DIMMs
Hard disks	6 x 2.5-inch SAS/SATA HDDs or SSDs or NVMe SSDs
RAID support	RAID 0 and RAID 1
LOM	2 x GE ports + 2 x 10GE ports
PCIe expansion	2 PCIe slots
Front ports	1 UCP (supports 1 VGA port, 3 USB 2.0 ports, and 1 serial port) + 1 GE management port

6 Installation, Deployment, and O&M

The FusionCube HCI features easy deployment and simple O&M. One-stop delivery is provided, and services can be rolled out on the day of system deployment. Simplified O&M reduces the skill requirements for IT management personnel. The FusionCube HCI provides FusionCube Builder (FCB) for automatic installation and deployment and FusionCube Center for unified O&M.

[6.1 Automatic Deployment](#)

[6.2 Unified O&M](#)

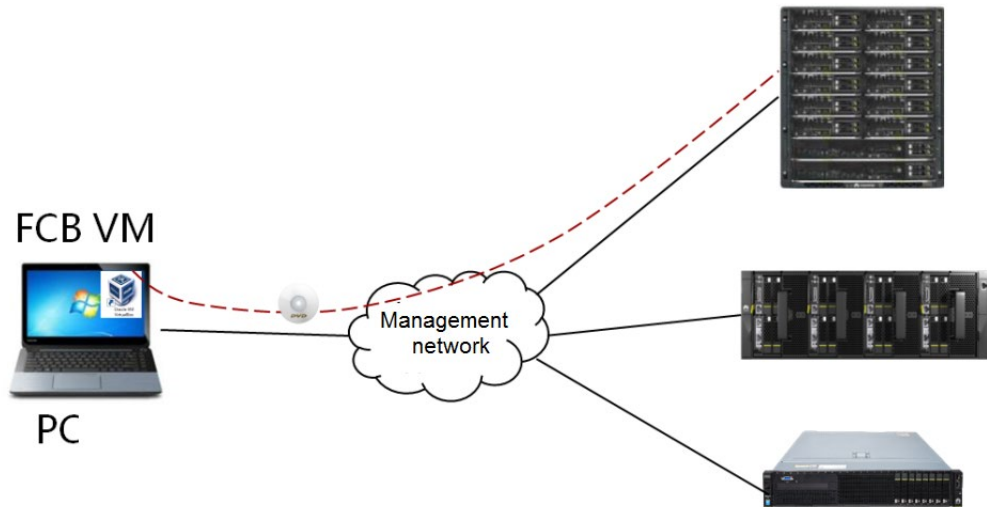
6.1 Automatic Deployment

In FusionSphere scenarios, FusionCube supports preinstallation of system software in factory environments. In other scenarios, FusionCube provides the quick installation and deployment tool FusionCube Builder (FCB for short) to install system software. FusionCube supports one-click system initialization. After basic system parameters are configured, the network configuration of each node is automatically completed, and management and storage clusters are created.

6.1.1 FusionCube Builder

FusionCube Builder (FCB) is a tool used to quickly install and deploy FusionCube system on site.

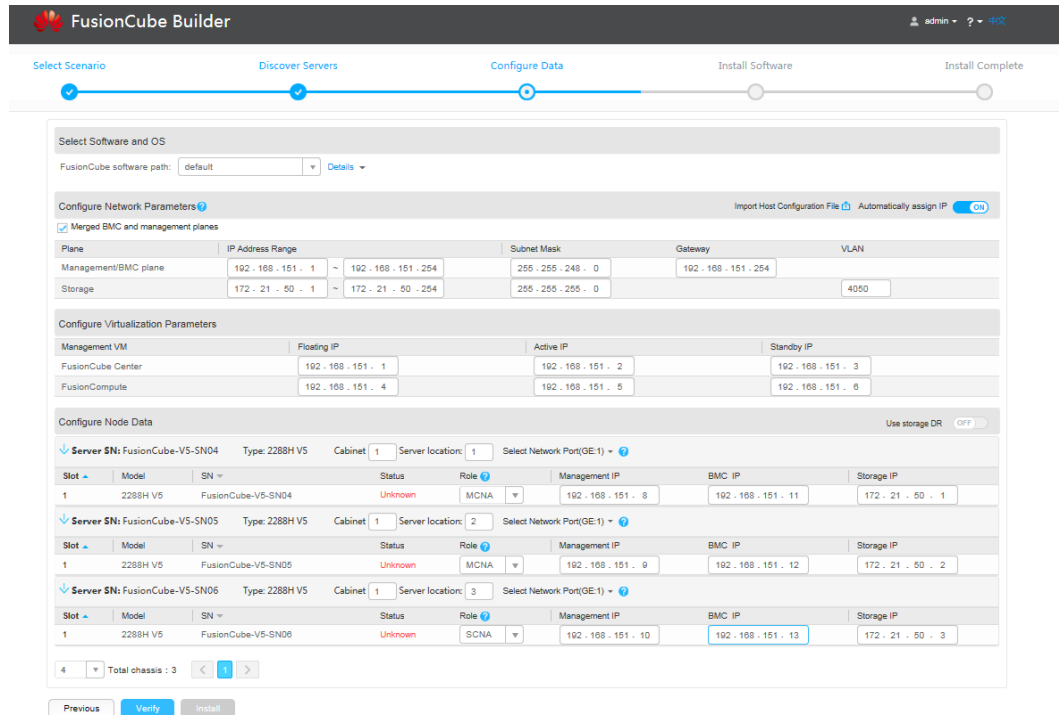
Figure 6-1 Software installation using FCB



- FCB can be installed on a PC or a virtual machine (VM).
- FCB discovers server by using the Simple Service Discovery Protocol (SSDP) or scanning IP address, and obtains the server information.
- FCB can connect to the server BMC, uses the KVM to mount the DVD drive to start the installation. A maximum of eight nodes can be installed concurrently.
- During the installation, the Network File System (NFS) is used to share the software packages and configuration.

FCB provides a unified installation and configuration interface to help users quickly complete data configuration. Then, FCB automatically completes software installation based on the data configuration.

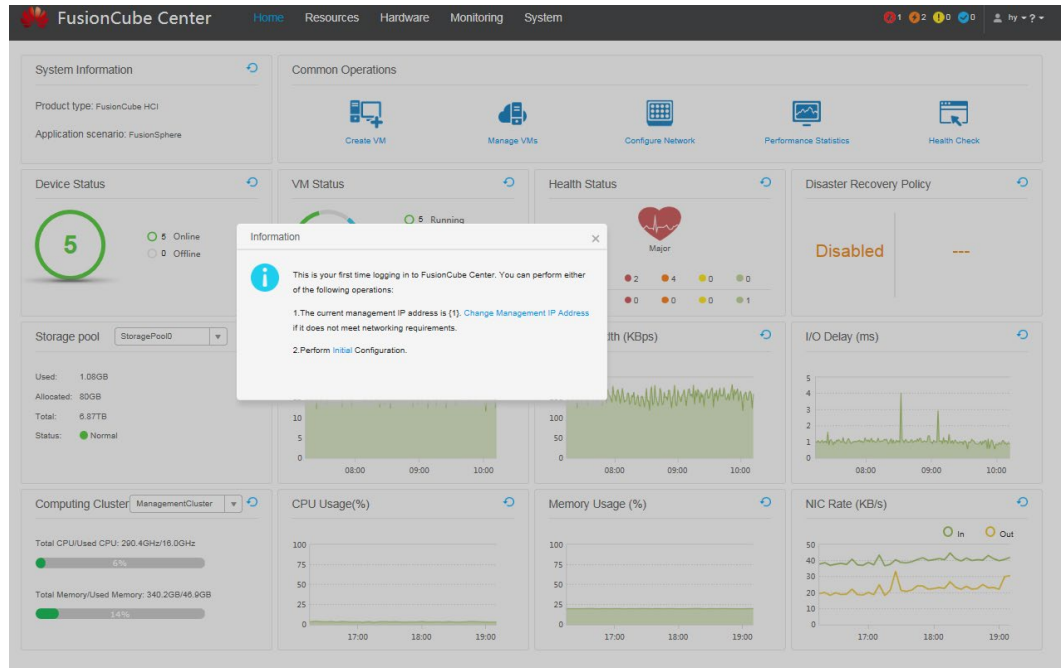
Figure 6-2 FCB installation wizard and configuration interface



6.1.2 System Initialization

When you log in to FusionCube Center for the first time, you can change the management IP address and perform system initialization.

Figure 6-3 First login to the FusionCube Center WebUI



The system initialization process is as follows:

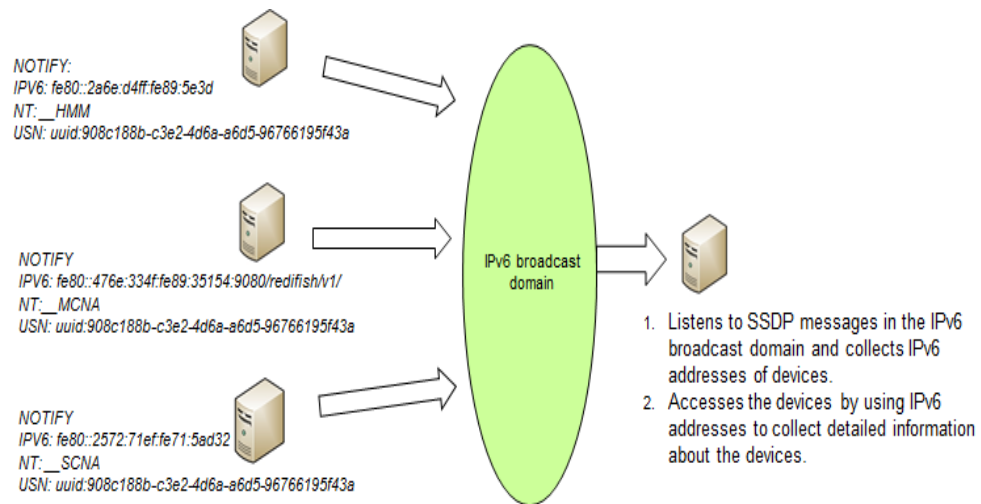
1. The system automatically discovers devices (including physical nodes and management VMs) and displays node information on the initialization parameter configuration page.
2. The user configures initialization parameters (including network parameters and storage pool parameters). After the parameter verification is successful, the system starts the initialization.
3. When the initialization is complete, the node network configuration is complete and the management cluster and storage cluster, and storage pool are created. In FusionSphere scenarios, computing clusters are also created and hosts are added to the computing clusters.

After the system initialization is complete, VMs can be provisioned. After the uplink network and NTP are configured, the FusionCube system can be used.

6.1.3 Automatic Device Discovery

FusionCube supports automatic device discovery during system installation, initialization, and capacity expansion. The SSDP is used to implement automatic device discovery.

Figure 6-4 Automatic device discovery using SSDP



The automatic device discovery process during the FCB installation is as follows:

1. The SSDP is embedded in the BMCs or MMs of the servers used by FusionCube. After the servers are powered on, SSDP messages containing the server IPv6 addresses will be automatically broadcast.
2. The SSDP server is deployed in FCB to monitor SSDP messages in the IPv6 broadcast domain and collect the IPv6 addresses of the servers.
3. FCB accesses the servers using the IPv6 addresses and obtains detailed information about the servers.

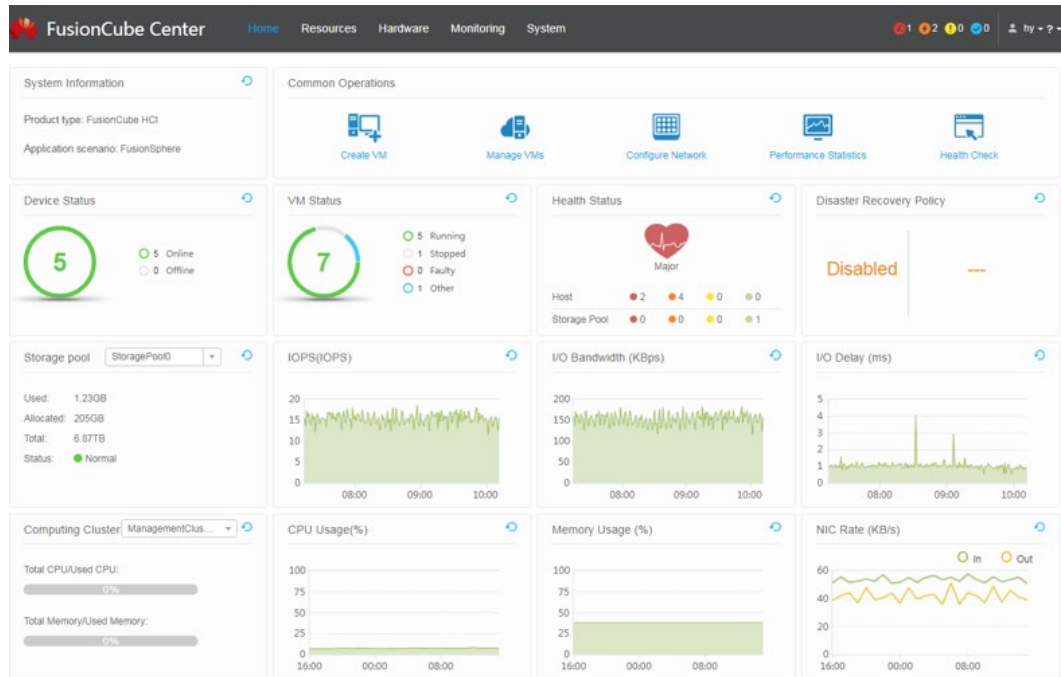
The automatic device discovery process during the system initialization and capacity expansion is as follows:

1. After the system is installed, the SSDP client is installed in the management VMs, Controller VMs (CVMs), and host OSs. The SSDP clients automatically broadcast the IPv6 address in SSDP messages.
2. The SSDP server is deployed in FusionCube Center to monitor SSDP messages in the IPv6 broadcast domain of the management plane and collect the IPv6 addresses of the servers.
3. FusionCube Center accesses the servers using the IPv6 addresses and obtains detailed information about the servers.

6.2 Unified O&M

FusionCube Center implements unified management of FusionCube. The unified management includes resource management, performance monitoring, alarm management, operation log management, rights management, hardware management, health check, and log collection.

Figure 6-5 FusionCube Center WebUI



The FusionCube Center WebUI consists of the following modules:

- Home
 - Provides information about alarms, capacity, performance, health, and tasks of the system and shortcuts for common operations.
- Resources
 - Monitors and manages resources including virtualization, storage, network, database, and VMware storage resources. The network management and database management are available only in FusionSphere scenarios. The VMware storage configuration is available only in VMware vSphere scenarios.
 - In FusionSphere scenarios, you can create management VMs on the **Virtualization** tab page. On the **Storage** tab page, you can create and mount VM volumes. On the **Network** tab page, you can configure VLAN pools and port groups. On the **Database** page, you can create and manage database nodes in hybrid scenarios.
 - In the VMware scenario, after the virtualization platform is added to vCenter, VMs can be monitored but not be provisioned or operated. On the **Storage** tab page, you can monitor the storage pool and log in to the storage management page using SSO. On the **VMware Storage Configuration** tab page, you can create volumes, mount the volumes to the ESXi hosts, and configure paths.
- Hardware
 - Monitors the hardware devices, including chassis, servers, and switches, used in the system. Displays the hardware device information, including node type, model, IP addresses, hardware information, and resource information.
- Monitoring
 - Implements alarm management and performance monitoring. It provides alarm list for all the components in the system, alarm statistics, and alarm settings. Performance monitoring data includes historical performance data and top performance statistics.

- System
 - Provides system configuration, rights management, task and log management, and system maintenance.
 - The system configuration provides time management, management data backup, email server configuration, eService configuration, SNMP configuration, desktop cloud access, and system timeout period.
 - Rights management includes management of users, user roles, password policies, and domain authentication.
 - The system maintenance provides one-click capacity expansion, health check, log collection, and upgrade.

6.2.1 Service Provisioning and Management

FusionCube HCI 3.2 supports VM-related service provisioning and management only in FusionSphere scenarios. The service provisioning and management include:

- VM provisioning and management
- Disk management
- Network management

VM Provisioning and Management

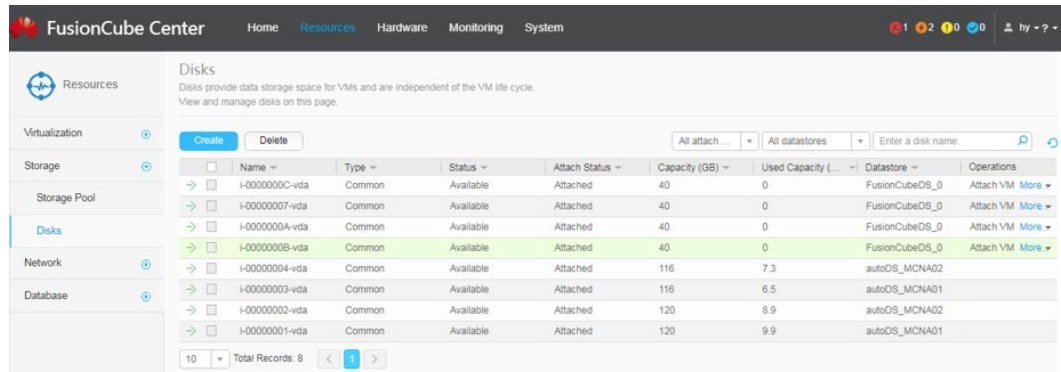
FusionCube Center provides VM provisioning and management for FusionSphere scenarios. The VM provisioning and management includes the following:

- VM creation and management (including powering on/off, restarting, migrating, and importing and exporting VMs)
- Adjustment of VM specifications
- VM performance monitoring
- Snapshot management
- VM template management

Name	ID	IP Address	Host	Status	CPU Usage	Memory Usage	Disk Usage	Operations
test-1	i-0000000B	192.170.58.170		Stop...	0%	0%	0%	
test-2	i-00000007	192.170.58.171	SCNA04	Runn...	0%	0%	0%	
test-3	i-0000000A	192.170.58.173		Hiber...	0%	0%	0%	
FCC02	i-00000004	192.170.58.30	MCNA02	Runn...	69.74%	51.02%	5.94%	
FCC01	i-00000003	192.170.58.31	MCNA01	Runn...	0.5%	6.3%	5.5%	
VRM02	i-00000002	192.170.58.164	MCNA02	Runn...	3%	15.59%	7.38%	
VRM01	i-00000001	192.170.58.163	MCNA01	Runn...	18.18%	80.5%	8.13%	

Disk Management

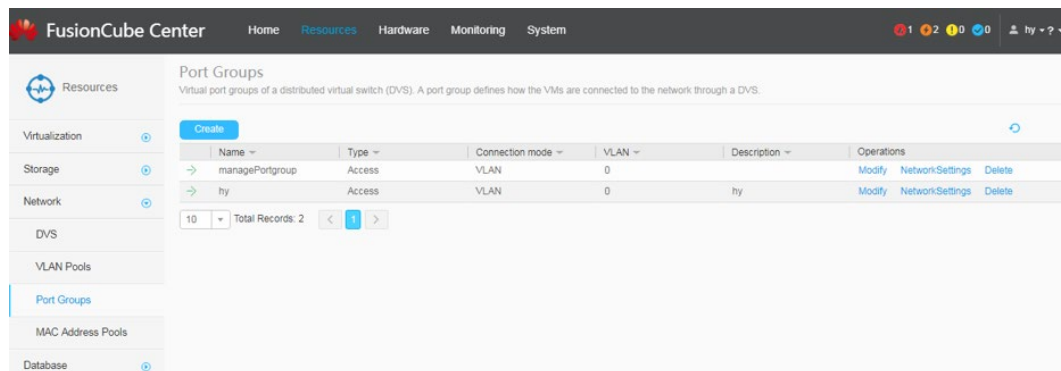
VM disk management includes creation of volumes and binding of volumes and hosts. The system provides common and shared volume devices supporting IDE, VIRTIO, and SCSI interfaces. By default, thin provisioning volumes are provided to improve disk utilization of the system.



Network Management

Network management allows network resources, such as the VLANs, port groups, and MAC addresses, to be configured for VM provisioning.

FusionCube Center supports creation and configuration of VLAN pools, port groups, and MAC address pools, and query of distributed switches (DVSs). The creation and management of DVSs can be performed only on the FusionCompute virtualization platform.



6.2.2 One-Click O&M

FusionCube Center provides one-click O&M functions to simplify O&M operations and improve O&M efficiency. The one-click O&M includes one-click capacity expansion, health check, log collection, and upgrade.

One-Click Capacity Expansion

The capacity expansion process is as follows:

- Step 1** Install the hardware and connect the newly added nodes to the system network.

Install the nodes, connect cables, start the nodes, configure the network data to connect the nodes to the system network.

In FusionSphere scenarios, the nodes have been installed before delivery. In the VMware scenario, use FusionCube Builder to install the OSs on the nodes to be added.

- Step 2** On the FusionCube Center WebUI, perform capacity expansion.

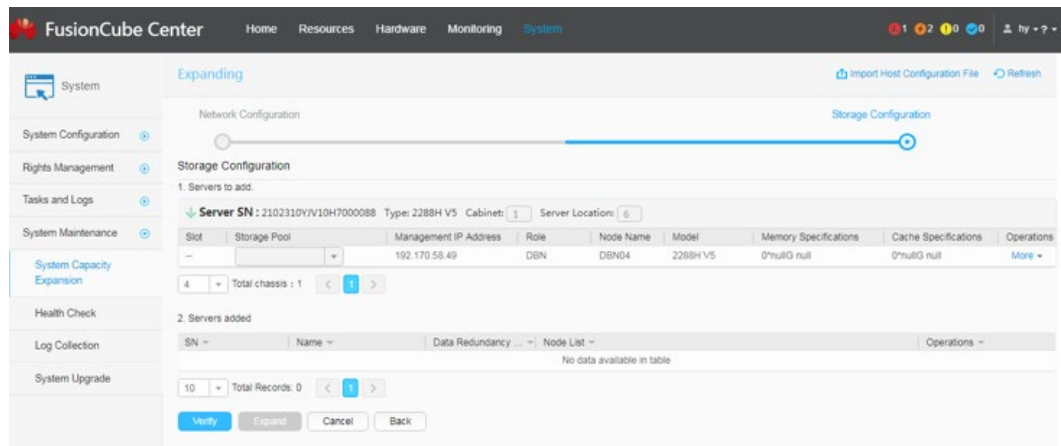
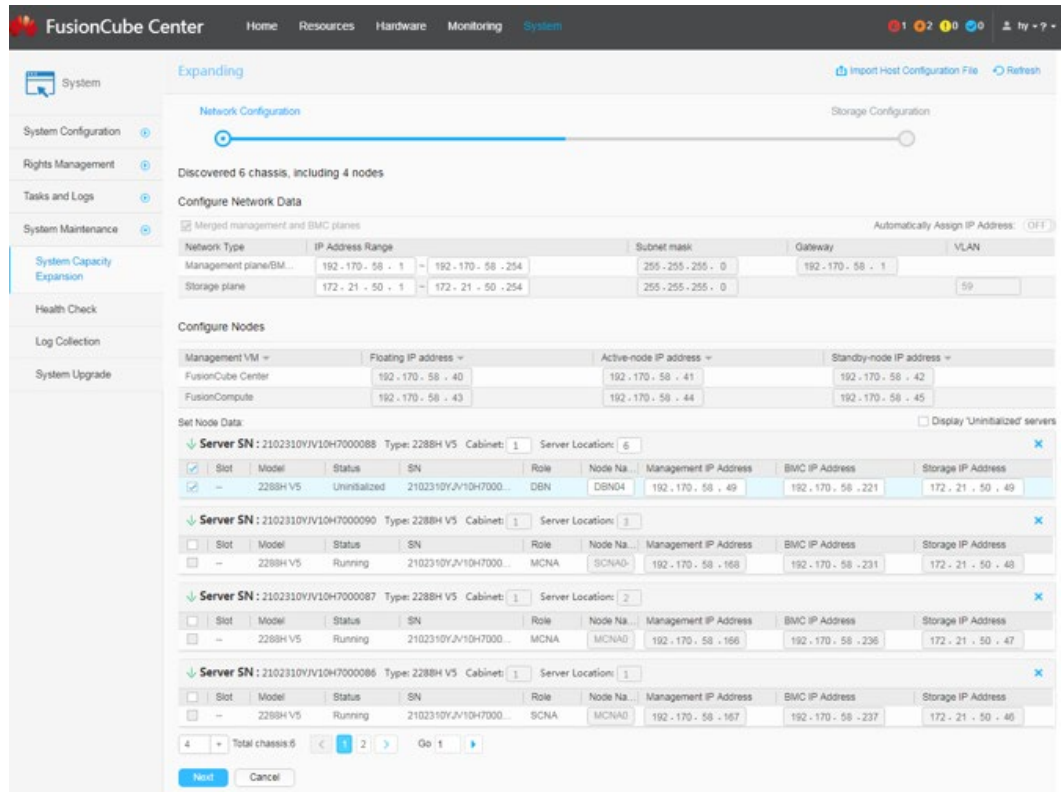
1. Choose **System > Capacity Expansion**.

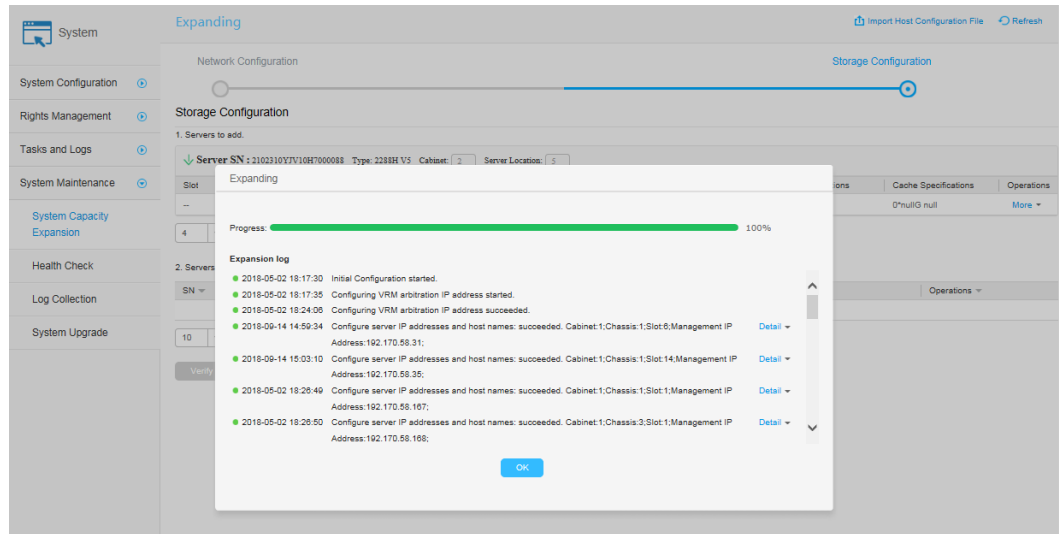
The system automatically discovers the nodes added.

2. Configure the IP addresses, host names, gateway, and storage pool for the nodes added.

3. Click **Verify** to verify the data configured.
4. If the verification is successful, click the **Capacity Expansion** button to add the nodes to the system cluster.

---End



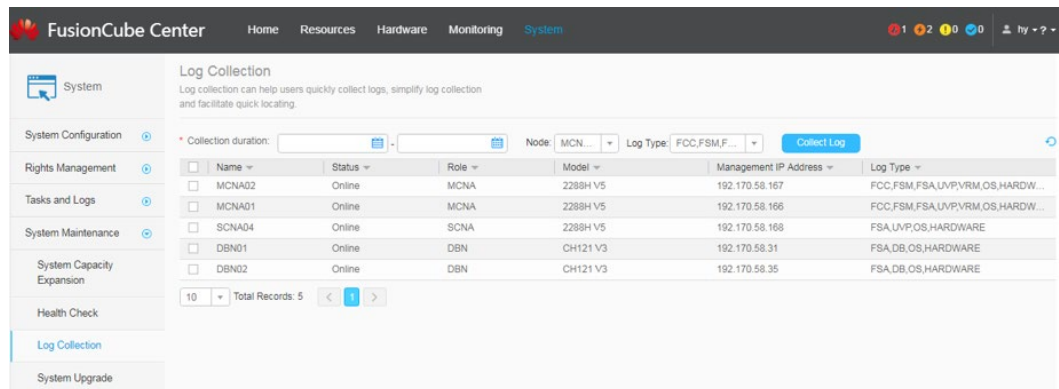


One-Click Log Collection

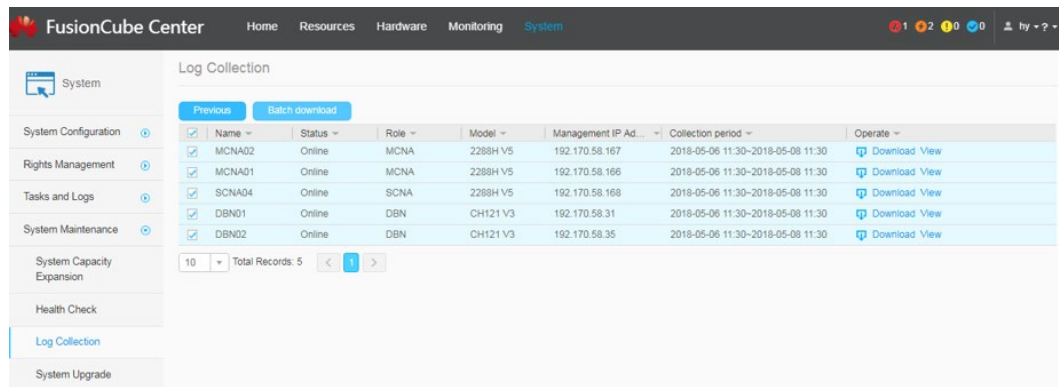
The FusionCube Center WebUI provides one-click log collection. Users can collect logs of all components or logs of specified components as required.

- FusionCube Center supports collection of the logs of the hardware (BMC, SSD, NICs, and IB components), distributed storage, FusionCompute, OSs, ESXi, FusionCube Center, and Oracle RAC.
- Only log files of two days can be collected at a time. Logs of multiple nodes can be collected at the same time.

On the FusionCube Center WebUI, choose **System** > **System Maintenance** > **Log Collection**, set the time segment, node type, log type, and nodes to be collected, and click **Collect Log**.



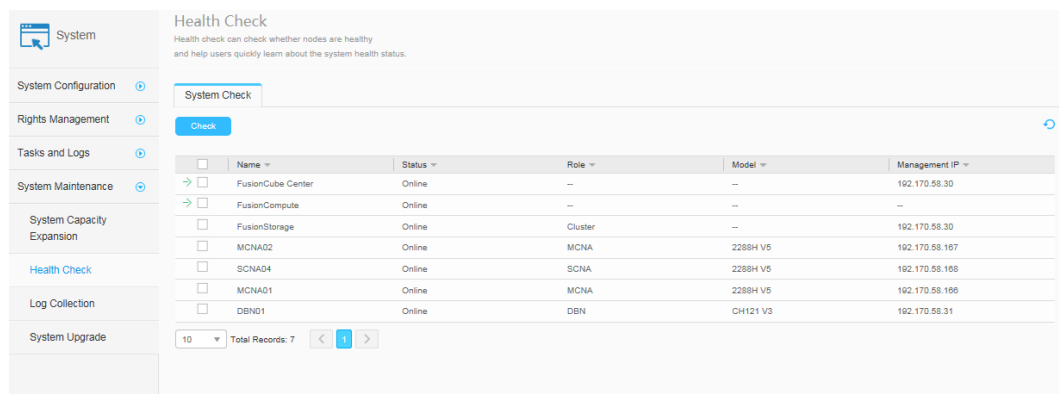
After collecting logs, you can download and analyze the logs.



One-Click Health Check

FusionCube Center provides one-click health check to check health status of system components and nodes, detects health risks or faults, and provides check reports and troubleshooting suggestions. The health check includes system check and hardware compatibility check. The system check includes health checks of the distributed storage, virtualization system, and hardware. The hardware compatibility check verifies whether the hardware, driver, and firmware versions meet the FusionCube version requirements.

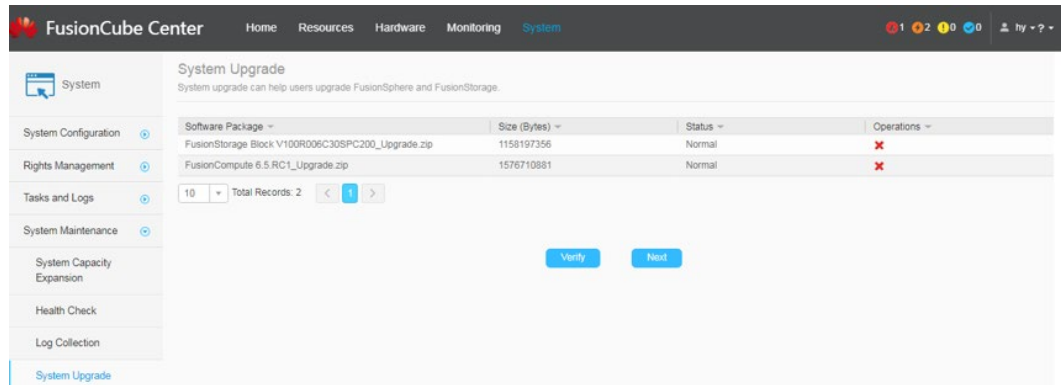
On the FusionCube Center WebUI, choose **System > System Maintenance > Health Check**, select the check to be performed and the components and nodes to be checked, and click **Check**.



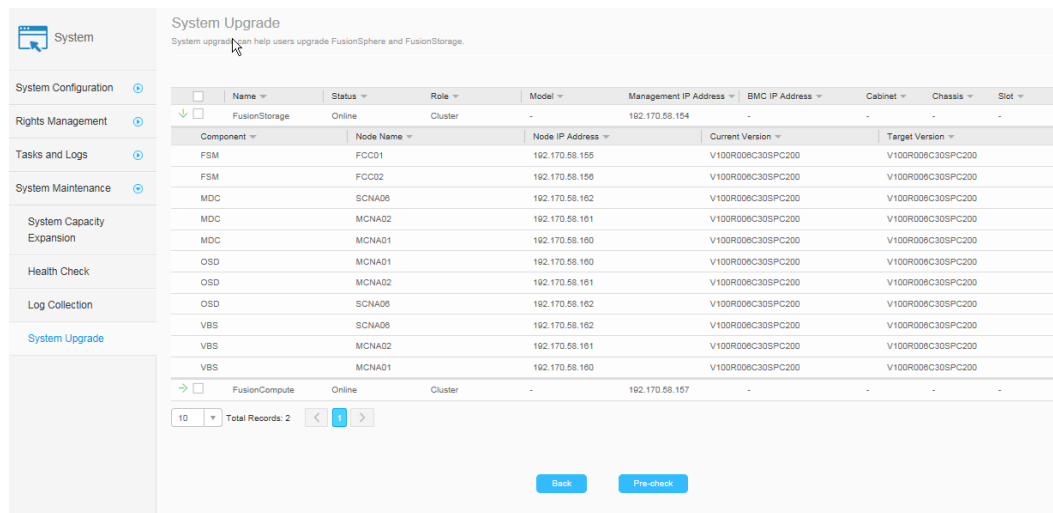
One-Click Upgrade

FusionCube Center provides one-click upgrade. The one-click upgrade supports upgrade of the system management platform, virtualization platform, distributed storage components, as well as the hardware devices, including the BMC, driver, and firmware of the hosts, NICs, SSDs, IB components, and RAID controller cards.

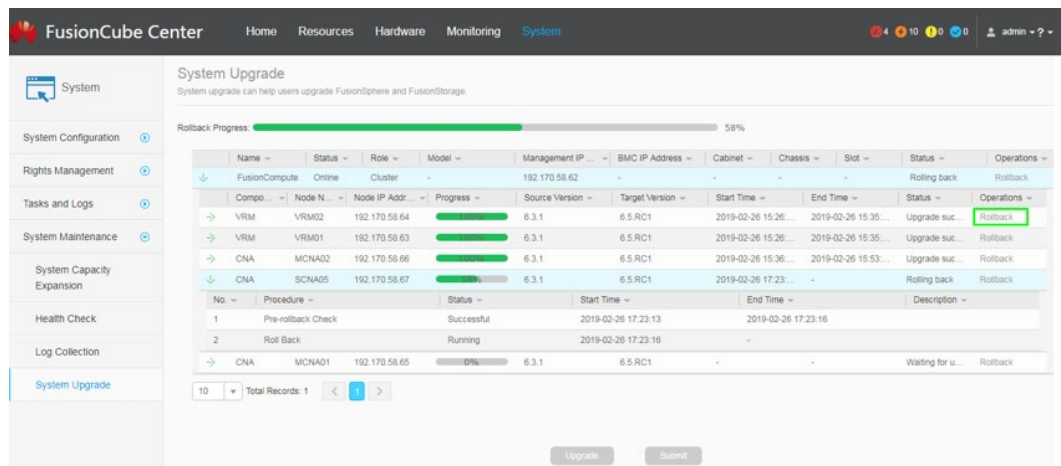
On the FusionCube Center WebUI, choose **System > System Maintenance > Upgrade**, upload the update package, and click **Verify**.



Select the software components or nodes to be upgraded, and perform a pre-upgrade check.

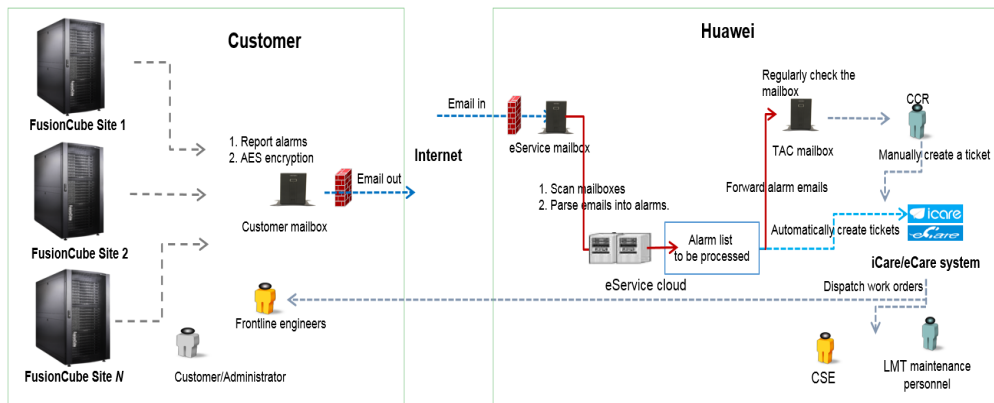


If the pre-check is successful, perform the upgrade.



6.2.3 Call Home

FusionCube Center provides the Call Home function, which allows the FusionCube alarms to be sent to Huawei Global Technical Assistance Center (GTAC) through the eServer. After receiving the alarm information, the GTAC personnel determines whether the fault needs to be handled immediately and provides related suggestions to the customer in a timely manner.



7 Performance and Scalability

[7.1 High Performance](#)

[7.2 Linear Expansion](#)

[7.3 Advantages of FusionCube Distributed Storage over Conventional SAN](#)

7.1 High Performance

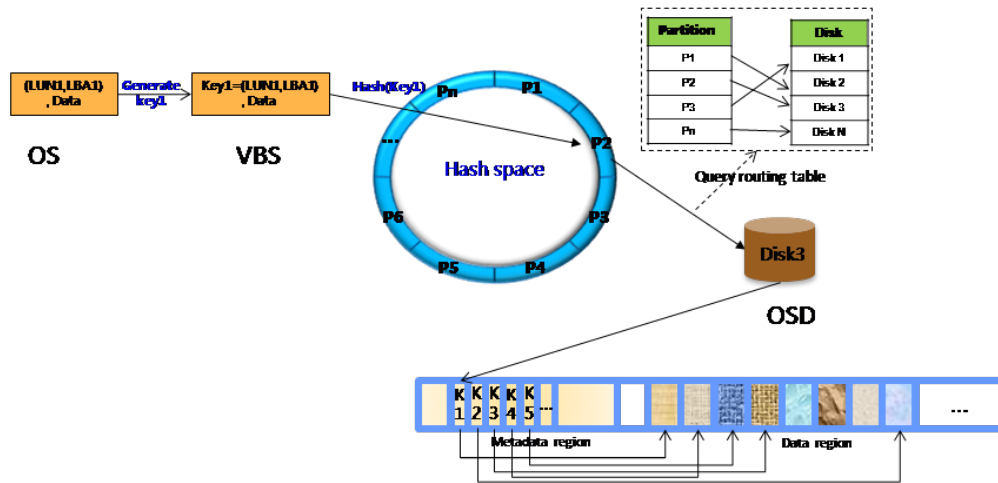
FusionCube distributed storage uses a distributed large resource pool architecture and does not have centrally accessed components or modules, eliminating performance bottlenecks caused by a single component or module. The distributed hash algorithm is used to evenly distribute service data to all disks in the resource pool, preventing a component or module from becoming a hotspot. All disks in the resource pool can be used as hot spare disks of the resource pool. If a component or hardware fault occurs, data can be quickly restored, ensuring service performance consistency. In the FusionCube distributed storage system, the SSDs on all storage nodes form a shared distributed SSD cache resource pool, which improves the I/O performance of the system.

In addition, FusionCube uses NVMe SSDs to offer higher performance and supported InfiniBand network devices to shorten network latency.

7.1.1 Distributed I/O Ring

The FusionCube distributed storage uses the DHT routing technology to rapidly locate the data on the hard disk for service I/O. This prevents searching and calculation of a great amount of data. The DHT routing technology uses Huawei-developed algorithms to ensure balanced data distribution among hard disks and rapid, automatic data adjustment when the hardware quantity increases (due to capacity expansion) or decreases (due to hardware faults). The DHT technology also ensures data migration validity, rapid automatic self-healing, and automatic resource balancing.

Figure 7-1 Data routing of the FusionCube distributed storage



During system initialization, the FusionCube distributed storage system sets partitions for each disk based on the value of N and the number of hard disks. For example, the default value of N is **3600** for two-copy backup. If the system has 36 hard disks, each hard disk has 100 partitions. The partition-disk mapping is configured during system initialization and dynamically adjusted based on the number of the hard disks. The mapping table requires only small space, and FusionStorage nodes store the mapping table in the memory for rapid routing.

FusionCube distributed storage logically slices each LUN based on the 1 MB size. For example, the LUN of 1 GB is sliced into 1024 x 1 MB slices. When an application accesses FusionStorage, the SCSI command carries the LUN ID, LBA ID, and data to be read/written. The OS forwards the message to the VBS of the local node. The VBS generates a key based on the LUN ID and LBA ID. The key contains rounding information of the LBA ID based on the unit of 1 MB. The result calculated using DHT hash indicates the partition. The specific hard disk is located based on the partition-disk mapping recorded in the memory. The VBS forwards the I/O operation to the OSD to which the hard disk belongs.

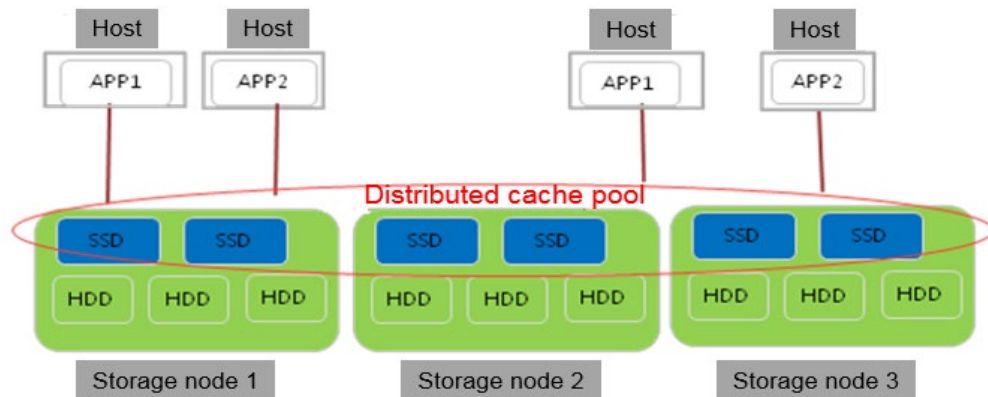
Each OSD manages a hard disk. During system initialization, the OSD divides the hard disk into slices of 1 MB and records the slice allocation information in the metadata management area of the hard disk. After receiving the I/O operation sent by the VBS, the OSD locates the data slice based on the key, obtains the data, and returns the data to the VBS.

7.1.2 Distributed SSD Cache Acceleration

Due to mechanical limitations, the performance of HDDs is basically unchanged for decades although the capacity increases greatly. The random I/O latency, ranging from several milliseconds to tens of milliseconds, severely affects user experience and system performance. Compared with HDDs, SSDs offer higher performance and also higher costs. Nowadays, SSDs are used as the system cache or tier layer to balance performance and costs.

The SSDs on each storage node of FusionCube form a cache resource pool for all services. In this way, SSD resources are fully utilized.

Figure 7-2 Distributed cache of FusionCube distributed storage



7.1.2.1 Read/Write Cache

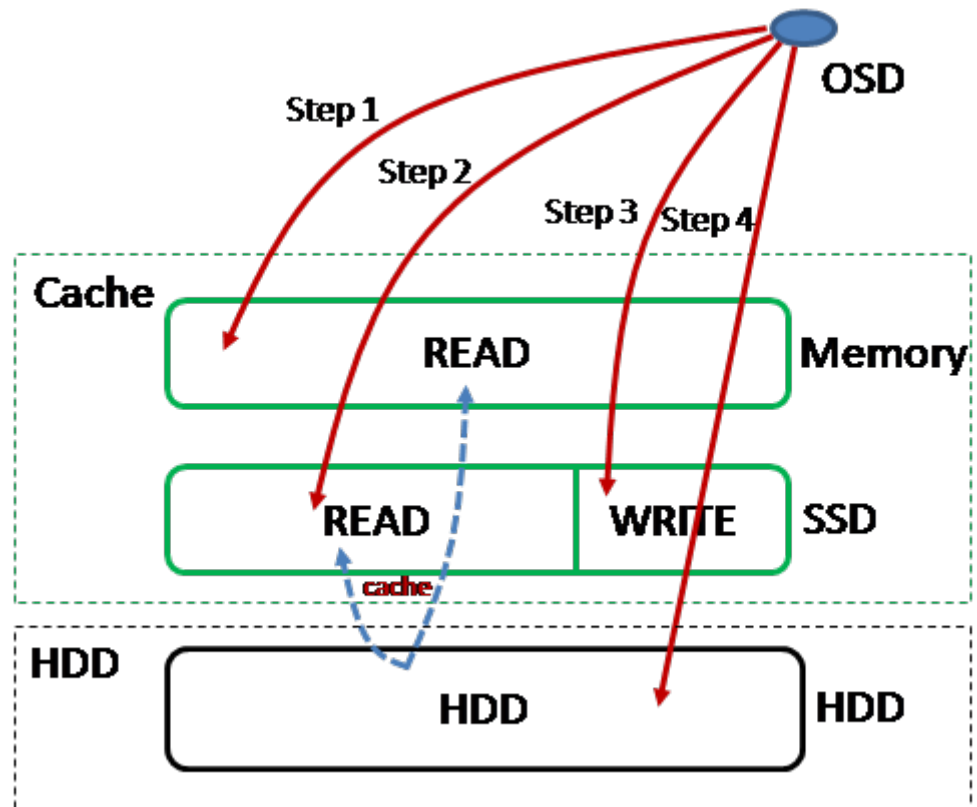
Read Cache

The FusionStorage block storage adopts a hierarchical mechanism for read cache. The first layer is the memory cache, which caches data using the LRU mechanism. The second layer is the SSD cache, which functions based on the hotspot read mechanism. The system collects statistics on each piece of read data and the hotspot access factor. When the threshold is reached, the system automatically caches data to the SSD and removes the data, which has not been accessed for a long time, from the SSD.

When receiving a read I/O request from the VBS, the OSD performs the following operations:

1. Search the read cache of the memory for the required I/O data.
 - If the I/O data is found, return it to the VBS and move the I/O data to the LRU queue head of the read cache.
 - If no, go to 2.
2. Search the read cache of the SSD for the required I/O data.
 - If the I/O data is found, return it directly in a copy scenario. If EC is used, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and add the hotspot access factor of the I/O data.
 - If no, go to 3.
3. Search the write cache of the SSD for the required I/O data.
 - If the I/O data is found, return it directly in a copy scenario. In an EC scenario, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and add the hotspot access factor of the I/O data. If the hotspot access factor reaches the threshold, the I/O data is recorded to the read cache of the SSD.
 - If no, go to Step 4.
4. Search the hard disk for the required I/O data. Return the data directly if multi-copy backup is used. If EC is used, obtain data blocks on other nodes, combine data by using the EC algorithm, return data, and add the hotspot access factor of the I/O data. If the hotspot access factor reaches the threshold, the I/O data is recorded to the read cache of the SSD.

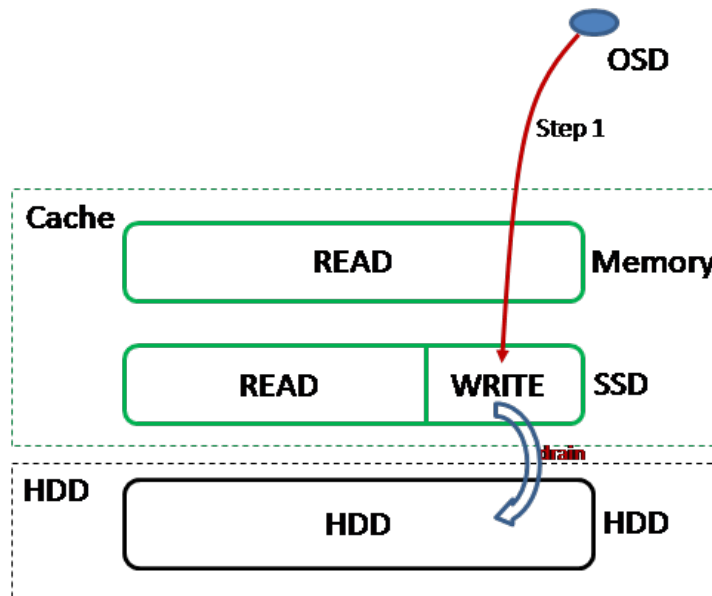
Figure 7-3 Read cache mechanism of the FusionCube distributed storage



Write Cache

When receiving a write I/O operation sent by the VBS, the OSD temporarily stores the write I/O in the SSD cache to complete the write operation on the local node. At the same time, the OSD processes I/Os in the background and rearranges the I/Os, and then writes the I/Os to the hard disk.

Figure 7-4 Write cache mechanism of the FusionCube distributed storage



7.1.2.2 Pass-Through of Large Blocks

The following table provides performance comparison of different media. For the random small I/O, SSDs provide a performance advantage of tens to hundreds of times than HDDs. However, the advantages in sequential I/O are not obvious.

Table 7-1 Performance comparison

Medium Type	4 KB Random Write IOPS	4 KB Random Read IOPS	1 MB Write Bandwidth	1 MB Read Bandwidth	Average Delay (ms)
SAS	180	200	150 MB	150 MB	3 to 5
NL-SAS	100	100	100 MB	100 MB	7 to 8
SATA	100	100	80 MB	80 MB	8 to 10
SSD disk	70,000	40,000	500 MB	500 MB	< 1
SSD card	600,000	800,000	2 GB	3 GB	< 1

The HDD disk performance data is measured in the condition that the HDD write cache is disabled. For the HDDs used to set up a storage system, the write cache must be disabled to ensure reliability.

The working principle of the HDD disks is similar. So is the performance of the HDDs from different vendors. The difference in performance is less than 10%.

There is a big difference in the performance of SSD disks and SSD cards. This table uses one type of SSD disks and SSD cards as an example. The SSD disk bandwidth performance is limited by the SAS/SATA interface bandwidth. The 6 Gbit/s SATA interface is commonly used for tests.

As indicated by the preceding performance data, SSDs have obvious performance advantages over HDDs in small-block random IOPS. In large-block sequential I/O, SSD cards provide large bandwidth advantages, but SSD disks do not have obvious advantages over HDDs due to the bandwidth limit on the SAS/SATA interface. If an SSD disk serves as the cache for more than five HDDs, directly accessing HDDs provides higher performance than accessing the SSD.

Huawei FusionStorage supports bypass SSD cache in large-block I/O. This feature provides the following advantages:

Higher performance in large-block I/O operations. The cache resources originally occupied by large-block I/O operations are released, and more small-block I/O operations can be cached. This increases the cache hit rate of small-block random I/O operations, enhances the overall system performance, allows more write I/O operations, and extends the service life of SSD cards.

7.1.3 Hardware Acceleration

SSD Disks/Cards

FusionCube provides all-flash SSD storage pools for high-performance applications. It uses Huawei ES3000 SSD V5 devices (including NVMe SSD disks/cards and SAS SSD disks) to provide higher read and write performance over traditional mechanical SATA/SAS disks.

The performance specifications of the ES3000 NVMe SSD V5 are as follows:

- Capacity: 800 GB/1.2 TB/1.6 TB/3.2 TB/6.4 TB
- IOPS: Max. read IOPS: 825,000@4KB; Max. write IOPS: 300,000@4KB
- Bandwidth: PCIe 3.0 x4 interface with a max. bandwidth of 3500 MBps
- Low latency: Huawei-developed Hi812E chip and multiple logical hardware collaboration tools reduce the read and write latency.

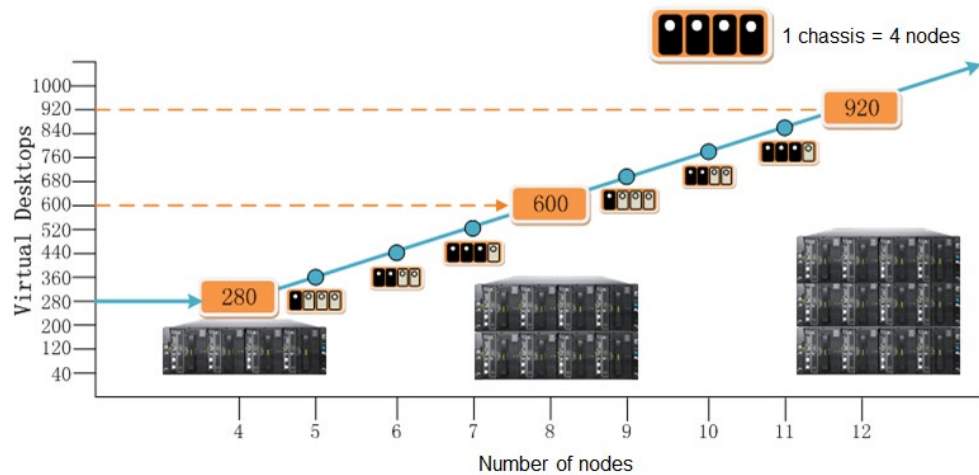
The performance specifications of the ES3000 SAS SSD V5 are as follows:

- Capacity:
ES3600S V5: 800 GB/1.2 TB/1.6 TB/3.2 TB/6.4 TB
ES3500S V5: 960 GB/1.92 TB/3.84 TB/7.68 TB
- IOPS:
Max. read IOPS: 430,000@4KB
Max. write IOPS: 155,000@4KB
- Bandwidth:
Max. read bandwidth: 2200 MB/s
Max. write bandwidth: 1750 MB/s
- Low latency:
Huawei-developed Hi812E chip and multiple logical hardware collaboration tools reduce the read and write latency.

7.2 Linear Expansion

FusionCube has good scalability. A FusionCube system has at least three nodes. It can be easily expanded by adding servers or nodes. A cluster supports a maximum of 256 nodes.

Figure 7-5 FusionCube system expansion

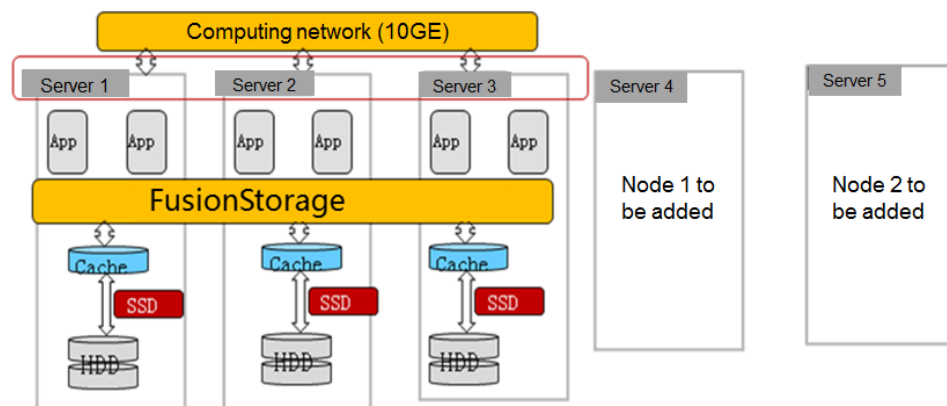


7.2.1 Smooth Storage Capacity Expansion

The distributed architecture of FusionCube distributed storage has good scalability and supports ultra-large storage capacity.

- Load balancing can be rapidly achieved after the capacity expansion without migrating a large amount of data.
- Computing nodes, hard disks, and storage nodes can be expanded separately or at the same time. Compute nodes and storage capacity can be added at the same time. After capacity expansion, computing and storage resources are still converged.
- Software engines, storage bandwidth, and cache are evenly distributed to every node. The system IOPS, throughput, and cache linearly increase as nodes are added.

Figure 7-6 Smooth storage capacity expansion



7.2.2 Linear Performance Expansion

The FusionCube distributed storage uses an innovative architecture to organize SATA HDDs into a storage pool like SAN providing a higher I/O than SAN devices and maximizing performance.

Distributed Engine

FusionCube distributed storage uses stateless software engines deployed on each node, eliminating the performance bottleneck of centralized engines. These software engines on nodes consume only a few CPU resources and provide higher IOPS and throughput than centrally deployed engines. For example, the system has 20 nodes that need to access storage resources provided by FusionStorage, and each node provides 2 x 10 Gbit/s bandwidth for the storage plane. If each node is deployed with a VBS module (a storage engine), there are 20 storage engines in the system and the total throughput can reach 400 Gbit/s (20 x 2 x 10 Gbit/s = 400 Gbit/s). The storage engines can be added linearly as the cluster capacity is expanded. This breaks the performance bottlenecks caused by centrally deployed engines in traditional dual-controller or multi-controller storage systems.

Distributed Cache

The cache and bandwidth of FusionCube distributed storage are evenly distributed to each node.

In the conventional storage system, a large number of disks share the limited bandwidth between computing devices and storage devices. In FusionCube, the hard disks of each node in the storage cluster use independent I/O bandwidth.

In the FusionCube distributed storage system, the nodes use some memory as the read cache and use the SSDs as the write cache. The data cache is evenly distributed to each node. The total cache capacity for nodes is far greater than the cache capacity provided by an external independent storage device. Even if large-capacity, low-cost SATA disks are used, the FusionCube distributed storage can still provide high I/O performance and improve the overall performance by 1 to 3 times.

The FusionCube distributed storage supports SSDs as data cache. In addition to the common write cache, FusionCube provides hot data statistics and caching functions to improve system performance.

Global Load Balancing

The DHT mechanism of the FusionCube distributed storage allows the I/O requests from upper-layer applications to be evenly distributed to different hard disks on different nodes, eliminating local hot spots and implementing global load balancing.

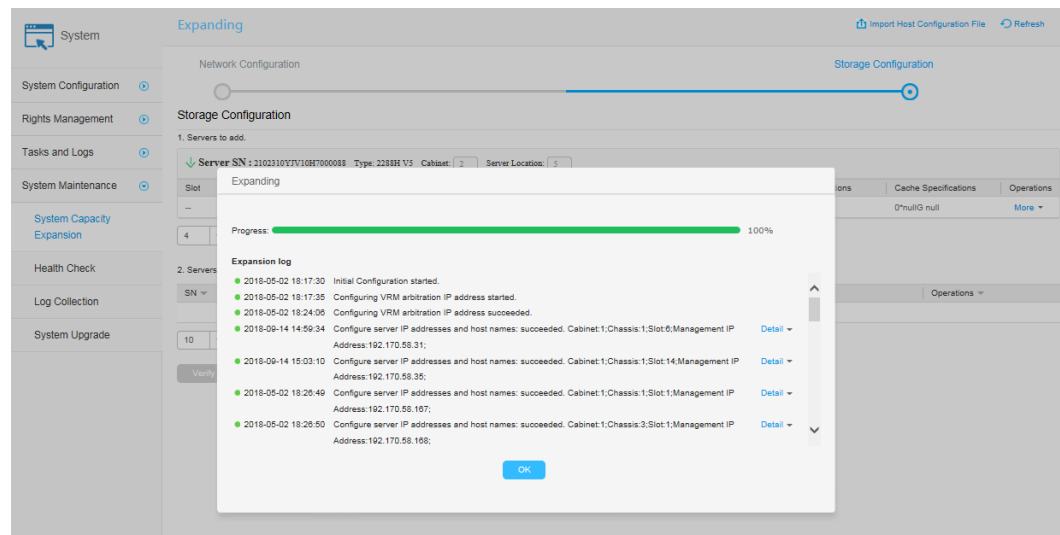
- The system automatically disperses data blocks on each volume and stores them on different hard disks. As a result, the data frequently or seldom accessed is evenly distributed on different nodes to prevent hot spots.
- The data slicing and distribution algorithm enables primary and secondary data copies to be evenly distributed on different hard disks of different nodes. In this way, the number of primary copies distributed on each hard disk is equal to the number of secondary copies.
- When a node is added or deleted due to a failure, the system uses the data rebuild algorithm to balance loads among all nodes after the rebuild.

7.2.3 One-Click Capacity Expansion

FusionCube provides the one-click capacity expansion function to simplify system capacity expansion operations.

1. On the FusionCube Center WebUI, choose **System > System Maintenance > Capacity Expansion**.
2. The system automatically discovers devices and displays the discovered devices.
3. Configure network and storage parameters and click **Verify**. If the verification is successful, click the **Capacity Expansion** button.
4. The system automatically completes the expansion configuration, including configuring network settings for nodes, adding nodes to storage clusters, and expanding the storage pool or creating a storage pool, based on the data configured.

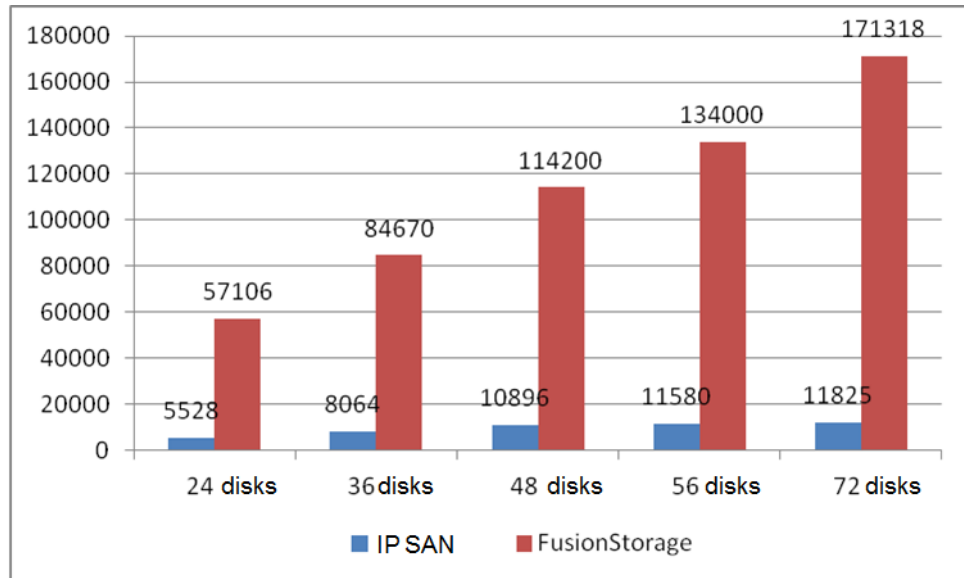
Figure 7-7 FusionCube Center capacity expansion



7.3 Advantages of FusionCube Distributed Storage over Conventional SAN

7.3.1 Higher performance

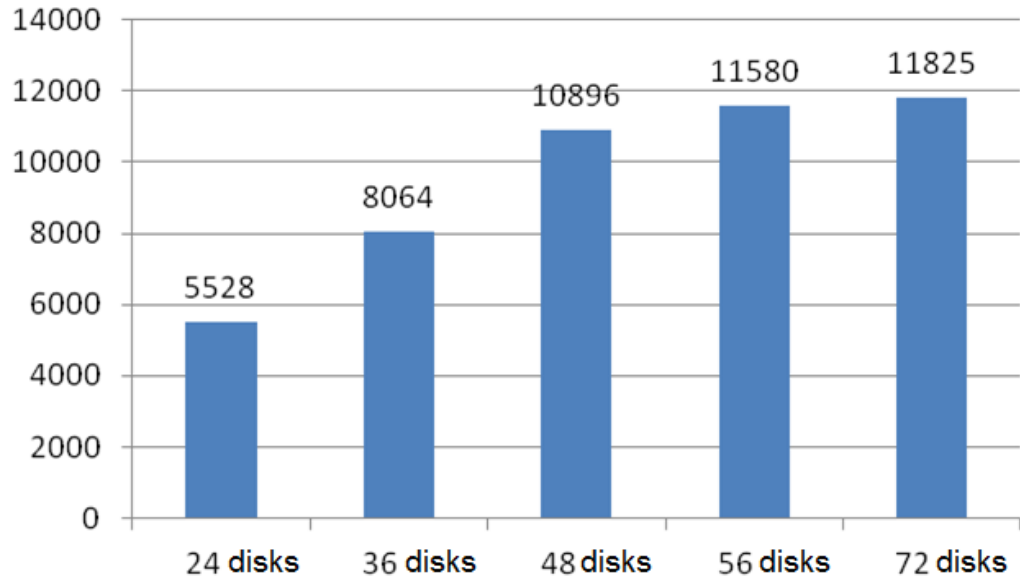
Figure 7-8 Comparison between FusionCube and IP SAN



Under the same conditions, FusionCube provides more than 10 times higher performance than the IP SAN. In addition, the performance increases with the number of disks.

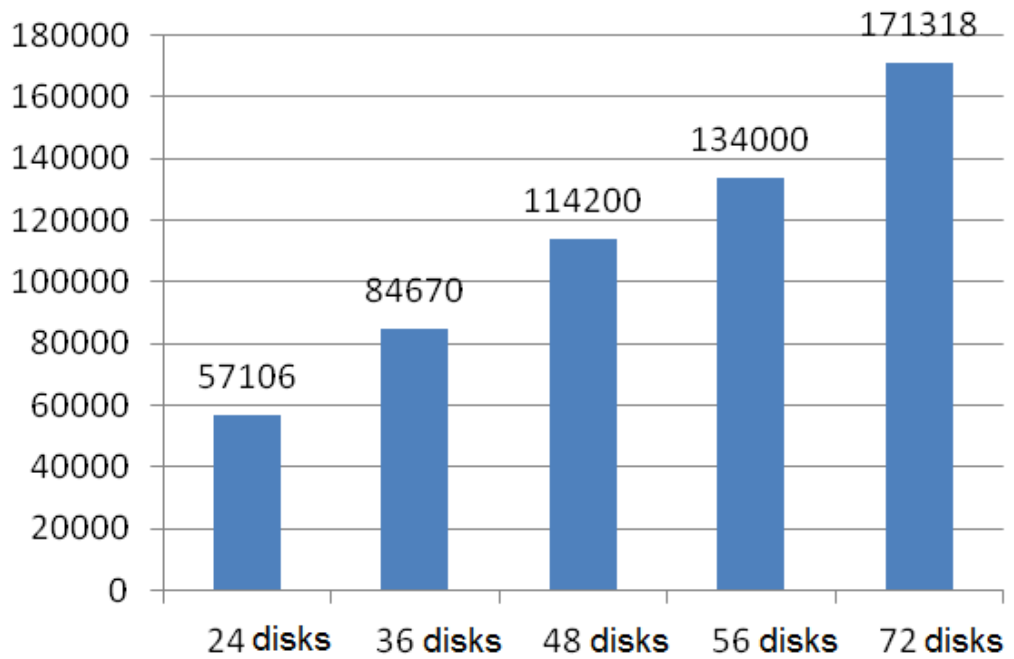
7.3.2 Linear Scale-up/Scale-out

Figure 7-9 Performance improvement with IP SAN capacity expansion



The IP SAN supports scale-up but not scale-out. In addition, the scale-up cannot ensure linear performance improvement. As shown in the preceding figure, the IP SAN support linear performance improvement within 48 disks. If the number of disks exceeds 48, the performance cannot be linearly improved due to the limitation on the engine processing capability. Therefore, the IP SAN high-performance configuration is not always configured with the maximum number of disk enclosures.

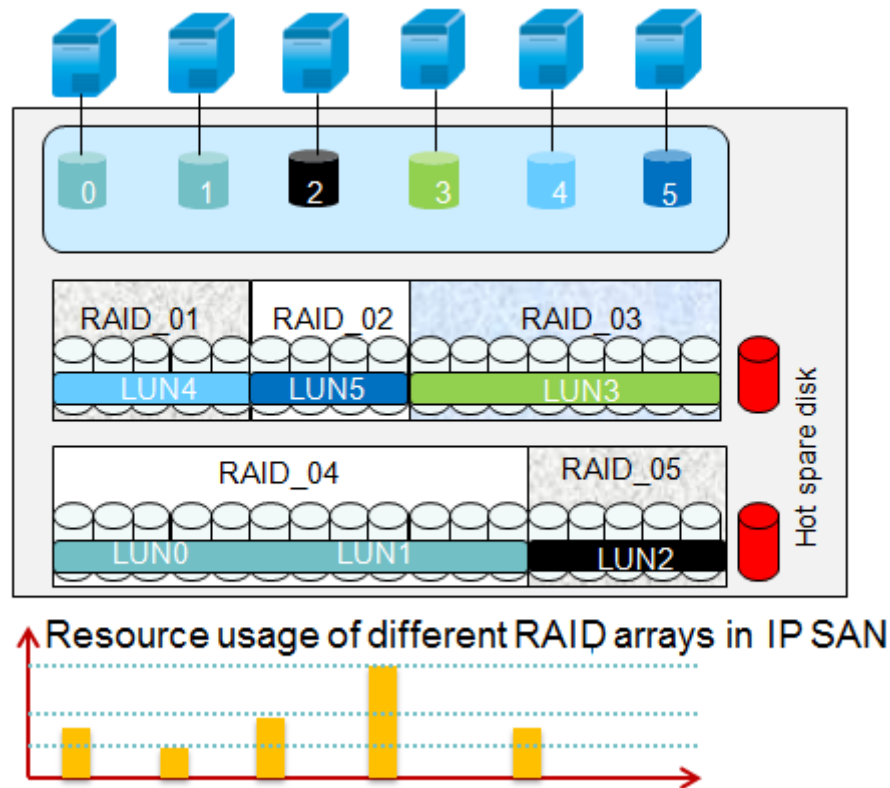
Figure 7-10 Performance improvement with FusionCube capacity expansion



The FusionCube distributed storage allows linear performance improvement as the number of disks increases.

7.3.3 Large Pool

Figure 7-11 Traditional IP SAN RAID

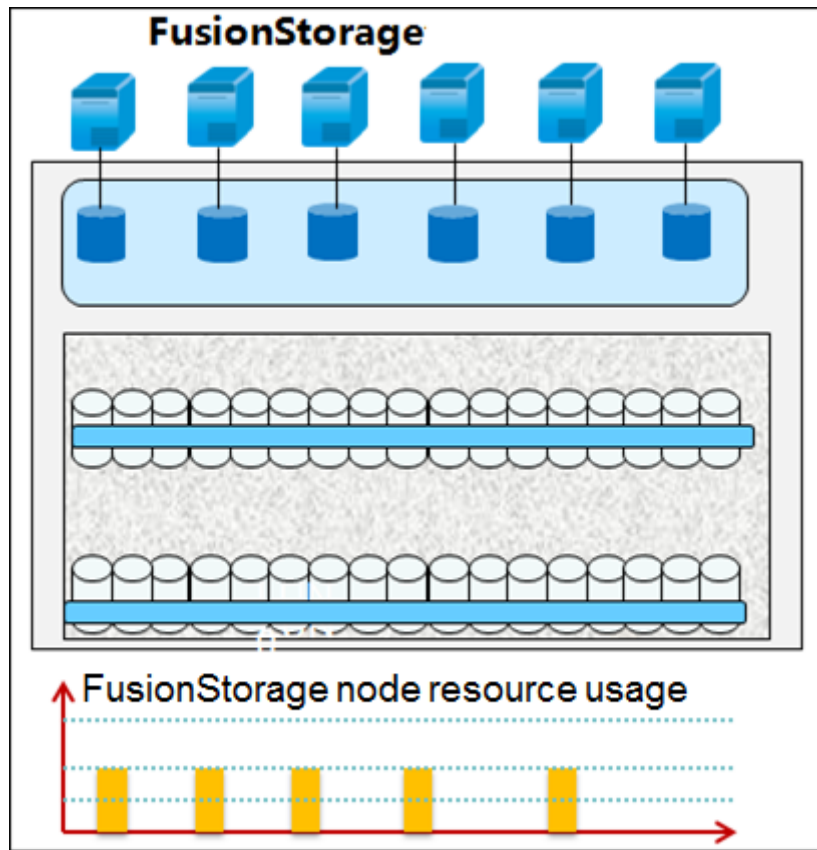


The conventional IP SAN provides services through the LUNs configured on RAID arrays. For reliability purposes, a RAID array is generally composed of a few disks. Therefore, the performance of a single RAID array is limited.

The traditional IP SAN has drawbacks in complex service planning and adjustment difficulty. If the performance cannot meet performance requirements, the LUNs need to be migrated to another RAID array. A large number of LUNs need to be adjusted when no RAID array can meet the performance requirements while the system overall performance still has allowance. The adjustment may pose risks to normal system running. In many cases, even large-scale adjustment cannot meet performance requirements and causes waste of resources. The system performance is wasted, and the performance of different RAID arrays cannot be shared. As shown in the preceding figure, the RAID arrays have different performance requirements. As a result, the system has redundant resources, but some services still have no resources to use. Resources are wasted in the following circumstances:

- Different services have different performance requirements.
RAID arrays are planned based on service requirements. The resources of different RAID arrays cannot be shared, resulting in waste of resources.
- A service has different performance requirements in different time periods.
The RAID array resources are planned based on the maximum performance required by only a very short period of time. As a result, the resources are underused and wasted in most of the time.

Figure 7-12 FusionCube distributed storage resource pool architecture



The FusionCube distributed storage uses a large resource pool. All hardware resources are involved in any service. Users can directly add services (new services or services performance improvement) as long as the overall system performance and capacity meet requirements. No extra performance planning or adjustment is required. Huawei FusionStorage can easily cope with IP SAN problems. It can maximize the system performance, minimize the maintenance investment, and reduce the service interruption risks.

7.3.4 SSD Cache vs SSD Tier

The SAN uses the memory of the engine as the I/O cache. However, the memory size is limited (8 GB, 16 GB, or 32 GB) and cannot meet service requirements, especially the applications with hybrid services and demanding high performance. Therefore, more and more storage vendors use SSDs to accelerate storage services. However, most storage vendors use SSDs as the SSD tier instead of SSD cache. The SSD tier is effective for single services with a fixed hotspot that lasts a long period of time. It does not apply to applications with hybrid services with frequently changed hotspots.

The FusionCube distributed storage uses SSDs as the cache layer to detect service hotspots in a timely manner and respond quickly to ensure high performance.

Table 7-2 Comparison between SSD cache and SSD tier

Description	SSD Cache	SSD Tier
Benefits	Hot data is stored in	Hot data is stored in

Description	SSD Cache	SSD Tier
	high-speed storage media (SSD cards or SSD disks) to improve the processing performance of the storage system.	high-speed storage media (SSD cards or SSD disks) to improve the processing performance of the storage system.
Data change	The hot data is migrated from HDDs to SSDs, but not deleted from HDDs. When hot data becomes cold, the SSD space is released.	The hot data is migrated from HDDs to SSDs and deleted from the HDDs. When hot data becomes cold, the SSD space is released and the data is written back to the HDDs.
Capacity	SSDs are only used as the cache and do not increase the total system capacity.	The SSD tier increases the total storage capacity provided by the system.
SSD space utilization	Hot data is backed up in HDDs. Therefore, reliability technologies such as RAID are not required in SSDs to ensure reliability. The space utilization is high.	Hot data is deleted from HDDs. When data is migrated to SSDs, RAID technology must be used to ensure reliability.
SSD performance	Data migrated to the SSDs is written directly without write penalty. After the hot data is cold, hot spots can be directly covered by new hot spots.	Write penalty will occur on the RAID arrays when data is migrated to SSDs. After hot data becomes cold, the data needs to be read and written back to the HDDs before being overwritten by new data.
Hotspot statistics period	It takes little time (few minutes) to detect a hotspot and the system responds quickly.	It takes a long time (several hours) to detect a hotspot, and the system responds slowly.
Size of the data management block	Small blocks, generally 8 KB, 16 KB, and 32 KB. The cache utilization is high.	Large block, generally 1 MB, 2 MB, and 4 MB. The cache space is wasted.
Applications	Hotspots change frequently. Hybrid services, especially services with changing hotspots.	Hotspots seldom change and lasts for a long time. Single service with fixed hotspots.

8 System Reliability

The FusionCube distributed storage system provides cross-node data protection. When multiple hard disks or nodes are faulty, the FusionCube distributed storage system can still provide services. Data is stored on different hard disks of different nodes in the same pool, achieving cross-node reliability and fast fault recovery. The hardware redundancy configuration also ensures high system availability.

- [8.1 Data Reliability](#)
- [8.2 Hardware Reliability](#)
- [8.3 Sub-Health Enhancement](#)
- [8.4 Backup and Restoration](#)
- [8.5 DR](#)

8.1 Data Reliability

8.1.1 Block Storage Cluster Reliability

The FusionCube distributed storage system uses cluster management to prevent SPOFs. When a node or hard disk is faulty, it will be automatically isolated from the cluster, minimizing adverse impact on system services. The cluster management is implemented as follows:

- **ZooKeeper**
ZooKeeper chooses the master MetaData Controller (MDC) and stores metadata generated during system initialization. The metadata includes data routing information, such as the mapping between partitions and hard disks. An odd number of ZooKeepers must be deployed in a system to form a cluster. A system must have a minimum of three ZooKeeper nodes, and more than half of the ZooKeeper nodes must be active and accessible. The number of ZooKeeper nodes cannot be added once the system is deployed.
- **MDC**
The MDC controls the status of the distributed clusters. A system must have a minimum of three MDCs. When a resource pool is added, an MDC will be automatically started or specified for the resource pool. ZooKeeper determines the master MDC from multiple MDCs. The master MDC monitors other MDCs. If the master MDC detects the fault of

an MDC, it restarts the MDC or specifies an MDC for the resource pool. When the master MDC is faulty, a new master MDC will be elected.

- OSD

The Object Storage Device (OSD) performs input/output operations. The OSDs work in active/standby mode. The MDC monitors the OSD status on a real-time basis. When the active OSD where a specified partition resides is faulty, services will be automatically switched over to the standby OSD to ensure service continuity.

Each node has multiple OSDs to manage drives or SSD virtual drives on the node. The OSDs are in one-to-one mapping with drives or SSD virtual drives, but do not bind with the drives or SSDs. The positions of the storage drives/SSDs in a node can be swapped. This can prevent misoperation during maintenance and improve system reliability.

8.1.2 Data Consistency

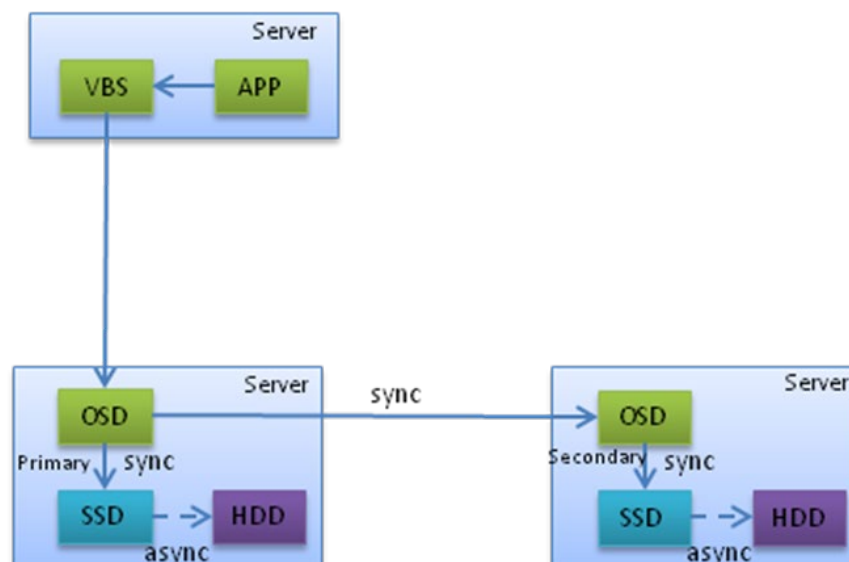
When data is written to the storage system, the data copies in the storage system must be consistent so that data read from any copy is the same.

FusionCube distributed storage uses the following methods to ensure data consistency between data copies:

- Synchronous write of data copies

When the VBS module sends a write request to the active OSD, the OSD synchronizes this write request to the standby OSD while writing it to the local hard disk. This synchronization process is implemented based on the I/O number to ensure the same sequence of the I/O operations on the active and standby OSDs. A success message is returned to the application only after the write operation is complete on the active and standby OSDs. Figure 8-1 shows the process.

Figure 8-1 Synchronous write of data copies



- Read repair

When a data read operation fails, the system determines the error type. If the read error occurs on a disk sector, the system automatically reads data from the copies stored on

other nodes and writes the data to the faulty node of the disk sector. This ensures that the total number of data copies and data consistency between copies.

8.1.3 Data Redundancy

The FusionCube distributed storage supports two data redundancy protection mechanisms: multi-copy backup and erasure code (EC).

The FusionCube distributed storage supports two-copy and three-copy backup. The two-copy backup allows normal system operation when a data disk or node is faulty. If two-copy backup is used, the SAS disks or SSDs must be used as the primary storage, and a storage pool supports a maximum of 96 disks. The three-copy backup allows normal system operation when any two data disks or nodes are faulty. The system must be configured with five ZooKeeper metadata management nodes. The three-copy backup supports all disk types. A storage pool supports 2048 disks. The three-copy backup is recommended for FusionCube.

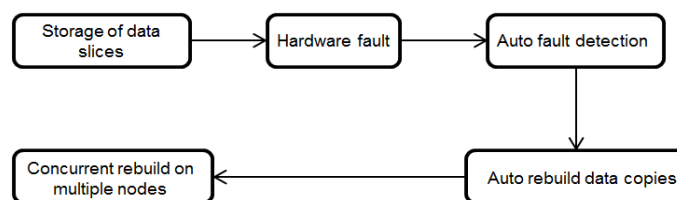
Compared with multi-copy backup, EC provides higher resource utilization but lower performance. FusionCube supports the following EC code rates: 2+2, 3+2 (:1), 4+2 (:1), 8+2 (:1), and 12+3 (:1). The code rate 4+2 is recommended. If the data turning algorithm is used, EC code rate 4+2 (:1) can be deployed on three nodes. However, the system reliability decreases. The system allows for the fault of two hard disks but not support the fault of two nodes at the same time.

8.1.4 Rapid Data Rebuild

Each hard disk in the FusionStorage system stores multiple data blocks (partitions), whose data copies are scattered on the nodes in the system based on certain distribution rules. If detecting a hard disk or server fault, the FusionCube distributed storage system automatically repairs data in the background. Since data copies are distributed on different storage nodes, data can be rebuilt on different nodes at the same time and each node has a minimum amount of data rebuilt. This mechanism prevents performance deterioration caused by restoration of a large amount of data on a single node, and therefore minimizes adverse impact on upper-layer services.

Figure 8-2 shows the automatic data rebuild process.

Figure 8-2 FusionCube data rebuild process



The FusionCube distributed storage supports concurrent and quick troubleshooting and data rebuild.

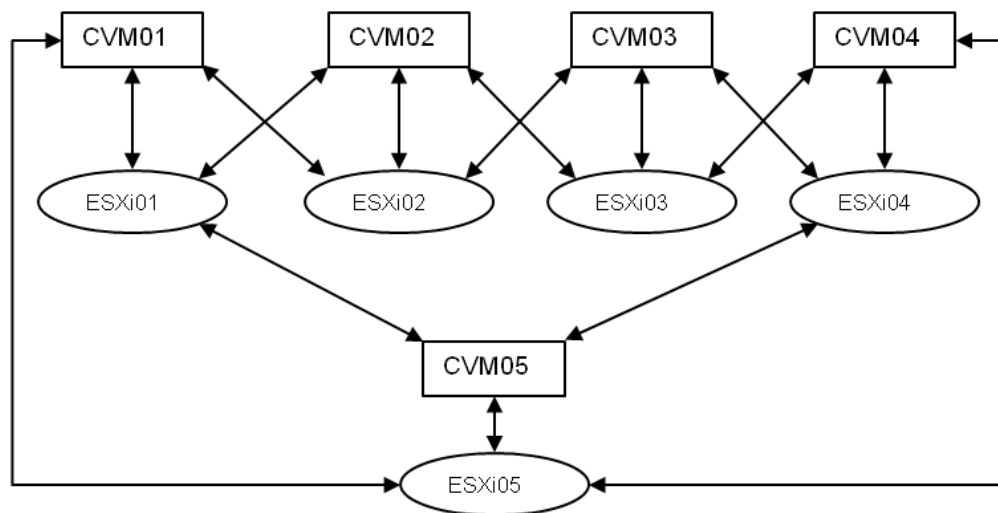
- Data blocks (partitions) and their copies are scattered in a resource pool. If a hard disk is faulty, data can be automatically rebuilt in the resource pool efficiently.
- Data is distributed to different nodes so that data can be obtained or rebuilt even if a node is faulty.

- Load can be automatically balanced between existing nodes in the event of node failures or capacity expansion. Larger capacity and higher performance can be achieved without modifying the application configuration.

8.1.5 Multipathing for Data Storage

If VMware vSphere is used, the FusionCube distributed storage is deployed on CVMs and provides the iSCSI target to communicate with iSCSI Initiator in ESXi. FusionCube supports multipathing configuration for iSCSI links to ensure that each host has more than two paths to access storage nodes.

Figure 8-3 iSCSI multipathing configuration



8.2 Hardware Reliability

FusionCube uses highly reliable Huawei-developed hardware in redundancy design to ensure system reliability. It has the following features:

- The cache is protected against power failures to ensure data security.
- Hot-swappable SAS disks in RAID 1 array are used as system disks.
- Servers are configured with redundant power supply modules and fans to ensure high system availability.
- Dual-plane design is used for network communication.
- The ES3000 NVMe SSD disks/cards support data integrity field (DIF) and provides data integrity check.

8.3 Sub-Health Enhancement

Sub-health is a status. If a component is in sub-health status, the component performance is severely lower than expected. FusionCube supports detection of the sub-health status of NICs, hard disks, memory, and CPUs. The components in sub-health status affect the overall system

performance directly or indirectly. FusionCube provides sub-health check and handling mechanisms for the following resources:

- Nodes
- Network
- Media

Node Sub-Health Check and Handling

The nodes in the OSD and replication clusters change to the sub-health status due to software or hardware problems, such as CPU underclocking and repeated memory error correction. If this occurs, the system service latency increases. The system locates the nodes in sub-health state based on the latency and isolates the nodes (or OSDs).

- **OSD cluster nodes**

Detection mechanism:

A module is deployed on the nodes accessing the OSDs to detect the access latency. If the access latency exceeds the threshold, an alarm will be reported to the control node.

The control node of the OSD cluster collects the OSD information reported by each access node and determines the sub-health status of the OSD according to the majority principle (if most of the nodes accessing an OSD report the OSD sub-health status, the OSD is determined to be in sub-health status).

Isolation mechanism:

The control node in the OSD cluster isolates the OSD node in sub-health status when the data redundancy requirements are met.

Restoration Mechanism:

After resolving the problem, the maintenance personnel use commands to add the isolated OSD to the cluster again.

- **Replica I/O path check and service handoff (Hint availability enhancement)**

Detection mechanism:

- A module is deployed on the nodes accessing the OSDs to detect the access latency in a short period of time. If the access latency exceeds the threshold, an alarm will be reported to the control node.

The control node of the OSD cluster collects the OSD information reported by each access node and determines the sub-health status of the OSD according to the majority principle (if most of the nodes accessing an OSD report the OSD sub-health status, the OSD is determined to be in sub-health status).

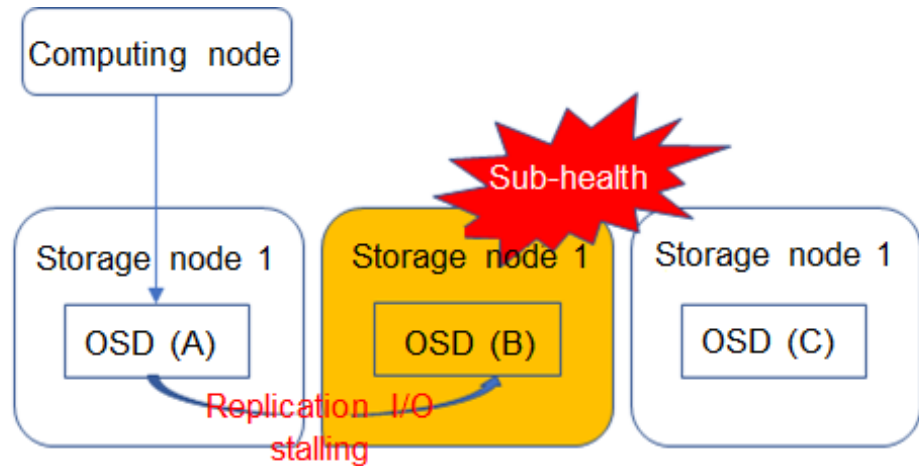
- When the latency of an I/O is too long, the control node determines the sub-health status of this OSD to ensure quick service recovery.

Handoff mechanism:

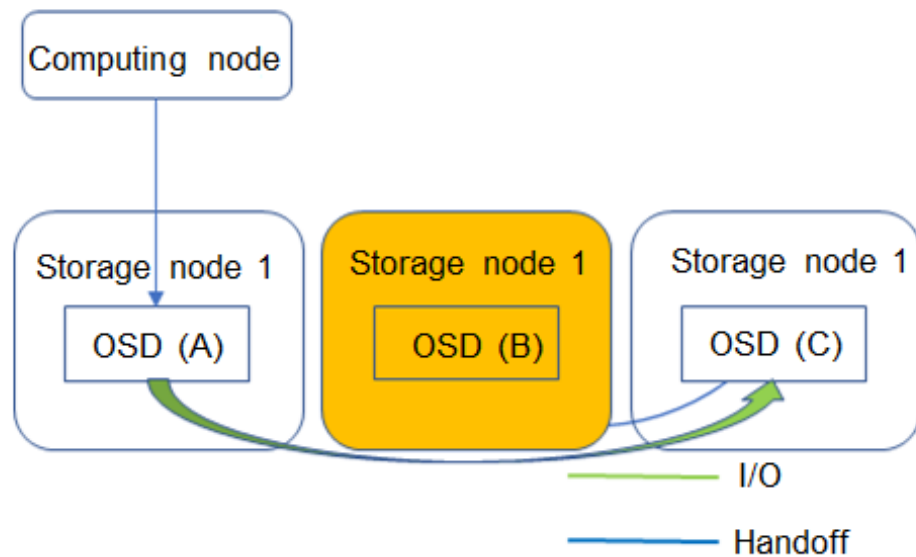
The control node in the OSD cluster generates a temporary view and switches over services from the OSD node in sub-health status to another OSD (Hint node) and writes the replication I/O of the master node to the Hint node.

Restoration Mechanism:

The system checks the access latency after a certain period of time. If the access latency decreases, the temporary data is pushed to the original OSD node through the background Handoff process. After data transmission is complete, the temporary view is deleted and the services are handed off to the original OSD. If the access latency is still long, the OSD node will be isolated.



Node sub-health scenario I/O flow



Node sub-health scenario I/O flow (hinted)

- **Replication cluster nodes**

Detection mechanism:

When accessing a replication node, the system detects the I/O path latency of the replication node and reports the latency information to the master node.

The master node collects the latency information reported by each node. If the ratio of timeout I/Os to the total number of I/Os is high, the replication node is in the sub-health state.

Isolation mechanism:

The replication master node isolates the node in sub-health status if the number of isolated nodes is within the limit.

Restoration Mechanism:

The master node adds the isolated replication node to the cluster after a period of time.

Network Sub-Health Check and Handling

The cluster network is in sub-health status when the network performance deteriorates due to decreased NIC speed or increased packet loss rate or packet error rate. The system locates the nodes based on the network resource status and performs an active/standby port handoff or isolates the node.

Detection mechanism:

System network sub-health check method:

Scenario		Check Method
Local node check	Intermittent disconnection of the network port	Check whether the number of intermittent disconnections within a unit time exceeds the threshold.
	Packet loss or error packets of the protocol stack	Check whether the packet loss rate or error packets per unit time exceeds the threshold.
	NIC bus width decrease	Check whether the NIC transmission rate decreases.
	PCIe speed decrease	Check whether the PCIe transmission rate decreases.
Inter-node communication check	Network between nodes	Ping the peer node to determine the network status.

The node reports the network sub-health status detected locally to the control node. The control node issues a command to switch over the network port or isolates the node.

The ping mechanism is used to check whether the peer node is reachable. Each node pings the peer node in the cluster and reports information about the node whose latency exceeds the threshold to the control node. The control node determines the network sub-health status of the node according to the majority principle (if most of the nodes that ping a node report the network sub-health status, the network of the node is in sub-health status). The control node performs port handoff or isolates the node in sub-health status.

Handoff and isolation mechanism:

If the network of a node is in sub-health status, the system attempts to hand off the active and standby network ports. If the network ports are not bonded in active/standby mode or the handoff cannot restore the network, the system isolates the node when the data redundancy conditions are met.

Restoration Mechanism:

The control node will add the isolated node to the cluster after a period of time.

Storage Media Sub-Health Check and Handling

The access to HDDs/SSDs slows down when the storage media in a cluster are in sub-health status due to firmware or mechanical problems. The system detects the storage media and isolates them in a timely manner.

Detection mechanism:

The sub-health status of a storage medium is detected using a threshold-based check mechanism and a homogeneous-medium comparison mechanism.

1. The node collects statistics on the I/O latency locally and determines the sub-health OSD based on the threshold.
2. The node reports the I/O latency statistics to the control node. The control node compares the statistics and identifies the OSD node that responds more slowly than other OSD nodes in the storage pool.

Isolation mechanism:

The control node isolates the OSD in sub-health status if the data redundancy conditions are met.

Restoration Mechanism:

After resolving the problem, the maintenance personnel use commands to add the isolated OSD to the cluster again.

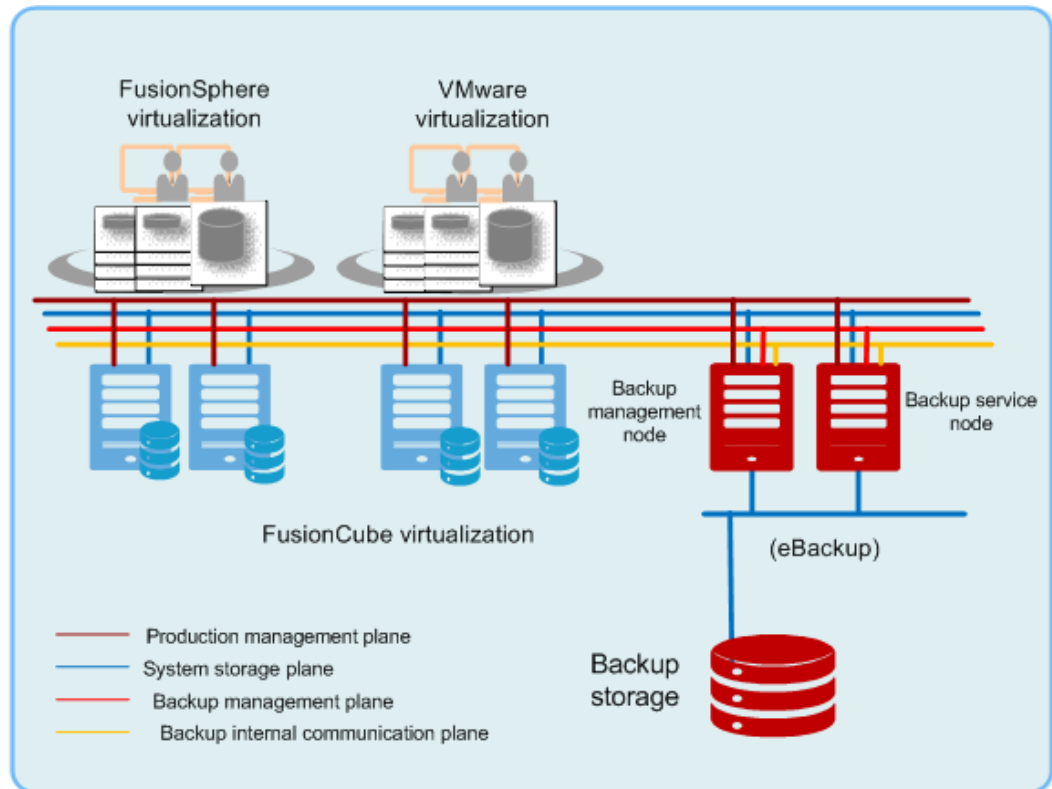
8.4 Backup and Restoration

The FusionCube HCI supports Huawei-developed backup software eBackup and mainstream third-party backup software like AnyBackup and Veeam. This section describes the eBackup software developed by Huawei.

Huawei eBackup applies to the FusionCube HCI that uses FusionSphere or VMware vSphere. It backs up massive VM data using VM snapshot and CBT technologies. eBackup supports VM backup and restoration, file-based recovery of the operating system, incremental restoration of disk data on VMs, and data backup to SAN, NAS, and S3 storage devices.

Figure 8-4 shows the architecture of the FusionCube+eBackup virtualization backup solution.

Figure 8-4 Architecture of the FusionCube+eBackup backup solution



The eBackup software consists of the following:

- Backup server: schedules and monitors backup and recovery tasks, manages backup storage and production systems, and receives requests from users. In addition, the backup server provides the backup agent function.
- Backup agent: receives backup and restoration tasks from the backup server and interacts with the backup storage and production system to perform the tasks. When the system backup workload is high, backup agents can be added to smoothly expand the backup performance.

In FusionCube FusionSphere scenarios, backup servers and agents must be deployed on servers. In VMware scenarios, backup servers and agents can be deployed on VMs.

The FusionCube+eBackup backup solution supports the following:

- Agent-free backup. No agent software needs to be installed on back VMs.
- Online VM backup. VM data can be backed up no matter whether the VMs are powered on or off.
- Data backup to SAN, NAS, and S3 storage.
- Multiple backup modes, including full backup, incremental backup, and batch backup.
 - Full backup supports valid data backup.
 - Incremental backup backs up only the changed data blocks since the last backup. Therefore, less data needs to be backed up, reducing VM backup costs and minimizing the backup window.
- A variety of restoration types, including restoring a single VM, multiple VMs in batches, disks of a single VM, and disks of multiple VMs in batches.

- Concurrent backup. A maximum of five concurrent backup tasks can be performed on an ESXi host, and a maximum of 40 concurrent backup tasks can be performed on a backup agent.

8.5 DR

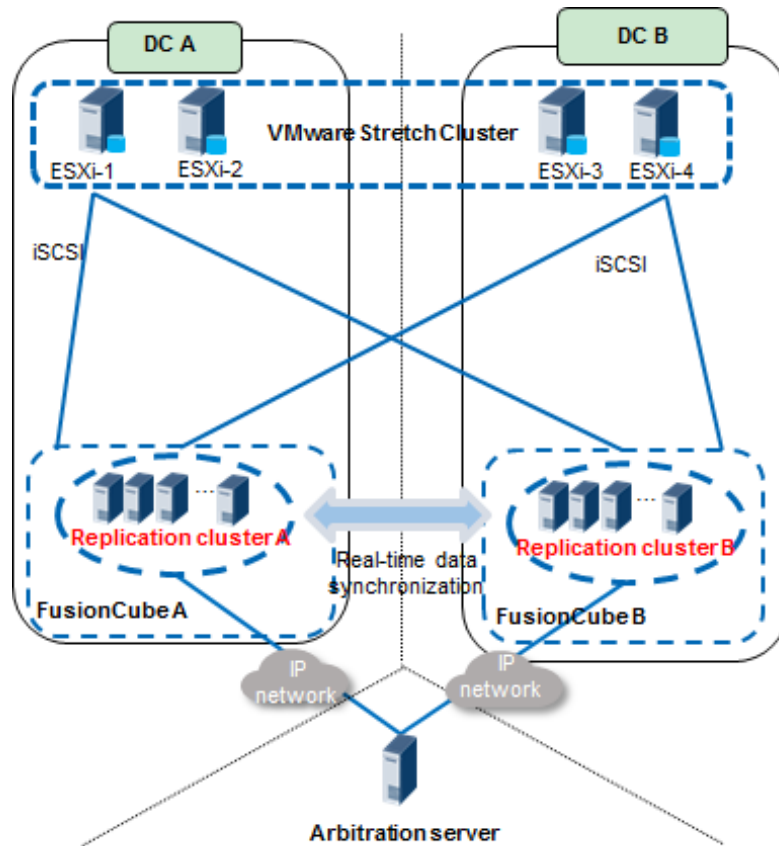
To ensure service continuity, enterprises need to consider how to protect data and quickly restore services in case of natural or man-made disasters, maloperations, or application failures. In addition to local backup, uploading data to a remote location is a good way to address these challenges. Great loss will be caused if there is no effective data protection and remote data upload measures. The DR system is a solution to the preceding challenges. Huawei FusionCube provides an end-to-end active-active data center solution based on active-active storage technologies and an active-standby DR solution based on asynchronous replication.

The end-to-end active-active solution implements load balancing and automatic failovers between two data centers. It offers optimal resource utilization. In the active-active solution, a success response is returned only after data is written to the two sites. Therefore, this solution allows almost zero recovery time objective (RTO) and recovery point objective (RPO). Currently, FusionCube supports only the active-active storage solution for VMware applications.

The DR solution based on asynchronous replication backs up data to the storage pool at the DR site in asynchronous mode. Therefore, this solution applies to remote deployment or sites allowing for long transmission latency.

8.5.1 Active-Active DR Solution

Figure 8-5 Logical architecture of the FusionCube VMware active-active solution



The Huawei FusionCube VMware active-active DR solution is an end-to-end disaster recovery solution based on the FusionStorage block storage active-active architecture and VMware's native multipathing software (NMP).

Two sets of FusionCube constitute active-active DR. A cross-site virtual volume is created from the volumes of two sets of FusionStorage. The data on the virtual volume is synchronized between the storage of the two sites on a real time basis. The storage of the two sites provides undifferentiated concurrent access for application servers and can simultaneously process the read and write requests from application servers. The VMware NMP implements cross-site networking. If the storage of FusionCube A is faulty, the storage of FusionCube B automatically takes over service processing. The upper-layer service system is unaware of the failovers between the two sites. If FusionCube A is faulty, the VMs of site A will be automatically migrated to FusionCube B, ensuring service continuity.

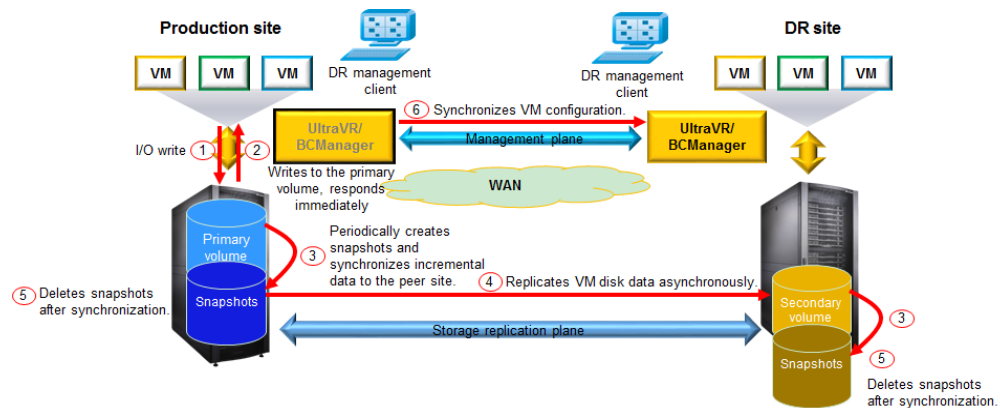
The system value-added service replication cluster provides the active-active capability for the active-active storage feature. Users can deploy replication clusters based on demand. The system resource consumption will not increase and the system performance will not be affected if no replication cluster is deployed.

The replication cluster has good scalability. A replication cluster supports 3 to 64 nodes. A FusionCube system supports a maximum of 128 clusters. Each replication cluster supports up to 38,000 active-active volumes and 9500 consistency groups. The whole system supports a maximum of 4,864,000 active-active volumes and 1,216,000 consistency groups.

The optimistic locking mechanism is used to prevent write conflicts between the two sites to reduce system complexity and improve performance. A compute node provides 80,000 active-active read/write performance.

The system supports two arbitration modes: third-party arbitration and static priority arbitration. It supports service-based arbitration to prevent unnecessary failover and improve service continuity.

8.5.2 Asynchronous Replication Solution



The Huawei asynchronous replication architecture consists of two sets of FusionCube distributed storage that build the asynchronous replication relationship and the UltraVR or BCManager DR management software. The data on the primary and secondary volumes are periodically synchronized based on the comparison of snapshots. All the data generated on the primary volume after the last synchronization will be written to the secondary volume in the next synchronization.

Storage DR clusters can be deployed based on service requirements. The storage DR cluster is a logical object that provides replication services. It manages cluster nodes, cluster metadata, replication pairs, consistency groups, and performs data migration. The DR cluster and system service storage are deployed on storage nodes. The DR cluster has good scalability. A DR cluster supports 3 to 64 nodes. The entire system supports a maximum of eight clusters. A DR cluster supports a maximum of 64000 volumes and 16000 consistency groups, meeting the requirements for rapid growth of DR services.

The UltraVR or BCManager manages DR services from the perspective of applications and protects the service VMs of the FusionCube system. It provides process-based DR service configuration, including one-click DR test, DR policy configuration, and fault recovery operations at the active site.

9 System Security

9.1 System Security Threats

9.2 Overall Security Framework

9.1 System Security Threats

Security Threats from External Networks

- Traditional IP attacks
Traditional IP attacks include port scans, IP address spoofing, land attacks, IP option attacks, IP routing attacks, IP fragmentation attacks, IP fragment packet attacks, and teardrop attacks.
- OS and software vulnerabilities
Numerous security bugs have been found in compute software, including third-part, commercial and free software. Hackers can control the OS by making use of minor programming errors or context dependency. Common OS and software vulnerabilities include buffer overflows, operations abusing privilege, and code download without integrity verification.
- Viruses, Trojan horses, and worms.
- SQL injection attacks
Attackers include portions of SQL statements in a web form entry field or query character strings of a page request in an attempt to get the node to execute malicious SQL statements. The forms, in which the contents entered by users are directly used to construct (or affect) dynamic SQL commands or the contents are used as input parameters for stored procedures, are particularly vulnerable to SQL injection attacks.
- Phishing attacks
Phishing is the act of attempting to acquire information, such as user names, passwords, and credit card details, by masquerading as a trustworthy entity in an electronic communication. Communications purporting to be from popular social web sites, auction sites, online payment processors or IT administrators are commonly used to lure the public. Phishing is typically carried out via e-mail or instant messaging.
- Zero-day attacks
Zero-day vulnerabilities are security vulnerabilities that have not been patched. Zero-day attacks are attacks that exploit zero-day vulnerabilities. It is difficult to install patches

immediately after a security vulnerability is found because it takes time to confirm, verify, evaluate, and fix the vulnerability. Therefore, zero-day vulnerabilities pose a great threat to network security.

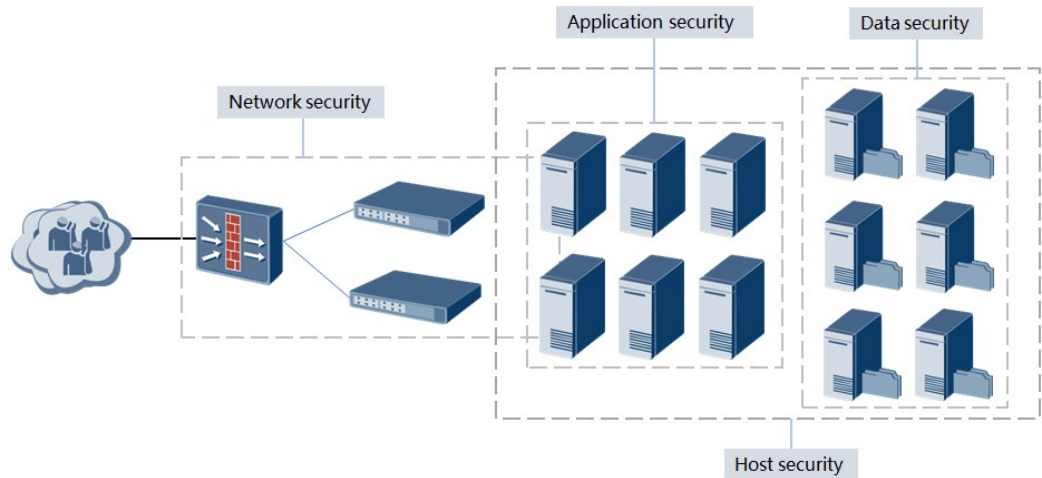
Security Threats from the Internal Network

- Ever-changing attacks pose security risks.
Internal network ARP spoofing and abuse of malicious plug-ins pose new security threats. The attacked intranet host may be used as "zombies" for penetration attacks or used as a DDOS tool to send a large number of attack packets to occupy network bandwidth. The system is vulnerable to attacks if malicious plug-ins are used or web pages that are implanted with viruses or Trojan horses are viewed.
- Worms and viruses are spread through loopholes if patches are not upgraded or the antivirus database is not updated in a timely manner.
If the OS, database, and application software of the hosts and devices on the network have security vulnerabilities and are not patched and the antivirus database of the hosts is not updated in a timely manner, viruses and worms will be spread. A large-scale worm outbreak may paralyze the intranet and interrupt services.
- Confidential information disclosure happens frequently because of unauthorized Internet access activities.
Enterprise employees can bypass the firewall and directly connect to the external network through the telephone, VPN, or GPRS. This may cause disclosure of important confidential information.
- Uncontrolled mobile device access challenges network border security.
The notebooks, pocket PCs, and other mobile devices of employees or temporary visitors are used in various network environments and may carry viruses and Trojan. If these devices access the intranet without being scanned, the intranet security will be threatened.
- Uncontrolled use of hardware and software threatens asset security.
If internal assets (such as CPUs, DIMMs, and hard disks) are replaced and modified without effective tracing measures and unified management, it is difficult to locate the fault once an attack or security incident occurs.
- Application software without monitoring mechanism poses new security risks.
Popular social communication applications, such as QQ and MSN, may spread viruses, worms, and Trojan horses. Using network tools, such as BitTorrent and eMule, to download movies, games, and software will affect the bandwidth for mission-critical services.
- Ineffective management of peripherals causes data leakage and virus spreading.
Peripherals, such as USB flash drives, CD drives, printer, infrared, serial port, and parallel ports, are prone to data leakage and virus infection. The peripherals, especially the USB flash drives, cannot be effectively managed by sealing the ports or introducing regulations. Technical measures are required to manage and control the peripherals.
- Security regulations without support of technical measures cannot be put into practice.

9.2 Overall Security Framework

FusionCube provides a security solution to address security risks and issues. Figure 9-1 shows the security framework. The FusionCube security framework ensures system security from the network, host, application, and data aspects.

Figure 9-1 FusionCube security solution framework



The FusionCube security is ensured from the following aspects:

- Network security
Network isolation is used to ensure normal data processing, storage security, and maintenance.
- Application security
Login user authentication, rights control, and audit control are adopted to ensure application security.
- Host security
OS security hardening has been performed to ensure normal operation of hosts.
- Data security
Cluster disaster recovery (DR), cluster backup, data integrity, and data confidentiality are used to ensure data security.

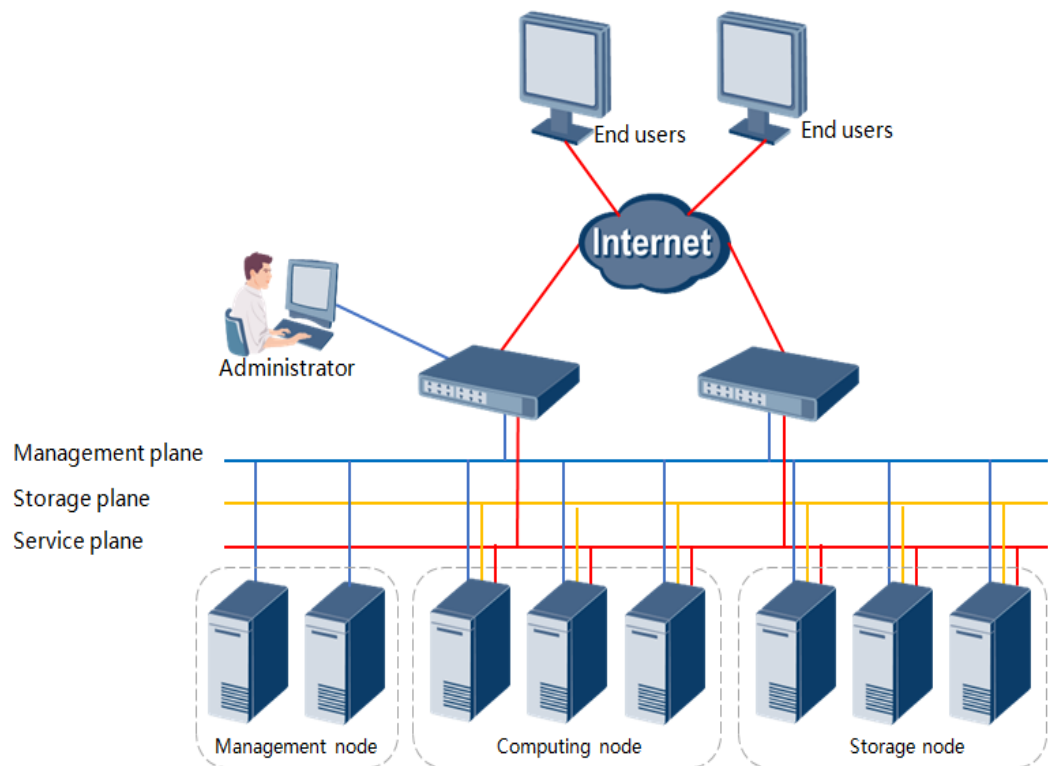
9.2.1 Network Security

The FusionCube network communication plane is divided into the following planes:

- Service plane
The service plane provides service channels and the communication plane for the virtual NICs of VMs.
- Storage plane
The storage plane enables VMs to access storage resources. The storage plane communicates with VMs through the virtualization platform.
- Management plane
The management plane provides communication channels for system management, routine maintenance, service configuration, and system loading.

For security purposes, isolate the three planes from each other. Figure 9-2 illustrates the isolation of the planes.

Figure 9-2 Plane isolation



9.2.2 Application Security

9.2.2.1 Rights Management

FusionCube supports rights-based management. Users are assigned different rights to ensure system security.

The operation rights of a user are defined by the role of the user. A user can have multiple roles, and a role can have various operation rights. The binding between a user and a role determines the operations that the user can perform. If a user has multiple roles, the user can perform all the operations defined for these roles.

9.2.2.2 Web Security

FusionCube implements the following web service security functions:

- Automatically converts user requests into HTTPS links.
 The web service platform automatically redirects user requests to HTTPS links. When a user accesses the web service platform using HTTP, the web service platform automatically converts the user requests into HTTPS requests to enhance access security.
- Prevents cross-site scripting.
 Cross-site scripting is a type of injection, in which attackers use insecure websites to attack website visitors.
- Prevents SQL injections.
 Attackers inject SQL commands to entry fields of web sheets or query character strings of page requests to enable servers to execute malicious SQL commands.

- Prevents cross-site request forgeries.
Malicious requests can exploit the browser's function of automatically sending authentication certificates. A cross-site request forgery attack deceives a logged-in user into loading a page with a malicious request in order to inherit the identity and privileges of the user. In this way, the attacker can make mischief for their own purposes, for example, changing the user's password or address information.
- Hides sensitive information for security purposes.
Sensitive information is hidden to prevent attackers' access.
- Restricts file upload and download.
Measures are taken to prevent confidential files from being downloaded and insecure files from being uploaded.
- Prevents unauthorized uniform resource locator (URL) access.
Users are prevented from accessing unauthorized URLs.
- Supports graphic verification codes for logins.
When a user attempts to log in to the web system, a random verification code will be generated. The user can log in to the system only when the user name, password, and random verification code are correct.

9.2.2.3 Database Hardening

The FusionCube management nodes use GaussDB database.

Basic security configurations are required to ensure database security. The security configurations for a GaussDB database include the following:

- Access source control
According to service requirements and security standards, the database allows local access only. All cross-server access requests are rejected to prevent external attacks.
- Principle of least privilege
Except the database super administrator, all the users are assigned roles based on the principle of least privilege.
- Folder protection
The owner of the data installation folder and its data area folders is the user who performs the installation, and the permission on the folders and its subfolders includes read, write, and execute.
- Protection of sensitive files
The owner of the database core configuration files is the user who performs the installation, and the permission on the files includes read and write.
- Restriction on the number of concurrent connections
By default, the system supports a maximum of 300 connections. The maximum number of connections can be modified in the configuration file to prevent malicious attacks.

To ensure data security, the data in the database must be backed up periodically. The database supports local online back and remote backup.

- Local backup: A script is executed at the specified time to back up data.
- Remote backup: Data is backed up to a third-party server.

9.2.2.4 Log Management

The following measures are used to ensure log security:

- Logs cannot be modified or deleted on the management systems.
- Only the users authorized to query log information can export logs.

9.2.3 Host Security

9.2.3.1 OS Security Hardening

All the compute nodes, storage nodes, and management nodes of FusionCube use the Linux OS. The following configuration must be performed to ensure OS security:

- Stop unnecessary services, such as Telnet service and file transfer protocol (FTP) service.
- Perform security hardening of the secure shell (SSH) service.
- Control the access permission on files and directories.
- Allow system access only for authorized users.
- Manage user passwords.
- Record operation logs.
- Detect system exceptions.

9.2.4 Data Security

FusionCube uses a variety of storage security technologies to ensure the security and reliability of user data.

- Data fragment storage
Data on FusionCube storage nodes is automatically stored in multiple copies. Different copies of each data slice are stored on different storage nodes. As a result, malicious users cannot obtain user data from a single storage node or physical disk.
- Encrypted storage of sensitive data
The AES-256 or SHA-256 algorithm is used to encrypt sensitive data (such as authentication information) stored in the database.