OceanStor Dorado 6.x & OceanStor 6.x Host Connectivity Guide for Connecting to Linux Hosts Using NVMe over Fabrics

Issue	15
Date	2024-02-01

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

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

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

Huawei Technologies Co., Ltd.

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

Website: https://e.huawei.com

Security Declaration

Product Lifecycle

Huawei's regulations on product lifecycle are subject to the *Product End of Life Policy*. For details about this policy, visit the following web page:
https://support.huawei.com/ecolumnsweb/en/warranty-policy

Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process*. For details about this process, visit the following web page:
https://www.huawei.com/en/psirt/vul-response-process
For vulnerability information, enterprise customers can visit the following web page:
https://securitybulletin.huawei.com/enterprise/en/security-advisory

Preconfigured Digital Certificate

The digital certificates preconfigured on Huawei devices are subject to the *Rights and Responsibilities of Preconfigured Digital Certificates on Huawei Devices*. For details about this document, visit the following web page:
https://support.huawei.com/enterprise/en/bulletins-service/ENEWS2000015789

Huawei Enterprise End User License Agreement

This agreement is the end user license agreement between you (an individual, company, or any other entity) and Huawei for the use of the Huawei Software. Your use of the Huawei Software will be deemed as your acceptance of the terms mentioned in this agreement. For details about this agreement, visit the following web page:
https://e.huawei.com/en/about/eula

Lifecycle of Product Documentation

Huawei after-sales user documentation is subject to the *Product Documentation Lifecycle Policy*. For details about this policy, visit the following web page:
https://support.huawei.com/enterprise/en/bulletins-website/ENEWS2000017761

Contents

1 About This Document

1.1 Purpose

This document details the configuration methods and precautions for connecting OceanStor Dorado storage systems to Linux hosts using NVMe over Fabrics (NVMe-oF).

The following table lists the product models that this document is applicable to.

Product Series	Product Model	Product Version
OceanStor Dorado[a]	OceanStor Dorado 3000	6.0.1 6.1.0 6.1.2 6.1.3 6.1.5 6.1.6 6.1.7
	OceanStor Dorado 5000	
	OceanStor Dorado 6000	
	OceanStor Dorado 8000	
	OceanStor Dorado 18000	
OceanStor	OceanStor 5310	6.1.3 6.1.5 6.1.6 6.1.7
	OceanStor 5510	
	OceanStor 5610	
	OceanStor 6810	

Product Series	Product Model	Product Version
	OceanStor 18510	
	OceanStor 18810	
	OceanStor 2200	6.1.6 6.1.7
	OceanStor 2600	
	OceanStor 2220	
	OceanStor 2620	
a: The 6.0.1 version does not support NVMe over RoCE.		

1.2 Audience

This document is intended for:

- Huawei technical support engineers
- Technical engineers of Huawei's partners
- Personnel who are involved in interconnecting Huawei storage systems and Linux servers using NVMe-oF.

Readers of this guide are expected to be familiar with the following topics:

- Huawei storage systems
- Linux
- NVMe

1.3 Related Documents

For the hosts, host bus adapters (HBAs), and operating systems that are compatible with Huawei storage devices, go to **info.support.huawei.com/storage/comp** .

For the latest Huawei storage product documentation, go to **support.huawei.com**.

For Linux-related documentation or support, go to the official website of the OS:

- Red Hat: **www.redhat.com/en/services/support**
- SUSE: **www.suse.com/support**

The commands used in this document are effective for specific OS releases. For the commands used by other OS releases, see their respective system administration guide on the official website.

1.4 Conventions

Symbol Conventions

Symbol	Description
⚠ DANGER	Indicates a hazard with a high level of risk which, if not avoided, will result in death or serious injury.
⚠ WARNING	Indicates a hazard with a medium level of risk which, if not avoided, could result in death or serious injury.
⚠ CAUTION	Indicates a hazard with a low level of risk which, if not avoided, could result in minor or moderate injury.
NOTICE	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.
🕮 NOTE	Supplements the important information in the main text. NOTE is used to address information not related to personal injury, equipment damage, and environment deterioration.

1.5 Change History

Changes between document issues are cumulative. The latest document issue contains all the changes made in earlier issues.

Issue 15 (2024-02-01)

This is the fifteenths official release. The updates are as follows:

Added the configuration guide for the OS Native Device Mapper and NVMe Native provided.

Issue 14 (2023-08-02)

This is the fourteenth official release. The updates are as follows:

Added the product models: OceanStor 2200, 2600, 2220, and 2620.

Issue 13 (2023-02-24)

This is the thirteenth official release. The updates are as follows:

Optimized descriptions in "Configuring Connectivity".

Issue 12 (2022-11-30)

This is the twelfth official release. The updates are as follows:

Optimized descriptions about some operations.

Issue 11 (2022-02-07)

This is the eleventh official release. The updates are as follows:

Provided support for OceanStor 6.x products.

Issue 10 (2021-11-15)

This is the tenth official release. The updates are as follows:

Optimized descriptions about some operations.

Issue 09 (2021-11-01)

This is the ninth official release. The updates are as follows:

Updated the section "Installing and Configuring OceanStor NOF Enabler."

Issue 08 (2021-07-15)

This is the eighth official release. The updates are as follows:

Updated the section "Configuring Multipathing."

Issue 07 (2021-04-20)

This is the seventh official release. The updates are as follows:

Added the example of domain configuration planning in switch networking.

1.6 Where To Get Help

Huawei support and product information can be obtained on the Huawei Online Support site.

Product Information

For documentation, release notes, software updates, and other information about Huawei products and support, go to the Huawei Online Support site (registration required) at **https://support.huawei.com/enterprise/**.

Technical Support

Huawei has a global technical support system, able to offer timely onsite and remote technical support service.

For any assistance, contact:

- Your local technical support

  https://e.huawei.com/en/branch-office-query

- Huawei company headquarters.

  Huawei Technologies Co., Ltd.

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

  Website: **https://e.huawei.com/**

Document Feedback

Huawei welcomes your suggestions for improving our documentation. If you have comments, send your feedback to infoit@huawei.com.

2 Introduction

2.1 Basic Concepts

2.1.1 NVMe Overview

Non-volatile memory express (NVMe) is a controller interface standard developed for PCI Express (PCIe) SSD systems. It defines specifications on the optimized controller register interface, command set, and I/O queue management to standardize the communication and data transmission between the SSD controller and the operating system. It does not define how the storage system meets upper-layer service requirements. NVMe aims to achieve high-performance access to flash media. It is originally used to access PCIe SSDs and is evolving to remote access based on other networks.

2.1.2 NVMe Highlights

NVMe has the following highlights:

1. Simple protocol, free of SCSI compatibility issues.
2. Supports a maximum of 65,535 I/O queues and 64,000 concurrent I/Os in each queue.
3. Supports priority queues.

The block device in the Linux kernel supports multiple queues. NVMe devices take the advantages of multiple queues, multi-core CPUs, and the lock-free mechanism to achieve high performance. The NVMe driver is simpler than the SCSI driver. Accessing back-end disks using NVMe does not pass through the SCSI layer or SAS

HBA. NVMe benefits from the host driver, which can implement multi-core and multi-queue processing to provide high performance.

2.1.3 NVMe over Fabrics

NVMe over Fabrics (NVMe-oF) carries the NVMe protocol over a network fabric (such as Ethernet, Fibre Channel, and InfiniBand) to remotely access SSDs with high performance and low latency. Currently, NVMe over RDMA, NVMe over FC, and NVMe over TCP have been released. NVMe-oF inherits the protocol model of NVMe over PCIe, including the subsystem, controller, and namespace. The major difference is that NVMe-oF converts PCIe register access and DMA access to different transport layer protocols, and adds the discovery mechanism.

2.1.4 Concepts Relevant to Hosts

2.1.4.1 Namespace

A namespace is a collection of logical blocks. On an NVMe SSD, a namespace is non-volatile memory formatted into logical blocks, similar to a LUN in the SCSI model.

Namespaces are identified by the IEEE Extended Unique Identifier (EUI-64) or Namespace Globally Unique Identifier (NGUID). An EUI-64 has 8 bytes and is used when there are only a small number of namespaces. An NGUID has 16 bytes and is used when there are a large number of namespaces. When a namespace is created, the controller specifies an EUI-64 or an NGUID or both to uniquely identify the namespace.

During data access, namespaces are addressed using namespace IDs, which are similar to host LUN IDs in the SCSI model. Namespace IDs cannot be 0. Before the host software delivers commands to a namespace, the namespace must be associated with the controller.

Namespaces can be classified into:

- Private namespace, which can be accessed by only one controller.
- Shared namespace, which can be accessed by multiple controllers.

2.1.4.2 Controller

The NVMe controller is a virtual function module that controls and processes NVMe queues and commands.

NVMe controllers are identified by controller IDs, which are allocated by storage vendors and must be unique in the NVMe subsystem. The NVMe protocol does not define how to allocate the controller IDs.

2.1.4.3 Subsystem

A subsystem is the NVMe target-side system, similar to a target in the SCSI system. A subsystem consists of one or more (a maximum of 65,535) NVMe controllers, one or more (a maximum of 65,535) NVMe subsystem ports, storage media, and ports between the controllers and the storage media.

For NVMe-oF, a subsystem NQN (SubNQN) is used to uniquely identify a subsystem. The storage vendors determine how to allocate the SubNQNs.

2.1.4.4 Port

Ports in the NVMe standard are classified into physical, transport, and subsystem ports.

- Physical port

  A physical port is a real port that connects a host to a subsystem, for example, a 10GE or 40GE port.

- Transport port

  A transport port is a TCP or UDP port, which is determined by the service protocol type. For example, for FTP services, the transport port number is 20 (TCP/UDP FTP data transfer) or 21 (TCP/SCTP/UDP FTP control). NVMe-oF generally uses port 4420.

- Subsystem port

  A subsystem port is a collection of one or more physical ports. It is a logical port that connects a host and a subsystem.

2.1.4.5 Session

Only one session can exist between an NVMe host and a controller, which is identified by a host session ID.

2.1.4.6 Connection

A connection is a queue pair between an NVMe host and a controller.

2.2 Connectivity

2.2.1 FC-NVMe

NVMe over Fibre Channel (FC-NVMe) transmits NVMe messages and commands over a Fibre Channel network between the host and the target storage subsystem.

Both FC-NVMe and FC-SCSI are based on the FCP. I/Os are based on exchange.

2.2.2 NVMe over RoCE

NVMe over RoCE is a type of the NVMe over RDMA protocol. RDMA can be RoCE, InfiniBand, or iWARP.

NVMe over RDMA maps the NVMe device's I/O queues to the RDMA queue pairs (QPs), and completes the I/O interactions using the RDMA SEND, RDMA WRITE, and RDMA READ semantics.

2.2.3 Multipath Connectivity

2.2.3.1 OS Native Multipathing Software

DM-Multipath is the native multipathing software in Linux.

DM-Multipath allows you to configure multiple I/O paths between a host and a storage system as one device. These I/O paths may contain independent physical devices such as cables, switches, and controllers.

DM-Multipath supports redundant paths and improves system performance.

- Redundancy

  DM-Multipath supports active/standby path configuration. This configuration creates a redundant path for each active path. The redundant paths are not used when the active paths work properly. Once an element (such as a cable, switch, or controller) on an active I/O path becomes faulty, DM-Multipath switches I/Os to a standby path.

- Performance enhancement

  DM-Multipath supports active-active paths, that is, I/Os are distributed to all paths based on the I/O scheduling algorithm. DM-Multipath can check I/O loads on paths and dynamically balance I/Os among the paths using the round-robin algorithm.

  Table 2-1 describes DM-Multipath components.

Table 2-1 DM-Multipath components

Component	Description
Kernel module	Redirects I/Os on paths and path groups and provides redundant paths.
mpathconf	A command used to configure and manage DM-Multipath (applicable in some operating systems)
multipath	A management command used to list and configure multipathing devices
multipathd	A daemon process that monitors paths. It initiates path switchover upon a path fault. This process also interactively modifies multipathing devices. This process is started before the **/etc/multipath.conf** file is modified.

2.2.3.2 UltraPath

UltraPath is a Huawei-developed multipathing software. It can manage and process disk creation/deletion and I/O delivery of operating systems.

UltraPath provides the following functions:

- Masking of redundant LUNs

  In a redundant storage network, an application server with no multipathing software detects a LUN on each path. Therefore, a LUN mapped through multiple paths is mistaken for two or more different LUNs. UltraPath installed on the application server masks redundant LUNs on the operating system driver layer to provide the application server with only one available LUN, the virtual LUN. In this case, the application server only needs to deliver data read and write operations to UltraPath that masks the redundant LUNs, and properly writes data into LUNs without damaging other data.

- Optimum path selection

  In a multipath environment, an application server with UltraPath accesses a LUN on the storage system through an optimum path, thereby obtaining the highest I/O speed.

- Failover and failback
  - Failover

    When a path fails, UltraPath fails over its services to another functional path.

  - Failback

    UltraPath automatically delivers I/Os to the first path again after the path recovers from the fault.

- I/O Load balancing

  UltraPath provides load balancing within a controller and across controllers.

  - For load balancing within a controller, I/Os poll among all the paths of the controller.
  - For load balancing across controllers, I/Os poll among the paths of all these controllers.

- Path test

  UltraPath tests the following paths:

  - Faulty paths

    UltraPath tests faulty paths with a high frequency to detect the path recover as soon as possible.

  - Idle paths

    UltraPath tests idle paths to identify faulty paths in advance, preventing unnecessary I/O retries. The test frequency is kept low to minimize impact on service I/Os.

2.2.3.3 NVMe Native Multipathing

NVMe native multipathing software allows hosts to access a shared namespace through two or more independent PCIe paths. That is, one or more hosts can access the same namespace through different NVMe controllers.

2.3 Interoperability Query

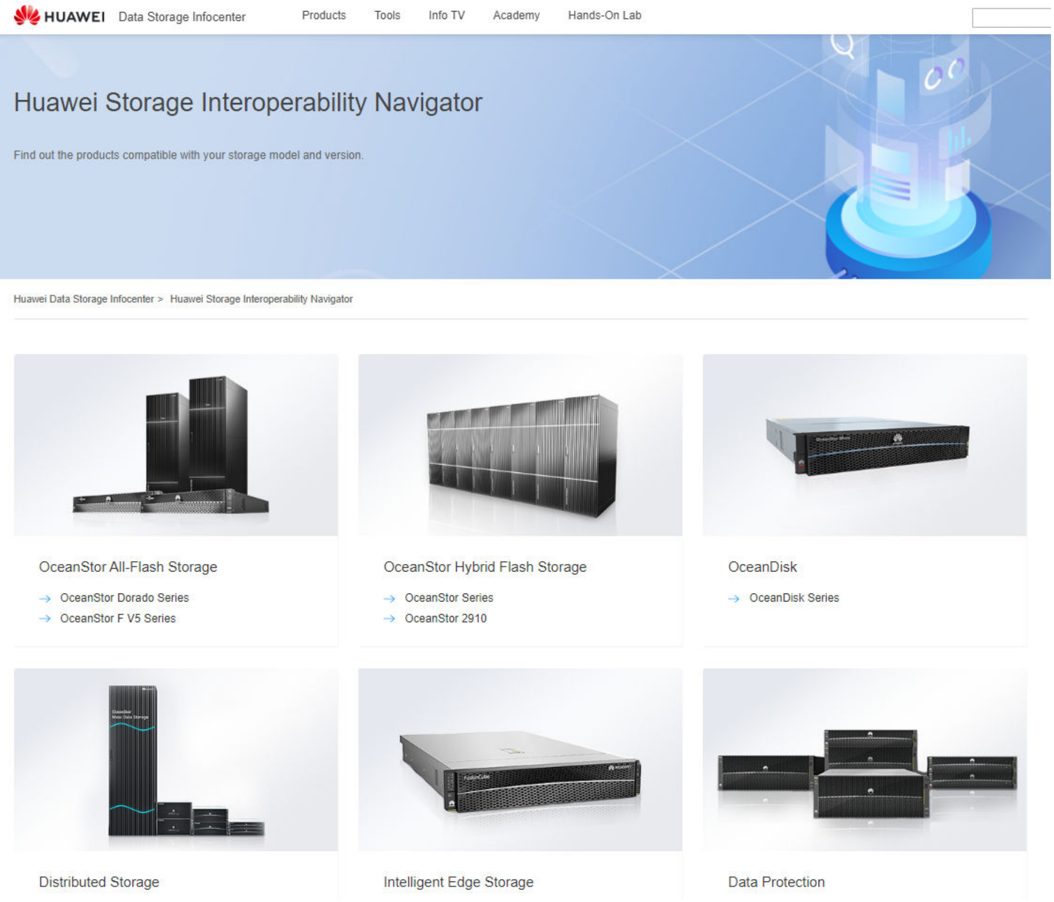
When connecting a storage system to a Linux host using NVMe-oF, consider the interoperability of components (such as the storage system and its front-end interface modules, host OS, HBA, HBA driver, switches, and multipathing software) in the environment.

You can query the latest compatibility information by performing the following steps:

Step 1 Log in to the website **info.support.huawei.com/storage/comp**.

Step 2 On the **Huawei Storage Interoperability Navigator** page, select the desired product. See **Figure 2-1**.

Figure 2-1 Huawei Storage Interoperability Navigator

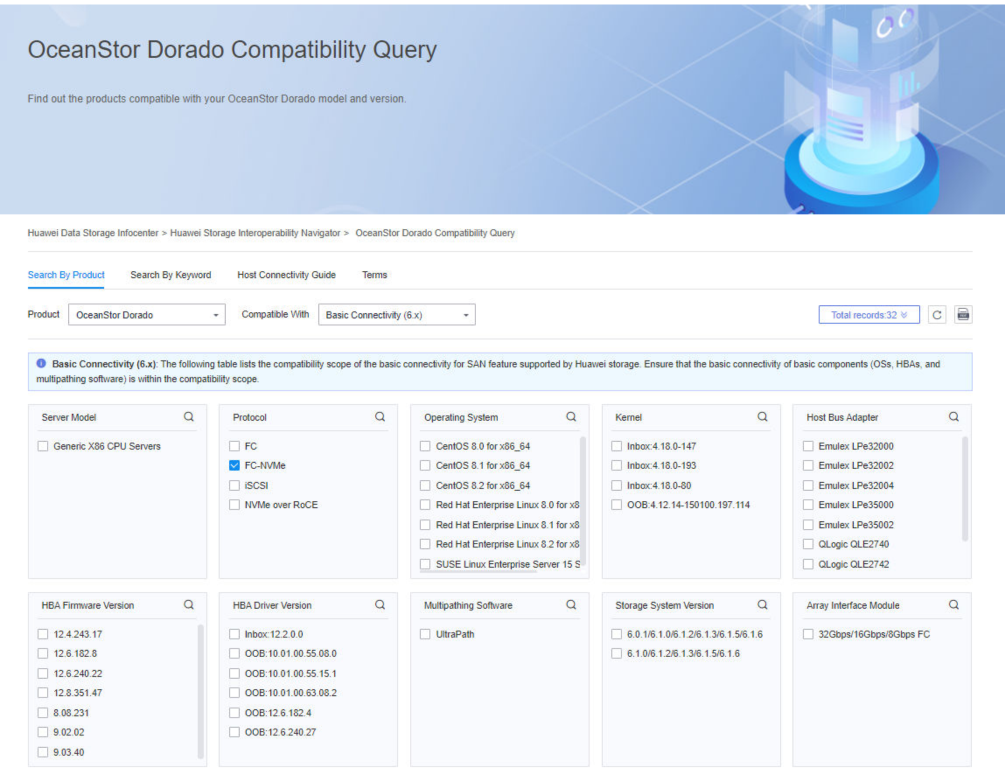


HUAWEI Data Storage Infocenter
Products
Tools
Info TV
Academy
Hands-On Lab
Huawei Storage Interoperability Navigator
Find out the products compatible with your storage model and version.
Huawei Data Storage Infocenter > Huawei Storage Interoperability Navigator
OceanStor All-Flash Storage
OceanStor Dorado Series
OceanStor F V5 Series
OceanStor Hybrid Flash Storage
OceanStor Series
OceanStor 2910
OceanDisk
OceanDisk Series
Distributed Storage
Intelligent Edge Storage
Data Protection

The page for querying the compatibility of this product is displayed.

Step 3 Select the component for query. **Figure 2-2** shows an example of querying the FC-NVMe protocol.

Figure 2-2 Query page



OceanStor Dorado Compatibility Query

Find out the products compatible with your OceanStor Dorado model and version.

Huawei Data Storage Infocenter > Huawei Storage Interoperability Navigator > OceanStor Dorado Compatibility Query

Product: OceanStor Dorado
Compatible With: Basic Connectivity (6.x)
Total records:32

Basic Connectivity (6.x): The following table lists the compatibility scope of the basic connectivity for SAN feature supported by Huawei storage. Ensure that the basic connectivity of basic components (OSs, HBAs, and multipathing software) is within the compatibility scope.

Server Model
□ Generic X86 CPU Servers

Protocol
□ FC
☑ FC-NVMe
□ iSCSI
□ NVMe over RoCE

Operating System
□ CentOS 8.0 for x86_64
□ CentOS 8.1 for x86_64
□ CentOS 8.2 for x86_64
□ Red Hat Enterprise Linux 8.0 for x8
□ Red Hat Enterprise Linux 8.1 for x8
□ Red Hat Enterprise Linux 8.2 for x8
□ SUSE Linux Enterprise Server 15 S

Kernel
□ Inbox:4.18.0-147
□ Inbox:4.18.0-193
□ Inbox:4.18.0-80
□ OOB:4.12.14-150100.197.114

Host Bus Adapter
□ Emulex LPe32000
□ Emulex LPe32002
□ Emulex LPe32004
□ Emulex LPe35000
□ Emulex LPe35002
□ QLogic QLE2740
□ QLogic QLE2742

HBA Firmware Version
□ 12.4.243.17
□ 12.6.182.8
□ 12.6.240.22
□ 12.8.351.47
□ 8.08.231
□ 9.02.02
□ 9.03.40

HBA Driver Version
□ Inbox:12.2.0.0
□ OOB:10.01.00.55.08.0
□ OOB:10.01.00.55.15.1
□ OOB:10.01.00.63.08.2
□ OOB:12.6.182.4
□ OOB:12.6.240.27

Multipathing Software
□ UltraPath

Storage System Version
□ 6.0.1/6.1.0/6.1.2/6.1.3/6.1.5/6.1.6
□ 6.1.0/6.1.2/6.1.3/6.1.5/6.1.6

Array Interface Module
□ 32Gbps/16Gbps/8Gbps FC

----End

2.4 Limitations and Constraints

1. HBAs of different vendors or different models from the same vendor cannot be used together on the host.
2. The LUN whose host LUN ID is 0 cannot be mapped.
3. The space reclamation granularity cannot exceed 64 MB. You must set **discard_max_bytes** of the namespace block device on the host to less than or equal to 64 MB. For details, see **7.4 How Can I Modify the Granularity for Reclaiming Thin LUNs on a Host?**. Manual space reclamation by **fstrim** is supported, but automatic space reclamation by running the **mount -o discard** command is not supported.
4. SAN boot is not supported.
5. Heterogeneous virtualization takeover is not supported.
6. A LUN cannot be shared by the SCSI and NVMe protocols.
7. The timeout processing mechanism of the NVMe protocol is immature. In the event of link disconnection and recovery or mode changes of storage interface modules, there is a possibility that the host cannot automatically establish the connection. In this case, you must manually remove and insert the cable or restart the host. For details, see the FAQ.
8. If you use FC-NVMe to connect the storage system to the host and create multiple logical hosts on the storage system for the same physical host (these logical hosts use initiators of different HBAs on the physical host), you must

assign different host LUN IDs when mapping LUNs to these logical hosts. Otherwise, LUNs with the same host LUN ID cannot be identified.

2.5 Common Commands

Table 2-2 lists the common management commands used in Linux hosts.

Table 2-2 Commands

Command	Function
df	Views the file system size and usage.
fdisk /dev/mapper/mpath#	Partitions disks.
ifconfig	Configures network port parameters.
lsscsi	Displays the hardware address, type, and manufacturer of each disk.
lvdisplay -v /dev/vgname/lvname	Views details about **lvname**.
mount	Mounts a logical volume.
shutdown -h now	Shuts down the host.
shutdown –ry 0	Restarts the host.
vgdisplay -v vgname	Views details about **vgname**.
vgscan	Scans for volume groups in the system.
nvme list	Queries NVMe disks.
nvme ns-rescan /dev/nvme#	Scans for NVMe disks.
nvme discover	Discovers NVMe targets (for NVMe over RoCE).
nvme connect	Accesses NVMe targets (for NVMe over RoCE).

NOTE

The pound (**#**) in the table indicates a number that can be specified based on actual conditions.

3 Planning Connectivity

Hosts and storage systems can be connected based on different criteria. **Table 3-1** describes the typical connection modes.

Table 3-1 Connection modes

Criteria	Connection Mode
Interface module type	FC-NVMe connection/NVMe over RoCE connection
Whether switches are used	Direct-attached connection/Fabric-attached connection (the switches must be included in the compatibility list)
Whether multiple paths exist	Single-path connection/Multi-path connection
Whether HyperMetro is configured	Non-HyperMetro

The following details the connections in various scenarios.

3.1 FC-NVMe Connectivity

You are advised to use **LLDesigner** to plan connectivity based on site requirements and export low level design (LLD) files. This section describes how to plan connectivity based on **LLDesigner**.

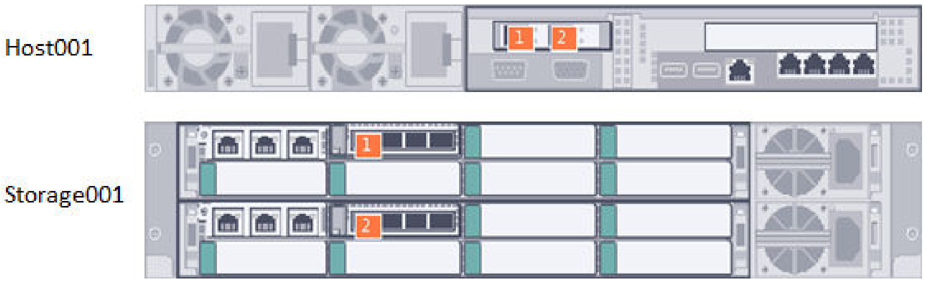
3.1.1 Direct-Attached FC-NVMe Connections

This section describes how to directly connect a host to a two-controller storage system and a four-controller storage system through FC-NVMe multi-path connections.

Two-Controller Storage System

Figure 3-1 shows how to directly connect a host to a 2 U two-controller storage system through FC-NVMe multi-path connections.

Figure 3-1 Direct-attached FC-NVMe multi-path connections (two-controller storage system)



NOTE

- In this connection diagram, each of the two controllers is connected to a host HBA port with an optical fiber. The cable connections are detailed in **Table 3-2**.
- For better performance and reliability, you are advised to deploy two dual-port FC HBAs on the host and two FC front-end interface modules on each storage controller. Then use a separate optical fiber to connect each of the four FC front-end interface modules on the storage system to each of the HBA ports on the host.

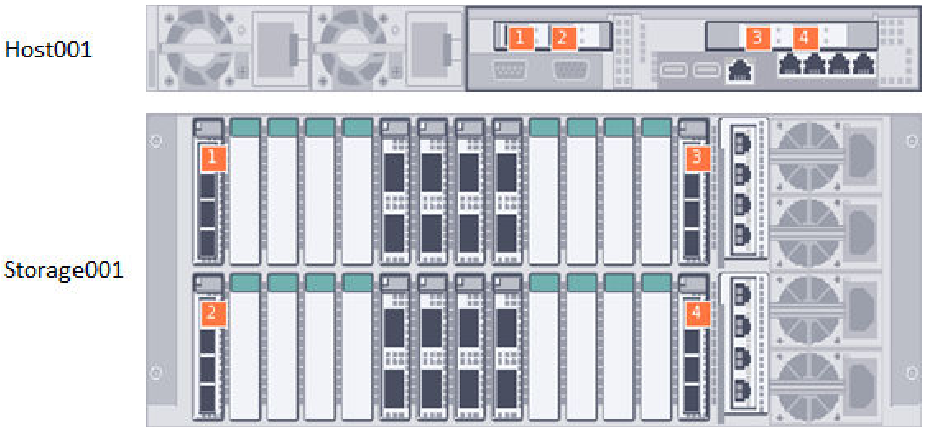
Table 3-2 Cable connections (two-controller storage system)

Cable No.	Description
1	Connects Port Slot1.P0 on Host001 to Port A.IOM0.P0 on Storage001.
2	Connects Port Slot1.P1 on Host001 to Port B.IOM0.P0 on Storage001.

Four-Controller Storage System

Figure 3-2 shows how to directly connect a host to a 4 U four-controller storage system through FC-NVMe multi-path connections.

Figure 3-2 Direct-attached FC-NVMe multi-path connections (four-controller storage system)



NOTE

In this connection diagram, each front-end interface module is fully interconnected with the four controllers and therefore can be accessed by all of them.

Table 3-3 Cable connections (four-controller storage system)

Cable No.	Description
1	Connects Port Slot1.P0 on Host001 to Port IOM.H0.P0 on Storage001.
2	Connects Port Slot1.P1 on Host001 to Port IOM.L0.P0 on Storage001.
3	Connects Port Slot2.P0 on Host001 to Port IOM.H13.P0 on Storage001.
4	Connects Port Slot2.P1 on Host001 to Port IOM.L13.P0 on Storage001.

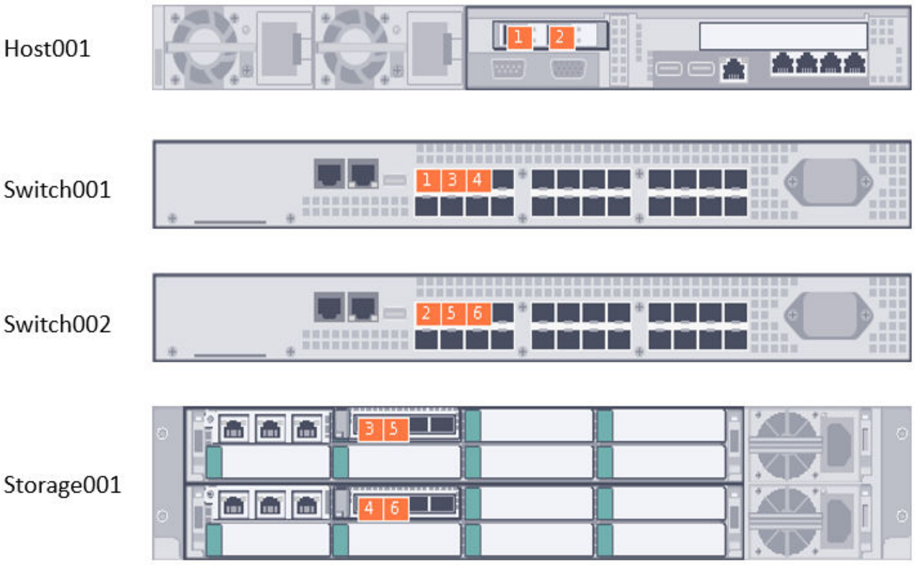
3.1.2 Fabric-Attached FC-NVMe Connections

This section describes how to connect a host to a two-controller storage system and a four-controller storage system through FC-NVMe multi-path connections using switches.

Two-Controller Storage System

Figure 3-3 shows how to connect a host to a 2 U two-controller storage system through FC-NVMe multi-path connections using switches.

Figure 3-3 Fabric-attached FC-NVMe multi-path connections (two-controller storage system)



NOTE

In this connection diagram, two controllers of the storage system and two ports of the host are connected to switches through optical fibers. On the switches, the ports connecting to a storage controller and to the host are grouped in a zone, ensuring connectivity between the host port and the storage system.

Table 3-4 Zone division on switches (two-controller storage system)

Switch Name	Zone Name	Zone Member
Switch001	Zone001	Ports 1 and 3
Switch001	Zone002	Ports 1 and 4
Switch002	Zone003	Ports 2 and 5
Switch002	Zone004	Ports 2 and 6

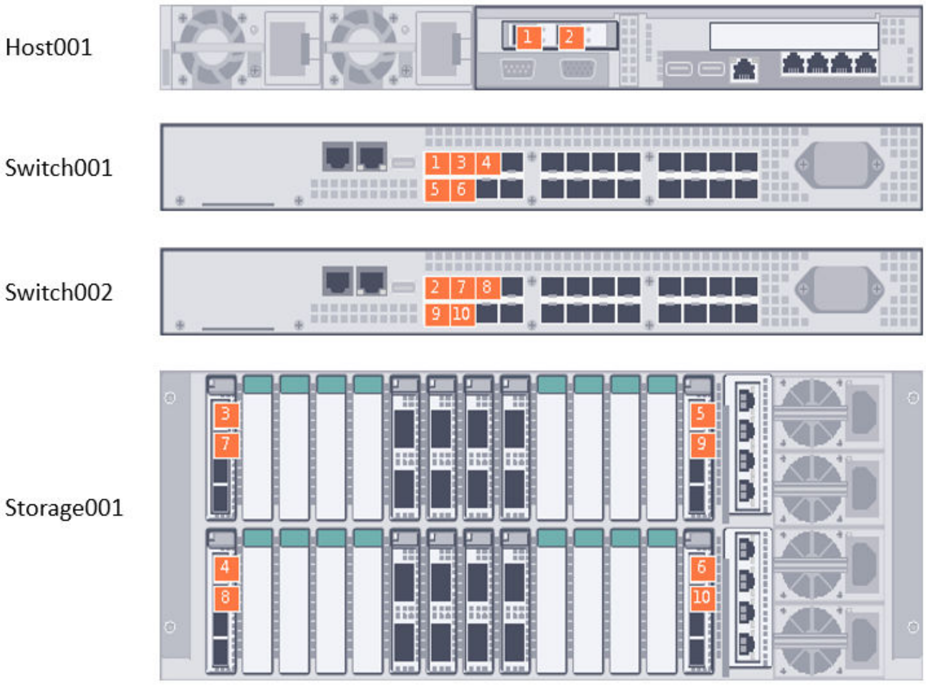
NOTE

- Port numbers in the **Zone Member** column in this table refer to numbers in **Figure 3-3** rather than switch port IDs.
- Zone division in this table is for reference only. Plan zones based on site requirements.
- If you use **LLDesigner** to plan connectivity, you can obtain zone division data from the **Zone Planning** worksheet in the exported LLD file.

Four-Controller Storage System

Figure 3-4 shows how to connect a host to a 4 U four-controller storage system through FC-NVMe multi-path connections using switches.

Figure 3-4 Fabric-attached FC-NVMe multi-path connections (four-controller storage system)



NOTE

In this connection diagram, four controllers of the storage system and two ports of the host are connected to switches through optical fibers. On the switches, the ports connecting to a storage controller and to the host are grouped in a zone, ensuring connectivity between the host port and the storage system.

Table 3-5 Zone division on switches (four-controller storage system)

Switch Name	Zone Name	Zone Member
Switch001	Zone001	Ports 1 and 3
Switch001	Zone002	Ports 1 and 4
Switch001	Zone003	Ports 1 and 5
Switch001	Zone004	Ports 1 and 6
Switch002	Zone005	Ports 2 and 7
Switch002	Zone006	Ports 2 and 8
Switch002	Zone007	Ports 2 and 9
Switch002	Zone008	Ports 2 and 10

NOTE

- Port numbers in the **Zone Member** column in this table refer to numbers in **Figure 3-4** rather than switch port IDs.
- Zone division in this table is for reference only. Plan zones based on site requirements.
- If you use **LLDesigner** to plan connectivity, you can obtain zone division data from the **Zone Planning** worksheet in the exported LLD file.

3.2 NVMe over RoCE Connectivity

You are advised to use **LLDesigner** to plan connectivity based on site requirements and export low level design (LLD) files. This section describes how to plan connectivity based on **LLDesigner**.

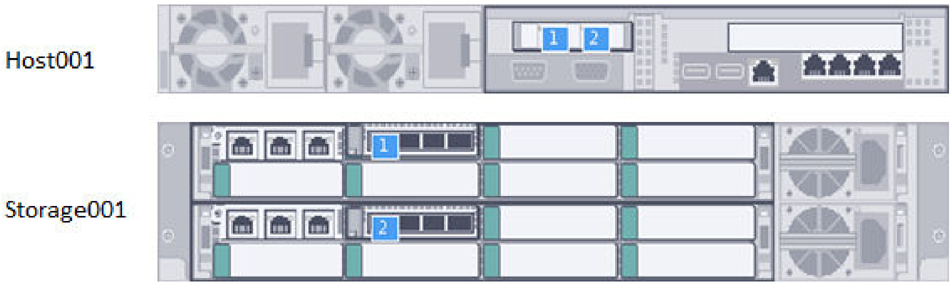
3.2.1 Direct-Attached NVMe over RoCE Connections

This section describes how to directly connect a host to a two-controller storage system and a four-controller storage system through NVMe over RoCE connections.

Two-Controller Storage System

Figure 3-5 shows how to directly connect a host to a 2 U two-controller storage system through NVMe over RoCE connections.

Figure 3-5 Direct-attached NVMe over RoCE connections (two-controller storage system)



NOTE

In this connection diagram, each of the two controllers is connected to a host NIC port with an Ethernet cable.

Table 3-6 IP address plan (two-controller storage system)

Port	Description	VLAN ID	IP Address	Subnet Mask
Port Slot1.P0 on Host001	Connects to Port A.IOM0.P0 on Storage001.	55	192.168.5.5	255.255.255.0

Port	Description	VLAN ID	IP Address	Subnet Mask
Port Slot1.P1 on Host001	Connects to Port B.IOM0.P0 on Storage001.	66	192.168.6.5	255.255.255.0
Port A.IOM0.P0 on Storage001	Connects to Port Slot1.P0 on Host001.	55	192.168.5.6	255.255.255.0
Port B.IOM0.P0 on Storage001	Connects to Port Slot1.P1 on Host001.	66	192.168.6.6	255.255.255.0

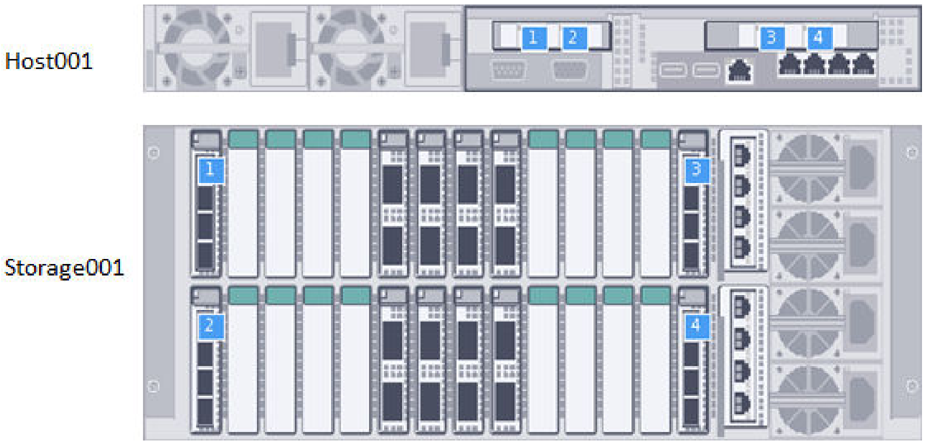
NOTE

- IP addresses in this table are for reference only. Plan IP addresses based on site requirements.
- If you use LLDesigner to plan connectivity, you can obtain IP address data from the **IP Address Planning** worksheet in the exported LLD file.

Four-Controller Storage System

Figure 3-6 shows how to directly connect a host to a 4 U four-controller storage system through NVMe over RoCE connections.

Figure 3-6 Direct-attached NVMe over RoCE connections (four-controller storage system)



NOTE

In this connection diagram, each of the four controllers is connected to a host NIC port with an Ethernet cable.

Table 3-7 IP address plan (four-controller storage system)

Port	Description	VLAN ID	IP Address	Subnet Mask
Port Slot1.P0 on Host001	Connects to Port IOM.H0.P0 on Storage001.	55	192.168.5.5	255.255.255.0
Port Slot1.P1 on Host001	Connects to Port IOM.L0.P0 on Storage001.	66	192.168.6.5	255.255.255.0
Port Slot2.P0 on Host001	Connects to Port IOM.H13.P0 on Storage001.	77	192.168.7.5	255.255.255.0
Port Slot2.P1 on Host001	Connects to Port IOM.L13.P0 on Storage001.	88	192.168.8.5	255.255.255.0
Port IOM.H0.P0 on Storage001	Connects to Port Slot1.P0 on Host001.	55	192.168.5.6	255.255.255.0
Port IOM.L0.P0 on Storage001	Connects to Port Slot1.P1 on Host001.	66	192.168.6.6	255.255.255.0
Port IOM.H13.P0 on Storage001	Connects to Port Slot2.P0 on Host001.	77	192.168.7.6	255.255.255.0
Port IOM.L13.P0 on Storage001	Connects to Port Slot2.P1 on Host001.	88	192.168.8.6	255.255.255.0

NOTE

- IP addresses in this table are for reference only. Plan IP addresses based on site requirements.
- If you use LLDesigner to plan connectivity, you can obtain IP address data from the **IP Address Planning** worksheet in the exported LLD file.

3.2.2 Fabric-Attached NVMe over RoCE Connections

This section describes how to connect a host to a two-controller storage system and a four-controller storage system through NVMe over RoCE connections using switches.

Two-Controller Storage System

Figure 3-7 shows how to connect a host to a 2 U two-controller storage system through NVMe over RoCE connections using switches.

Figure 3-7 Fabric-attached NVMe over RoCE connections (two-controller storage system)

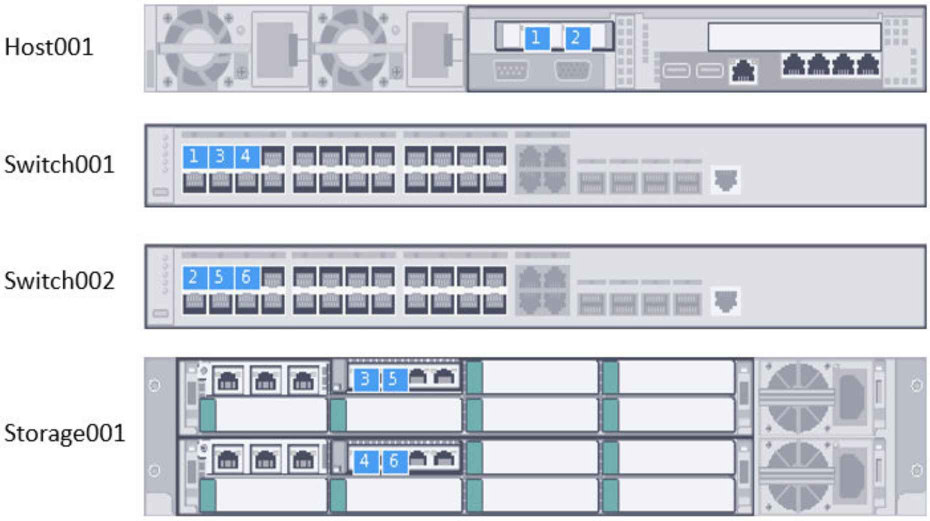


Table 3-8 IP address plan (two-controller storage system)

Port	Description	VLAN ID	IP Address	Subnet Mask
Port Slot1.P0 on Host001	Connects to port A.IOM0.P0 and port B.IOM0.P0 of Storage001 through Switch001.	55	192.168.5.5	255.255.255.0
Port Slot1.P1 on Host001	Connects to port A.IOM0.P1 and port B.IOM0.P1 of Storage001 through Switch002.	66	192.168.6.5	255.255.255.0
Port A.IOM0.P0 on Storage001	Connects to port Slot1.P0 on Host001 through Switch001.	55	192.168.5.6	255.255.255.0
Port A.IOM0.P1 on Storage001	Connects to port Slot1.P1 on Host001 through Switch002.	66	192.168.6.6	255.255.255.0
Port B.IOM0.P0 on Storage001	Connects to port Slot1.P0 on Host001 through Switch001.	55	192.168.5.7	255.255.255.0
Port B.IOM0.P1 on Storage001	Connects to port Slot1.P1 on Host001 through Switch002.	66	192.168.6.7	255.255.255.0

Table 3-9 Domain plan (two-controller storage system)

Domain Name	Domain Member
domain001	192.168.5.5, 192.168.5.6
domain002	192.168.5.5, 192.168.5.7
domain003	192.168.6.5, 192.168.6.6
domain004	192.168.6.5, 192.168.6.7

NOTE

- In this connection diagram, two controllers of the storage system and two NIC ports of the host are connected to switches through Ethernet cables, ensuring the connectivity between the host ports and the storage.
- If you use **LLDesigner** to plan connectivity, you can obtain IP address data from the **IP Address Planning** worksheet in the exported LLD file.

Four-Controller Storage System

Figure 3-8 shows how to connect a host to a 4 U four-controller storage system through NVMe over RoCE connections using switches.

Figure 3-8 Fabric-attached NVMe over RoCE connections (four-controller storage system)

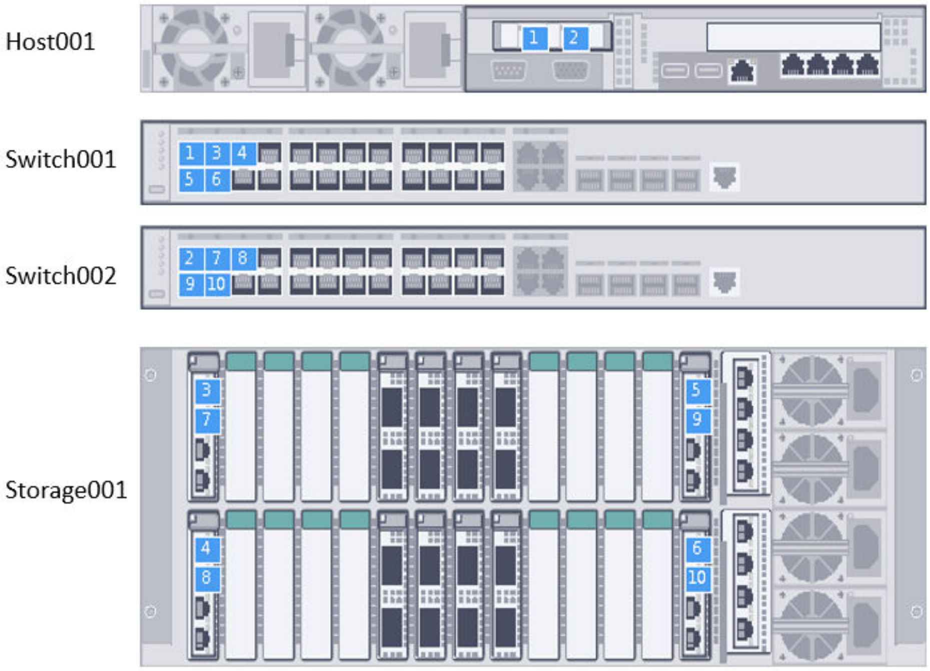


Table 3-10 IP address plan (four-controller storage system)

Port	Description	VLAN ID	IP Address	Subnet Mask
Port Slot1.P0 on Host001	Connects to ports IOM.H0.P0, IOM.L0.P0, IOM.H13.P0, and IOM.L13.P0 of Storage001 through Switch001.	55	192.168.5.5	255.255.255.0
Port Slot1.P1 on Host001	Connects to ports IOM.H0.P1, IOM.L0.P1, IOM.H13.P1, and IOM.L13.P1 of Storage001 through Switch002.	66	192.168.6.5	255.255.255.0
Port IOM.H0.P0 on Storage001	Connects to port Slot1.P0 on Host001 through Switch001.	55	192.168.5.6	255.255.255.0
Port IOM.L0.P0 on Storage001	Connects to port Slot1.P0 on Host001 through Switch001.	55	192.168.5.7	255.255.255.0
Port IOM.H13.P0 on Storage001	Connects to port Slot1.P0 on Host001 through Switch001.	55	192.168.5.8	255.255.255.0
Port IOM.L13.P0 on Storage001	Connects to port Slot1.P0 on Host001 through Switch001.	55	192.168.5.9	255.255.255.0
Port IOM.H0.P1 on Storage001	Connects to port Slot1.P1 on Host001 through Switch002.	66	192.168.6.6	255.255.255.0
Port IOM.L0.P1 on Storage001	Connects to port Slot1.P1 on Host001 through Switch002.	66	192.168.6.7	255.255.255.0
Port IOM.H13.P1 on Storage001	Connects to port Slot1.P1 on Host001 through Switch002.	66	192.168.6.8	255.255.255.0
Port IOM.L13.P1 on Storage001	Connects to port Slot1.P1 on Host001 through Switch002.	66	192.168.6.9	255.255.255.0

Table 3-11 Domain plan (four-controller storage system)

Domain Name	Domain Member
domain001	192.168.5.5, 192.168.5.6
domain002	192.168.5.5, 192.168.5.7
domain003	192.168.5.5, 192.168.5.8
domain004	192.168.5.5, 192.168.5.9
domain005	192.168.6.5, 192.168.6.6
domain006	192.168.6.5, 192.168.6.7
domain007	192.168.6.5, 192.168.6.8
domain008	192.168.6.5, 192.168.6.9

NOTE

- In this connection diagram, four controllers of the storage system and two NIC ports of the host are connected to switches through Ethernet cables, ensuring the connectivity between the host ports and the storage.
- If you use **LLDesigner** to plan connectivity, you can obtain IP address data from the **IP Address Planning** worksheet in the exported LLD file.

4 Preparing for Configuration

This chapter describes the preparations on switches, storage systems, and hosts. Before the configuration, ensure that all of the components are included in the compatibility list. For details on how to query compatibility, see **2.3 Interoperability Query**.

4.1 Preparations for FC-NVMe Connections

4.1.1 FC Switch Configuration

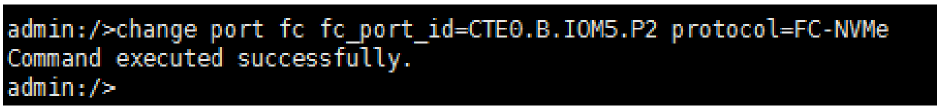
The method and requirements for configuring FC switches for FC-NVMe connections are the same as those for FC-SCSI connections. Follow the correct configuration guide based on your network type and switch model.

4.1.2 Storage System Configuration

Create storage pools, LUNs/LUN groups, and hosts/host groups on the storage system and map the LUNs/LUN groups to the hosts/host groups according to your service requirements. For details, see the *Basic Storage Service Configuration Guide for Block*.

Change the protocol to FC-NVMe for the storage ports that connect to the host, as shown in **Figure 4-1**.

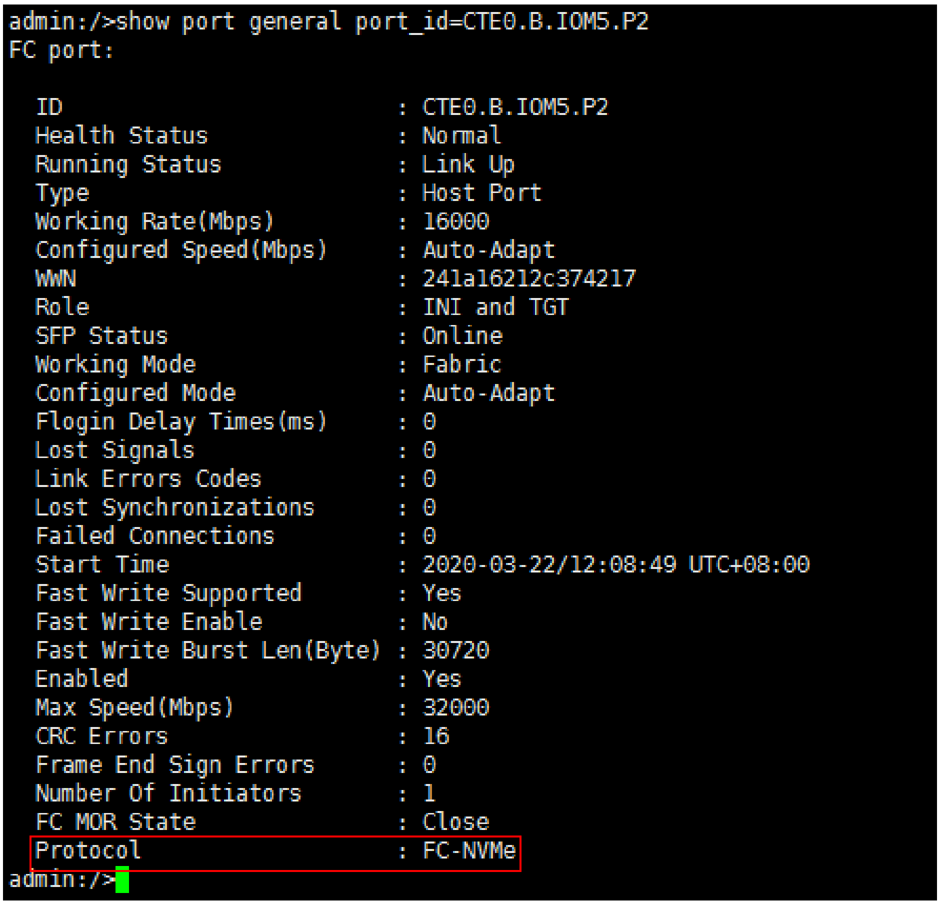
Figure 4-1 Changing the port mode on the storage system



```
admin:/>change port fc fc_port_id=CTE0.B.IOM5.P2 protocol=FC-NVMe
Command executed successfully.
admin:/>
```

After the change, run the following command to verify that the port mode is correctly set.

Figure 4-2 Verifying the port mode on the storage system



```
admin:/>show port general port_id=CTE0.B.IOM5.P2
FC port:

  ID                          : CTE0.B.IOM5.P2
  Health Status               : Normal
  Running Status              : Link Up
  Type                        : Host Port
  Working Rate(Mbps)          : 16000
  Configured Speed(Mbps)      : Auto-Adapt
  WWN                         : 241a16212c374217
  Role                        : INI and TGT
  SFP Status                  : Online
  Working Mode                : Fabric
  Configured Mode             : Auto-Adapt
  Flogin Delay Times(ms)      : 0
  Lost Signals                : 0
  Link Errors Codes           : 0
  Lost Synchronizations       : 0
  Failed Connections          : 0
  Start Time                  : 2020-03-22/12:08:49 UTC+08:00
  Fast Write Supported        : Yes
  Fast Write Enable           : No
  Fast Write Burst Len(Byte)  : 30720
  Enabled                     : Yes
  Max Speed(Mbps)             : 32000
  CRC Errors                  : 16
  Frame End Sign Errors       : 0
  Number Of Initiators        : 1
  FC MOR State                : Close
  Protocol                    : FC-NVMe
admin:/>
```

NOTICE

- The storage ports must be idle FC ports and services can be configured after the port protocol has been changed. If the storage ports have been carrying services using another protocol, the services will become abnormal after the protocol is changed.
- To enable FC-NVMe for storage ports on OceanStor Dorado 6.0.1, apply for the corresponding patch from Huawei support website and install it. For details on how to apply for and install the patch, contact Huawei technical support engineers. In OceanStor Dorado 6.1.0 and later versions, you do not need to install the patch.
- In HyperMetro storage scenarios, the Fast Write function cannot be enabled on both the storage devices and switches (this function is called Fast Write on Brocade switches and Write Acceleration on Cisco switches).

4.1.3 Identifying FC HBAs of the Host

Before connecting a host to a storage system using FC-NVMe, make sure that the host HBAs have been identified and are functioning properly. You also need to obtain the world wide names (WWNs) of HBA ports for subsequent storage system configurations.

The following describes how to query the attributes of QLogic and Emulex HBAs, including their models, driver versions, firmware versions, WWNs, port topologies, and port rates. To query other attributes or other vendors' HBAs, it is recommended that you use the management software provided by the respective HBA vendor. See the operation guide of the corresponding HBA management software for detailed operations.

4.1.3.1 Emulex FC HBA

To query the Emulex HBA model:

```
[root@localhost ~]# cat /sys/class/scsi_host/host*/modelname
LPe32002-M2
LPe32002-M2
```

To query the Emulex HBA driver version:

```
[root@localhost ~]# cat /sys/class/scsi_host/host*/lpfc_drvr_version
Emulex LightPulse Fibre Channel SCSI driver 12.6.182.4
Emulex LightPulse Fibre Channel SCSI driver 12.6.182.4
```

To query the Emulex HBA firmware version:

```
[root@localhost ~]# cat /sys/class/scsi_host/host*/fwrev
12.4.243.17, sli-4:2:c
12.4.243.17, sli-4:2:c
```

To query the Emulex HBA WWN:

```
[root@localhost ~]# cat /sys/class/fc_host/host*/port_name
0x100000109b32a3c0
0x100000109b32a3c1
```

To query the Emulex HBA topology:

```
[root@localhost ~]# cat /sys/class/fc_host/host*/port_type
NPort (fabric via point-to-point)
NPort (fabric via point-to-point)
```

To query the Emulex HBA port rate:

```
[root@localhost ~]# cat /sys/class/fc_host/host*/speed
32 Gbit
32 Gbit
```

4.1.3.2 QLogic FC HBA

To query the QLogic HBA model:

```
[root@localhost ~]# cat /sys/class/scsi_host/host*/model_name
QLE2742
QLE2742
```

To query the QLogic HBA driver version:

```
[root@localhost ~]# cat /sys/class/scsi_host/host*/driver_version
10.01.00.55.08.0-k-debug
10.01.00.55.08.0-k-debug
```

To query the QLogic HBA firmware version:

```
[root@localhost ~]# cat /sys/class/scsi_host/host*/optrom_fw_version
8.08.231 2019217621
8.08.231 2019217621
```

To query the QLogic HBA WWN:

```
[root@localhost ~]# cat /sys/class/fc_host/host*/port_name
0x2100000e1e1e0091
0x2100000e1e1e0090
```

To query the QLogic HBA topology:

```
[root@localhost ~]# cat /sys/class/fc_host/host*/port_type
NPort (fabric via point-to-point)
NPort (fabric via point-to-point)
```

To query the QLogic HBA port rate:

```
[root@localhost ~]# cat /sys/class/fc_host/host*/speed
32 Gbit
32 Gbit
```

4.2 Preparations for NVMe over RoCE Connections

4.2.1 Ethernet Switch Configuration

NVMe over RoCE has high requirements on the network. Switches must support lossless Ethernet and priority-based flow control (PFC) deadlock detection, suppression, and isolation.

- PFC

  Set the PFC priority to 3, enable PFC on all switch ports to be used, and configure hardware-based PFC deadlock detection.

- (Optional, recommended) AI-ECN

  Configure the low-latency fabric, disable dynamic ECN, and then enable and activate AI-ECN.

- VLAN

  Configure VLANs based on the network plan.

- (Optional) iNOF (mandatory if the OceanStor NOF Director is used)

  OceanStor NOF Director simplifies service deployment on the conventional Ethernet (lossless RoCE network) and can quickly detect faults and perform service switchover. If OceanStor NOF Director is used, iNOF must be configured on the switches. To configure iNOF on a switch, enable LLDP on the switch and configure the iNOF reflector and client. You must create user-defined domains on the switch according to the service plan. Currently, this feature is supported only by Huawei CE6866 and CE8851 switches. For the detailed configuration methods, see the **iNOF configuration guide** in the switch documentation.

NOTICE

- It is recommended that dual planes be configured to ensure high network reliability. Do not use leaf stacking or M-LAG.
- For details about the switch models supported by the NVMe over RoCE feature, refer to the **Huawei Storage Interoperability Navigator**.
- The configuration methods for switches of different models and vendors may vary. For details, see the configuration guide released by the switch vendor.

The configurations of an Ethernet network vary with scenarios and service requirements. This section only provides the basic configurations of common Huawei switches for the NVMe over RoCE connections. In practice, configure the network based on the customer requirements and network plan.

4.2.1.1 Huawei CE6866, CE8851, CE6860-SAN, and CE8850-SAN

The methods for configuring Huawei CE6866, CE8851, CE6860-SAN, and CE8850-SAN switches are the same. The following uses CE6866 as an example to describe how to configure PFC, AI-ECN, VLANs, and iNOF.

PFC Configuration

Step 1 Enable PFC priority 3.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]dcb pfc
[~CE6866-48S8CQ-P-dcb-pfc-default]priority 3
[*CE6866-48S8CQ-P-dcb-pfc-default]commit
[~CE6866-48S8CQ-P-dcb-pfc-default]quit
[~CE6866-48S8CQ-P]
```

Run the following command to verify that the PFC priority configuration has taken effect:

```
[~CE6866-48S8CQ-P]display dcb pfc-profile default
------------------------------------------------------------------------------
PFC-profile Name                    Priority
------------------------------------------------------------------------------
default                          3
------------------------------------------------------------------------------
[~CE6866-48S8CQ-P]
```

Step 2 Enable PFC on the switch ports.

The following uses port 25GE1/0/26 as an example. Run the **dcb pfc enable mode manual** command to enable PFC on the port.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]interface 25GE1/0/26
[~CE6866-48S8CQ-P-25GE1/0/26]dcb pfc enable mode manual
[*CE6866-48S8CQ-P-25GE1/0/26]commit
[~CE6866-48S8CQ-P-25GE1/0/26]quit
[~CE6866-48S8CQ-P]
```

Run the following command to verify that PFC has been enabled on the 25GE1/0/26 port:

```
[~CE6866-48S8CQ-P]display dcb
M:Manual;    A:Auto
-------------------------------------------------------------------------------
Interface         PFC Name     PFC Status  ETS Name    ETS Status App-Profile
-------------------------------------------------------------------------------
25GE1/0/26       default      ENABLE(M)   -           -          -
-------------------------------------------------------------------------------
[~CE6866-48S8CQ-P]
```

NOTE

Repeat this operation on all switch ports to be used, including the ports connecting to the host and storage system.

Step 3 Configure PFC deadlock detection.

Configure PFC deadlock detection for the PFC priority, and set the detection time to 1000 ms and recovery time to 1500 ms.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]dcb pfc deadlock-detect timer 1000
[*CE6866-48S8CQ-P]dcb pfc deadlock-recovery timer 1500
[*CE6866-48S8CQ-P]commit
```

Verify the configuration.

```
[~CE6866-48S8CQ-P]display this | include dcb
dcb pfc deadlock-detect timer 1000
dcb pfc deadlock-recovery timer 1500
[~CE6866-48S8CQ-P]
```

Step 4 Set the threshold for disabling PFC on a port to 20 deadlocks in 1 minute.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]dcb pfc

[*CE6866-48S8CQ-P-dcb-pfc-default]priority 3 turn-off threshold 20
[*CE6866-48S8CQ-P-dcb-pfc-default]commit
[~CE6866-48S8CQ-P-dcb-pfc-default]quit
[~CE6866-48S8CQ-P]
```

Verify the configuration.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]dcb pfc
[~CE6866-48S8CQ-P-dcb-pfc-default]display this
#
dcb pfc

 priority 3 turn-off threshold 20
#
return
[~CE6866-48S8CQ-P-dcb-pfc-default]quit
[~CE6866-48S8CQ-P]
```

----End

(Optional, Recommended) AI-ECN Configuration

You are advised to configure AI-ECN to dynamically and intelligently set the threshold based on the network traffic.

Step 1 Before the configuration, check whether the switch has the AI-ECN license.

If the value of **CE-LIC-AFRD** is **YES**, the switch has the AI-ECN license. Otherwise, you must load the license.

```
[~CE6866-48S8CQ-P]display license
MainBoard:
 Active License    : flash:/LICCloudEngine6800_V300R020_20200819SDQG5M.xml
 License state     : Demo
 Revoke ticket     : No ticket
 License mode      : common

 RD of Huawei Technologies Co., Ltd.

 Product name      : CloudEngine 6800
 Product version   : V300R020
 License Serial No : LIC20200819SDQG5M
 Creator           : Huawei Technologies Co., Ltd.
 Created Time      : 2020-08-19 16:01:01
 ----------------------------------------------------------------
 Feature name      : Trial0
 Authorize type    : demo
 Expired date      : 2020-11-12
 Trial days        : --

 Item name          Item type      Value      Description
 ----------------------------------------------------------------
 N1-CE68LIC-AKA      --             1          N1-AK Advanced SW License for CloudEngine 6800
  CE-LIC-NSH        Function       YES        CE-LIC-NSH
  CE-LIC-TLM        Function       YES        CE-LIC-TLM
  CE-LIC-BASE       Function       YES        CE-LIC-BASE
 CE68-LIC-AFRD       --             1          CloudEngine 6800 RDMA AI Fabric Application Acceleration Basic 
Function
  CE-LIC-AFRD       Function       YES        CE-LIC-AFRD
 N1-CE68LIC-iNOF     --             1          N1-CloudEngine 6800 AI Fabric Value-added Package for the iNOF 
Storage Scenarios
  CE-LIC-AFRD       Function       YES        CE-LIC-AFRD
  CE-LIC-iNOF       Function       YES        CE-LIC-iNOF
 CE68-LIC-AFV        --             1          CE6800 Anyflow Visibility Function
  CE-LIC-AFV        Function       YES        CE-LIC-AFV
 CE68-LIC-TLM        --             1          CE6800 Telemetry Function
  CE-LIC-TLM        Function       YES        CE-LIC-TLM
 N1-CE68LIC-PLLV     --             1          N1-CloudEngine 6800 Packet Loss and Latency Visibility Function
  CE-LIC-PLLV       Function       YES        CE-LIC-PLLV
 CE68-LIC-iNOF       --             1          CloudEngine 6800 for AI Fabric iNOF storage
  CE-LIC-iNOF       Function       YES        CE-LIC-iNOF
 N1-CE68LIC-AFV      --             1          N1-CloudEngine 6800 Anyflow Visibility Function
  CE-LIC-AFV        Function       YES        CE-LIC-AFV
 N1-CE68LIC-CFAD     --             1          N1-CloudFabric Advanced SW License for CloudEngine 6800
  CE-LIC-NSH        Function       YES        CE-LIC-NSH
  CE-LIC-TLM        Function       YES        CE-LIC-TLM
  CE-LIC-PTP        Function       YES        CE-LIC-PTP
  CE-LIC-BASE       Function       YES        CE-LIC-BASE
 N1-CE68UPG-M-A      --             1          N1-CloudEngine 6800 Upgrade SW License:Management to 
Advanced
  CE-LIC-NSH        Function       YES        CE-LIC-NSH
  CE-LIC-TLM        Function       YES        CE-LIC-TLM
  CE-LIC-PTP        Function       YES        CE-LIC-PTP
 N1-CE68LIC-AFRD-2   --             1          N1-CloudEngine 6800 AI Fabric RDMA Application Acceleration 
Function 2
  CE-LIC-AFRD       Function       YES        CE-LIC-AFRD
 N1-CE68LIC-IAF      --             1          N1-CE6800 Intelligent Analysis Function
  CE-LIC-TLM        Function       YES        CE-LIC-TLM
 N1-CE68LIC-CFFD     --             1          N1-CloudFabric Foundation SW License for CloudEngine 6800
  CE-LIC-TLM        Function       YES        CE-LIC-TLM
  CE-LIC-PTP        Function       YES        CE-LIC-PTP
  CE-LIC-BASE       Function       YES        CE-LIC-BASE
 N1-CE68UPG-F-A      --             1          N1-CloudEngine 6800 Upgrade SW License:Foundation to Advanced
  CE-LIC-NSH        Function       YES        CE-LIC-NSH
 N1-CE68LIC-CFMM     --             1          N1-CloudFabric Management SW License for CloudEngine 6800
  CE-LIC-BASE       Function       YES        CE-LIC-BASE
```

```
N1-CE68LIC-BS      --             1          N1-CloudEngine 6800 Basic Function
 CE-LIC-BASE       Function       YES        CE-LIC-BASE
CE68-LIC-PTP       --             1          CE6800 Precision Time Protocol Function
 CE-LIC-PTP        Function       YES        CE-LIC-PTP
CE68-LIC-PLLV      --             1          CE6800 Packet Loss and Latency Visibility Function
 CE-LIC-PLLV       Function       YES        CE-LIC-PLLV
CE68-LIC-BASE      --             1          CE6800 Basic Software Function
 CE-LIC-BASE       Function       YES        CE-LIC-BASE
N1-CE68LIC-ADA     --             1          N1-CloudEngine 6800 Advantage Function A
 CE-LIC-TLM        Function       YES        CE-LIC-TLM
```

Step 2 Enable AI-ECN and specify the lossless queues with AI-ECN enabled.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]ai-service
[*CE6866-48S8CQ-P-ai-service]ai-ecn
[*CE6866-48S8CQ-P-ai-service-ai-ecn]ai-ecn enable
[*CE6866-48S8CQ-P-ai-service-ai-ecn]assign queue 3
[*CE6866-48S8CQ-P-ai-service-ai-ecn]commit
[~CE6866-48S8CQ-P-ai-service-ai-ecn]quit
[~CE6866-48S8CQ-P-ai-service]quit
```

Step 3 Run the following command to verify that AI-ECN has been activated:

```
[~CE6866-48S8CQ-P]display ai-ecn calculated state
AI-ECN active model: AI_ECN, version 1.0.0, actived time: 2020-08-19 17:25:49
------------------------------------------------------------------------------
Interface       Queue   Low-Threshold   High-Threshold   Probability   Mode
                          (Byte)          (Byte)           (%)
------------------------------------------------------------------------------
25GE1/0/26        3          5120          409600           5    BBR
25GE1/0/29        3          5120          409600           5    BBR
25GE1/0/30        3          5120          409600           5    BBR
25GE1/0/31        3          5120          409600           5    BBR
25GE1/0/32        3          5120          409600           5    BBR
25GE1/0/33        3          5120          409600           5    BBR
------------------------------------------------------------------------------
[~CE6866-48S8CQ-P]
```

----End

VLAN Configuration

Step 1 Configure global VLANs. The following example uses VLAN 55 and VLAN 66.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.

[~CE6866-48S8CQ-P]vlan batch 55 66
[*CE6866-48S8CQ-P]commit
[~CE6866-48S8CQ-P]quit
<CE6866-48S8CQ-P>
```

Step 2 Configure VLANs for ports.

The following uses port 25GE1/0/26 as an example. Configure the port to work in trunk mode and then add the port to VLANs 55 and 66. Then set the port to an edge port.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
Warning: The current device is single master board. Exercise caution when performing this operation.
[~CE6866-48S8CQ-P]interface 25GE1/0/26
[~CE6866-48S8CQ-P-25GE1/0/26]port link-type trunk
[*CE6866-48S8CQ-P-25GE1/0/26]port trunk allow-pass vlan 55 66
[*CE6866-48S8CQ-P-25GE1/0/26]stp edged-port enable
[*CE6866-48S8CQ-P-25GE1/0/26]commit
```

```
[~CE6866-48S8CQ-P]quit
<CE6866-48S8CQ-P>
```

Run the following command to verify that port 25GE1/0/26 has been added to VLANs 55 and 66:

```
[~CE6866-48S8CQ-P]display vlan

The total number of vlans is : 3
--------------------------------------------------------------------------------
U: Up;         D: Down;         TG: Tagged;         UT: Untagged;
MP: Vlan-mapping;               ST: Vlan-stacking;
#: ProtocolTransparent-vlan;    *: Management-vlan;
MAC-LRN: MAC-address learning;  STAT: Statistic;
BC: Broadcast; MC: Multicast;   UC: Unknown-unicast;
FWD: Forward;  DSD: Discard;
--------------------------------------------------------------------------------

VID         Ports
--------------------------------------------------------------------------------
   1        UT:100GE1/0/1(D)   100GE1/0/2(D)   100GE1/0/3(D)   100GE1/0/4(D)
               100GE1/0/5(D)   100GE1/0/6(D)   100GE1/0/7(D)   100GE1/0/8(D)
               25GE1/0/1(D)    25GE1/0/2(D)    25GE1/0/3(D)    25GE1/0/4(D)
               25GE1/0/5(D)    25GE1/0/6(D)    25GE1/0/7(D)    25GE1/0/8(D)
               25GE1/0/9(D)    25GE1/0/10(D)   25GE1/0/11(D)   25GE1/0/12(D)
               25GE1/0/13(D)   25GE1/0/14(D)   25GE1/0/15(D)   25GE1/0/16(D)
               25GE1/0/17(D)   25GE1/0/18(D)   25GE1/0/19(D)   25GE1/0/20(D)
               25GE1/0/21(D)   25GE1/0/22(D)   25GE1/0/23(D)   25GE1/0/24(D)
               25GE1/0/25(D)   25GE1/0/26(U)   25GE1/0/27(D)   25GE1/0/28(D)
               25GE1/0/29(U)   25GE1/0/30(U)   25GE1/0/31(U)   25GE1/0/32(U)
               25GE1/0/33(U)   25GE1/0/34(D)   25GE1/0/35(D)   25GE1/0/36(D)
               25GE1/0/37(D)   25GE1/0/38(D)   25GE1/0/39(D)   25GE1/0/40(D)
               25GE1/0/41(D)   25GE1/0/42(D)   25GE1/0/43(D)   25GE1/0/44(D)
               25GE1/0/45(D)   25GE1/0/46(D)   25GE1/0/47(D)   25GE1/0/48(D)
 55         TG:25GE1/0/26(U)   25GE1/0/29(U)   25GE1/0/30(U)   25GE1/0/31(U)
               25GE1/0/32(U)   25GE1/0/33(U)
 66         TG:25GE1/0/26(U)   25GE1/0/29(U)   25GE1/0/30(U)   25GE1/0/31(U)
               25GE1/0/32(U)   25GE1/0/33(U)
```

NOTE

Repeat this operation on all switch ports to be used, including the ports connecting to the host, storage system, and other switches (if any).

----End

(Optional) PHB Mapping Configuration (Mandatory for Layer 3 Switching Networks)

If the switch network uses Layer 3 switching, add the following configuration to the uplink and downlink ports of the switch to perform PHB mapping for the DSCP values of outgoing packets. Layer 2 switching networks do not need this configuration.

```
[~CE6866-48S8CQ-P]interface 25GE1/0/26
[~CE6866-48S8CQ-P-25GE1/0/26]qos phb marking dscp enable
[*CE6866-48S8CQ-P-25GE1/0/26]commit
[~CE6866-48S8CQ-P-25GE1/0/26]quit
```

(Optional) DSCP Configuration

Step 1 Change the trust mode to DSCP.

```
<CE6866-48S8CQ-P>system-view
Enter system view, return user view with return command.
[~CE6866-48S8CQ-P]interface 25GE1/0/1
```

```
[~CE6866-48S8CQ-P-25GE1/0/1]trust dscp
[*CE6866-48S8CQ-P-25GE1/0/1]commit
[~CE6866-48S8CQ-P-25GE1/0/1]quit
[~CE6866-48S8CQ-P]
```

Step 2 Run the following command to verify that DSCP is enabled for the 25GE1/0/1 port.

```
[~CE6866-48S8CQ-P]interface 25GE1/0/1
[~CE6866-48S8CQ-P-25GE1/0/1]display this
#
interface 25GE1/0/1
 trust dscp
#
```

----End

NOTICE

The switch ports, host ports, and storage ports must use the same mode.

(Optional) iNOF Configuration (Mandatory If the OceanStor NOF Director Is Used)

To configure iNOF on a switch, enable LLDP on the switch and configure the iNOF reflector and client. You must create user-defined domains on the switch according to the service plan. For the detailed configuration methods, see the **iNOF configuration guide** in the switch documentation.

NOTICE

The iNOF system has a default domain in which members cannot be manually added. By default, the switch allows automatic adding of members to the iNOF default domain. If no user-defined domain is configured, hosts and storage ports with SNSD enabled are automatically added to the default domain. If iNOF domain isolation (hard-zone) is enabled on the switch, devices can access each other only when they have been added to the same iNOF domain.

4.2.1.2 Huawei CE6865 and CE8861

The methods for configuring Huawei CE6865 and CE8861 switches are the same. The following uses CE8861 as an example to describe how to configure PFC, AI-ECN, and VLANs.

PFC Configuration

Step 1 Enable PFC priority 3.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]dcb pfc
[~CE8861EI-dcb-pfc-default]priority 3
[*CE8861EI-dcb-pfc-default]commit
Committing....done.
[~CE8861EI-dcb-pfc-default]quit
```

Run the following command to verify that the PFC priority configuration has taken effect:

```
[~CE8861EI]display dcb pfc-profile default
--------------------------------------------------------------------------------
PFC-profile Name                    Priority
--------------------------------------------------------------------------------
default                          3
--------------------------------------------------------------------------------
 [~CE8861EI]quit
```

Step 2 Enable PFC on the switch ports.

The following uses port 100GE1/1/1 as an example. Run the **undo flow-control** command to disable the pause function (which conflicts with PFC), and run the **dcb pfc enable mode manual** command to enable PFC on the port.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]interface 100GE1/1/1
[~CE8861EI-100GE1/1/1]undo flow-control
[*CE8861EI-100GE1/1/1]dcb pfc enable mode manual
[*CE8861EI-100GE1/1/1]commit
[~CE8861EI-100GE1/1/1]quit
[~CE8861EI]quit
```

Run the following command to verify that PFC has been enabled on the 100GE1/1/1 port:

```
<CE8861EI>display dcb
M:Manual;    A:Auto
--------------------------------------------------------------------------------
Interface         PFC Name     PFC Status  ETS Name    ETS Status App-Profile
--------------------------------------------------------------------------------
100GE1/1/1        default      ENABLE(M)   -           -          -
--------------------------------------------------------------------------------
<CE8861EI>
```

NOTE

Repeat this operation on all switch ports to be used.

Step 3 Configure PFC deadlock detection.

Set the PFC deadlock detection interval and deadlock recovery interval to 100 ms. Configure PFC deadlock detection for PFC priority 3, and set the total detection period to 10 x 100 ms and total recovery period to 15 x 100 ms. Set the threshold for disabling PFC on a port to 20 deadlocks in 1 minute.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]dcb pfc deadlock-detect interval 100
[~CE8861EI]commit
[~CE8861EI]dcb pfc
[*CE8861EI-dcb-pfc-default]priority 3 deadlock-detect time 10 deadlock-recovery time 15
[*CE8861EI-dcb-pfc-default]priority 3 turn-off threshold 20
Info: PFC will be disabled if the threshold for the number of detected deadlocks is exceeded. If you need to
use PFC again, reconfigure PFC on an interface.
[*CE8861EI-dcb-pfc-default]commit
[~CE8861EI-dcb-pfc-default]quit
[~CE8861EI]quit
```

Verify the configuration.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]dcb pfc
[~CE8861EI-dcb-pfc-default]display this
#
```

```
dcb pfc
priority 3 deadlock-detect time 10 deadlock-recovery time 15
priority 3 turn-off threshold 20
#
return
[~CE8861EI-dcb-pfc-default]quit

[~CE8861EI]quit
```

----End

(Optional, Recommended) AI-ECN Configuration

You are advised to configure AI-ECN to dynamically and intelligently set the threshold based on the network traffic.

Step 1 Before the configuration, check whether the switch has the AI-ECN license.

If the value of **CE-LIC-LLETH** is **YES**, the switch has the AI-ECN license. Otherwise, you must load the license.

```
[~CE8861EI]display license
MainBoard:
 Active License    : flash:/LICCloudEngine8800_V200R019_20200616QV6G50.dat
 License state     : Demo
 Revoke ticket     : No ticket

 RD of Huawei Technologies Co., Ltd.

 Product name      : CloudEngine 8800
 Product version   : V200R019
 License Serial No : LIC20200616QV6G50
 Creator           : Huawei Technologies Co., Ltd.
 Created Time      : 2020-06-16 20:24:36
 SnS End Date      : --
 ------------------------------------------------------------
 Feature name      : NEMAL
 Authorize type    : demo
 Expired date      : 2020-08-31
 Trial days        : --
 ------------------------------------------------------------
 Feature name      : ACPHFDMA
 Authorize type    : demo
 Expired date      : 2020-08-31
 Trial days        : --
 ------------------------------------------------------------
 Feature name      : CELIC
 Authorize type    : demo
 Expired date      : 2020-08-31
 Trial days        : --

 Item name          Item type      Value      Description
 ------------------------------------------------------------
 CE-LIC-VXLAN       Function       YES        CE-LIC-VXLAN
 CE-LIC-FCF-PORT    Resource       512        CE-LIC-FCF-PORT
 CE-LIC-FCF-ALL     Function       YES        CE-LIC-FCF-ALL
 CE-LIC-NPV         Function       YES        CE-LIC-NPV
 CE-LIC-NS          Function       YES        CE-LIC-NS
 CE-LIC-NQA         Function       YES        CE-LIC-NQA
 CE-LIC-DHCP-S      Function       YES        CE-LIC-DHCP-S
 CE-LIC-VXLAN-M     Function       YES        CE-LIC-VXLAN-M
 CE-LIC-TLM         Function       YES        CE-LIC-TLM
 CE-LIC-LLETH       Function       YES        CE-LIC-LLETH
 CE-LIC-PTP         Function       YES        CE-LIC-PTP
 CE-LIC-FINHA       Function       YES        CE-LIC-FINHA
 CE-LIC-25G01       Resource       1          CE-LIC-25G01
 CE-LIC-40G01       Resource       1          CE-LIC-40G01
 CE-LIC-100G01      Resource       1          CE-LIC-100G01
```

Step 2 Configure the low-latency fabric.

This function takes effect after the switch is restarted.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]low-latency fabric
Info: Please save the configuration and reboot the system to enable the operation.
[*CE8861EI-low-latency-fabric]quit
[*CE8861EI]quit
Warning: Uncommitted configurations found. Are you sure to commit them before exiting? [Y(yes)/N(no)/
C(cancel)]:y
<CE8861EI>save
Warning: The current configuration will be written to the device. Continue? [Y/N]:y
Now saving the current configuration to the slot 1 ..
Info: Save the configuration successfully.
<CE8861EI>reboot
slot 1:
Next startup system software: flash:/CE8861EI-V200R019C10SPC800.cc
Next startup saved-configuration file: flash:/hs.cfg
Next startup paf file: default
Next startup patch package: NULL
Warning: The system will reboot. Continue? [Y/N]:y
```

Step 3 After the restart, the switch automatically enables the intelligent lossless network functions. Disable dynamic ECN and then enable AI-ECN.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]low-latency fabric
[~CE8861EI-low-latency-fabric]undo qos dynamic-ecn-threshold enable
[*CE8861EI-low-latency-fabric]quit
[*CE8861EI]ai-service
[*CE8861EI-ai-service]ai-ecn
[*CE8861EI-ai-service-ai-ecn]ai-ecn enable
Info: This function takes effect only for the queue with the highest priority configured in the pfc default
profile.
[*CE8861EI-ai-service-ai-ecn]commit
[~CE8861EI-ai-service-ai-ecn]quit
[~CE8861EI-ai-service]quit
[~CE8861EI]quit
<CE8861EI>
```

Step 4 Run the following command to verify that AI-ECN has been activated:

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]ai-se
[~CE8861EI]ai-service
[~CE8861EI-ai-service]dis ai-ecn calculated state
*: Indicates the queue where AI ECN takes effect.
AI-ECN State: enabled
--------------------------------------------------------------------
Interface       Queue   Low-Threshold   High-Threshold   Probability
                         (Byte)          (Byte)           (%)
--------------------------------------------------------------------
100GE1/1/1       0           0              0           0
                 1           0              0           0
                 2           0              0           0
                *3        5120         512000           1
                 4           0              0           0
                 5           0              0           0
                 6           0              0           0
                 7           0              0           0
--------------------------------------------------------------------
100GE1/1/2       0           0              0           0
                 1           0              0           0
                 2           0              0           0
                *3        5120         512000           1
                 4           0              0           0
                 5           0              0           0
```

```
              6            0             0           0
              7            0             0           0
-------------------------------------------------------------------
```

----End

VLAN Configuration

Step 1 Configure global VLANs. The following example uses VLAN 55 and VLAN 66.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
 [~CE8861EI]vlan batch 55 66
Info: Operating, please wait for a moment......done.
[*CE8861EI]commit
[~CE8861EI]quit
```

Step 2 Configure VLANs for ports.

The following uses port 100GE1/1/1 as an example. Configure the port to work in trunk mode and then add the port to VLANs 55 and 66. Then set the port to an edge port.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]inte
[~CE8861EI]interface 100GE1/1/1
[~CE8861EI-100GE1/1/1]port link-type trunk
[*CE8861EI-100GE1/1/1]port trunk allow-pass vlan 55 66
[*CE8861EI-100GE1/1/1]stp edged-port enable
[*CE8861EI-100GE1/1/1]commit
[~CE8861EI-100GE1/1/1]quit
[~CE8861EI]quit
```

Run the following command to verify that port 100GE1/1/1 has been added to VLANs 55 and 66:

```
<CE8861EI>display vlan
The total number of vlans is : 3
--------------------------------------------------------------------------------
U: Up;         D: Down;         TG: Tagged;         UT: Untagged;
MP: Vlan-mapping;               ST: Vlan-stacking;
#: ProtocolTransparent-vlan;    *: Management-vlan;
MAC-LRN: MAC-address learning;  STAT: Statistic;
BC: Broadcast; MC: Multicast;   UC: Unknown-unicast;
FWD: Forward;  DSD: Discard;
--------------------------------------------------------------------------------

VID         Ports
--------------------------------------------------------------------------------
  1        UT:100GE1/1/1(U)   100GE1/1/2(U)   100GE1/1/3(D)   100GE1/1/4(D)
              100GE1/1/5(D)   100GE1/1/6(D)   100GE1/1/7(D)   100GE1/1/8(D)
              100GE1/2/1(D)   100GE1/2/2(D)   100GE1/2/3(D)   100GE1/2/4(D)
              100GE1/2/5(D)   100GE1/2/6(D)   100GE1/2/7(D)   100GE1/2/8(D)
              100GE1/3/1(D)   100GE1/3/2(D)   100GE1/3/3(D)   100GE1/3/4(D)
              100GE1/3/5(D)   100GE1/3/6(D)   100GE1/3/7(D)   100GE1/3/8(D)
              100GE1/4/1(D)   100GE1/4/2(D)   100GE1/4/3(D)   100GE1/4/4(D)
              100GE1/4/5(D)   100GE1/4/6(D)   100GE1/4/7(D)   100GE1/4/8(D)
 55        TG:100GE1/1/1(U)   100GE1/1/2(U)
 66        TG:100GE1/1/1(U)   100GE1/1/2(U)
```

NOTE

Repeat this operation on all switch ports to be used, including the ports connecting to the host, storage system, and other switches (if any).

----End

(Optional) PHB Mapping Configuration (Mandatory for Layer 3 Switching Networks)

If the switch network uses Layer 3 switching, add the following configuration to the uplink and downlink ports of the switch to perform PHB mapping for the DSCP values of outgoing packets. Layer 2 switching networks do not need this configuration.

```
[~CE8861EI]interface 100GE1/1/1
[~CE8861EI-100GE1/1/1]qos phb marking dscp enable
[*CE8861EI-100GE1/1/1]commit
[~CE8861EI-100GE1/1/1]quit
```

(Optional) DSCP Configuration

Step 1 Change the trust mode to DSCP.

```
<CE8861EI>system-view
Enter system view, return user view with return command.
[~CE8861EI]interface 100GE1/1/1
[~CE8861EI-100GE1/1/1]trust dscp
[*CE8861EI-100GE1/1/1]commit
[~CE8861EI-100GE1/1/1]quit
[~CE8861EI]
```

Step 2 Run the following command to verify that DSCP is enabled for the 25GE1/0/1 port.

```
[~CE8861EI]interface 100GE1/1/1
[~CE8861EI-100GE1/1/1]display this
#
interface 100GE1/1/1
 trust dscp
#
```

----End

NOTICE

The switch ports, host ports, and storage ports must use the same mode.

4.2.2 Storage System Configuration

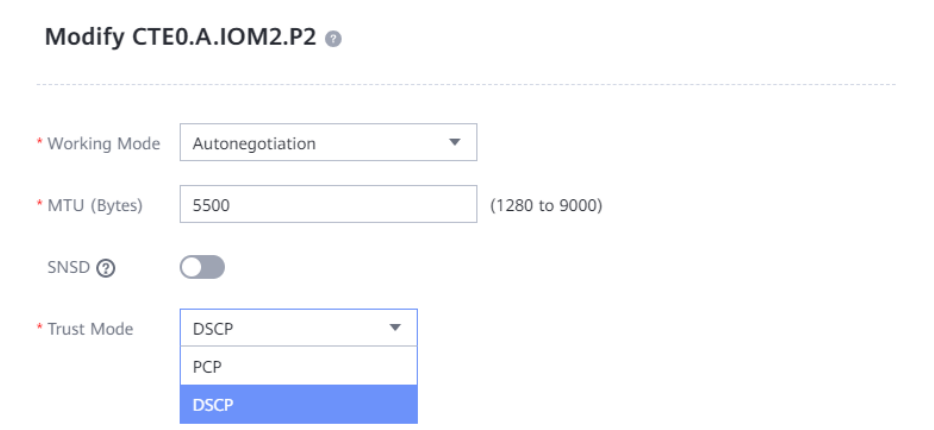
Create storage pools, LUNs/LUN groups, and hosts/host groups on the storage system and map the LUNs/LUN groups to the hosts/host groups according to your service requirements. For details, see the *Basic Storage Service Configuration Guide for Block*.

Configure the RoCE ports on the storage system as follows:

Step 1 (Optional) Change the trust mode of storage ports.

Log in to DeviceManager of the storage system and choose **Services** > **Network** > **RoCE Network**. Click the desired physical port to modify it.

Figure 4-3 Changing the trust mode of a storage RoCE port



Modify CTE0.A.IOM2.P2

- * Working Mode: Autonegotiation
- * MTU (Bytes): 5500 (1280 to 9000)
- SNSD: off
- * Trust Mode: DSCP

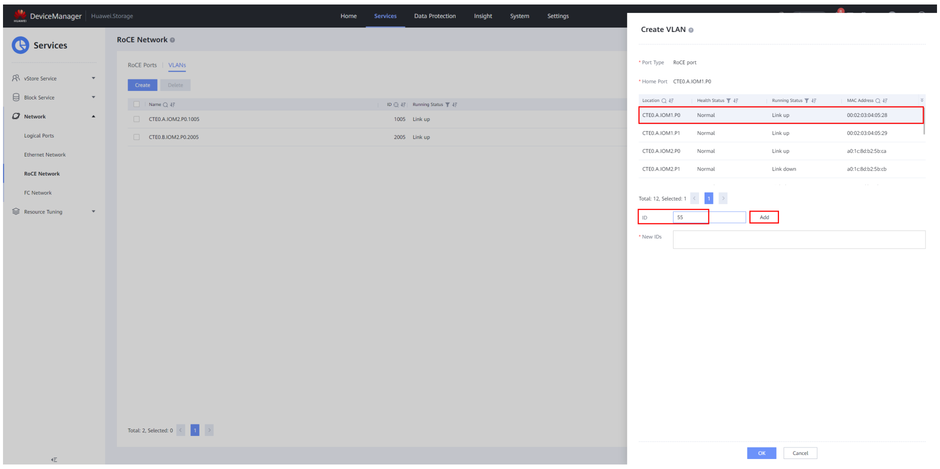
NOTICE

- For details about the interface modules that support DSCP, see "Managing the RoCE Network" in the *Basic Storage Service Configuration Guide for Block*.
- The default trust mode is **PCP**. If you want to create a logical port directly on the RoCE port, select **DSCP**.
- Only 6.1.5 and later versions can set the trust mode of a port. The switch ports, host ports, and storage ports must use the same mode.
- In PCP mode, logical ports can be configured only on VLAN ports. In DSCP mode, logical ports can be configured on RoCE and VLAN ports.

Step 2 (Optional for DSCP but mandatory for PCP) Configure VLANs.

Choose **Services** > **Network** > **RoCE Network**. On the **VLANs** tab, click **Create**. On the **Create VLAN** page, select the desired RoCE port, enter the planned VLAN ID in the **ID** text box, click **Add**, and then click **OK**.

Figure 4-4 Creating VLANs for storage RoCE ports



Name	ID	Running Status
CTE0.A.IOM2.P0.1005	1005	Link up
CTE0.B.IOM2.P0.2005	2005	Link up

Total: 2, Selected: 0

Create VLAN

- * Port Type: RoCE port
- * Home Port: CTE0.A.IOM1.P0

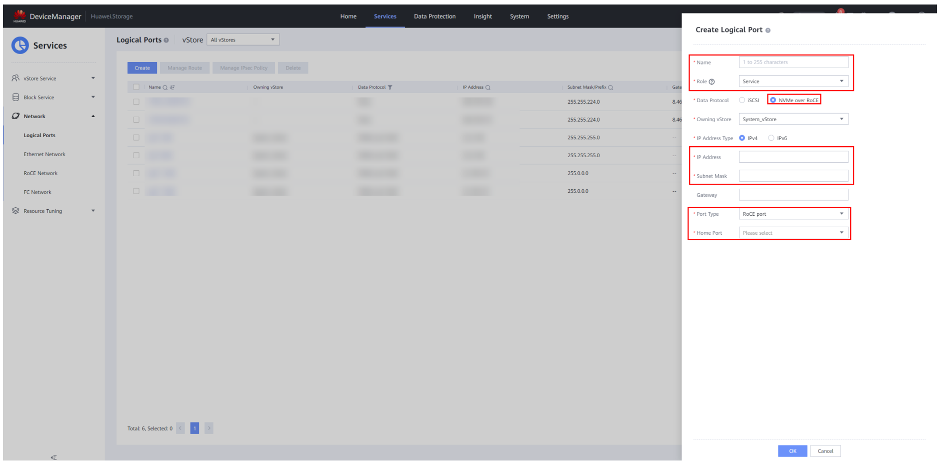
Location	Health Status	Running Status	MAC Address
CTE0.A.IOM1.P0	Normal	Link up	00:02:03:04:05:28
CTE0.A.IOM1.P1	Normal	Link up	00:02:03:04:05:29
CTE0.A.IOM2.P0	Normal	Link up	a0:1c:8d:b2:5b:ca
CTE0.A.IOM2.P1	Normal	Link down	a0:1c:8d:b2:5b:cb

Total: 12, Selected: 1

- ID: 55 Add
- * New IDs:

Step 3 Choose **Services** > **Network** > **Logical Ports**. Click **Create**. On the **Create Logical Port** page, set the logical port parameters. Set **Role** to **Service**, **Data Protocol** to **NVMe over RoCE**, **Port Type** to **RoCE port** or **VLAN**, and **Home Port** to the physical port to be configured. Set other parameters based on the network plan.

Figure 4-5 Creating a logical port



Create Logical Port

* Name: 1 to 255 characters

* Role: Service

* Data Protocol: ○ iSCSI ● NVMe over RoCE

* Owning vStore: System_vStore

* IP Address Type: ● IPv4 ○ IPv6

* IP Address:

* Subnet Mask:

Gateway:

* Port Type: RoCE port

* Home Port: Please select

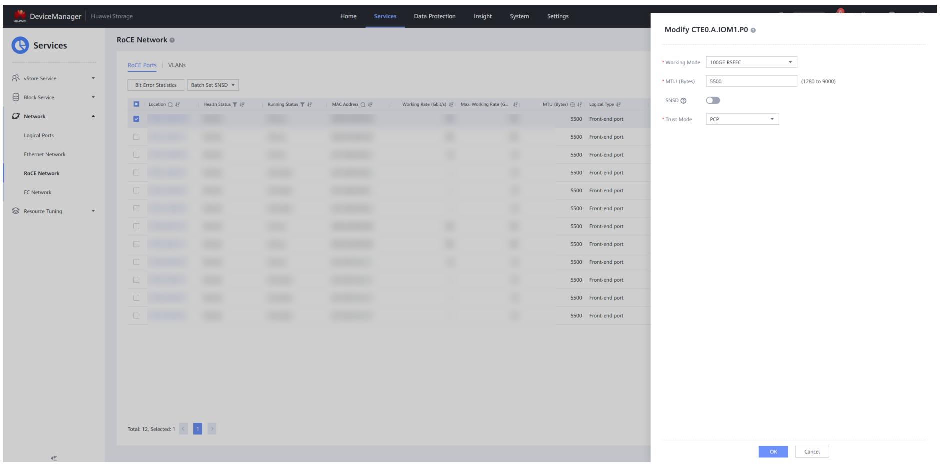
NOTICE

RoCE port is available only when DSCP is configured for the port.

Step 4 (Optional) Modify the MTU of the storage port.

1. For better performance, you are advised to change the MTU of the storage RoCE port and VLAN to 4200 or higher (5500 is recommended, which is the default value for 6.1.2 and later versions of the storage system). When NVMe over RoCE connections are set up on the host, set the MTU of the physical ports and VLAN ports on the host to 4200 or higher (5500 is recommended).
2. To use OceanStor NOF Director, enable SNSD for the port.

Figure 4-6 Modifying the MTU and SNSD of the storage RoCE port



Modify CTE0.A.IOM1.P0

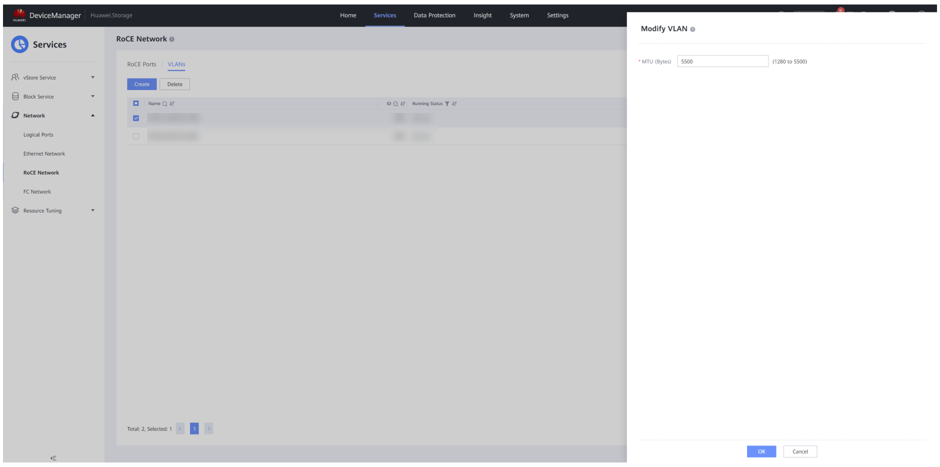
* Working Mode: 100GE RSFEC

* MTU (Bytes): 5500 (1280 to 9000)

SNSD

* Trust Mode: PCP

Figure 4-7 Modifying the MTU of the storage VLAN



Modify VLAN

* MTU (Bytes) 5500 (1280 to 5500)

Step 5 Repeat **Step 1** to **Step 4** to configure other storage RoCE ports based on the network plan.

----End

4.2.3 Identifying NICs of the Host

Before connecting a host to a storage system using NVMe over RoCE, make sure that the host's NICs have been identified and are working properly.

The following describes how to query the attributes of Mellanox NICs, including their driver versions, firmware versions, port rates, and port connection status. To query other attributes or other vendors' NICs, it is recommended that you use the management software provided by the respective NIC vendor. See the operation guide of the corresponding NIC management software for detailed operations.

4.2.3.1 Mellanox NIC

For Mellanox NICs, run the following commands to query the NIC bus information and then query the NIC model using the bus information:

```
[root@localhost ~]# lspci |grep Mellanox
81:00.0 Ethernet controller: Mellanox Technologies MT28800 Family [ConnectX-5 Ex]
81:00.1 Ethernet controller: Mellanox Technologies MT28800 Family [ConnectX-5 Ex]
[root@localhost ~]# lspci -vv -s 81:00.0 | grep "Part number"
                [PN] Part number: MCX516A-CDAT
[root@localhost ~]# lspci -vv -s 81:00.1 | grep "Part number"
                [PN] Part number: MCX516A-CDAT
```

Check the NIC driver version and firmware version as follows (uses the **enp129s0f0** port as an example):

```
[root@localhost ~]# ethtool -i enp129s0f0
driver: mlx5_core
version: 5.0-0
firmware-version: 16.27.1016 (MT_0000000013)
expansion-rom-version:
bus-info: 0000:81:00.0
supports-statistics: yes
supports-test: yes
supports-eeprom-access: no
supports-register-dump: no
supports-priv-flags: yes
```

The preceding information indicates that the Mellanox HBA driver version is 5.0-0 and the firmware version is 16.27.1016.

Check the port rate and connection status of the NIC as follows (uses the **enp129s0f0** port as an example):

```
[root@localhost ~]# ethtool enp129s0f0
Settings for enp129s0f0:
        Supported ports: [ FIBRE ]
        Supported link modes:   1000baseKX/Full
                                10000baseKR/Full
                                40000baseKR4/Full
                                40000baseCR4/Full
                                40000baseSR4/Full
                                40000baseLR4/Full
                                25000baseCR/Full
                                25000baseKR/Full
                                25000baseSR/Full
                                50000baseCR2/Full
                                50000baseKR2/Full
                                100000baseKR4/Full
                                100000baseSR4/Full
                                100000baseCR4/Full
                                100000baseLR4_ER4/Full
        Supported pause frame use: Symmetric
        Supports auto-negotiation: Yes
        Supported FEC modes: Not reported
        Advertised link modes:  1000baseKX/Full
                                10000baseKR/Full
                                40000baseKR4/Full
                                40000baseCR4/Full
                                40000baseSR4/Full
                                40000baseLR4/Full
                                25000baseCR/Full
                                25000baseKR/Full
                                25000baseSR/Full
                                50000baseCR2/Full
                                50000baseKR2/Full
                                100000baseKR4/Full
                                100000baseSR4/Full
                                100000baseCR4/Full
                                100000baseLR4_ER4/Full
        Advertised pause frame use: Symmetric
        Advertised auto-negotiation: Yes
        Advertised FEC modes: Not reported
        Speed: 100000Mb/s
        Duplex: Full
        Port: FIBRE
        PHYAD: 0
        Transceiver: internal
        Auto-negotiation: on
        Supports Wake-on: d
        Wake-on: d
        Current message level: 0x00000004 (4)
                               link
        Link detected: yes
```

The preceding information indicates that the port is properly connected and the rate is 100 Gbit/s.

4.3 Installing NVMe Software Packages

To run NVMe commands on Linux hosts, you must install the NVMe-CLI tool.

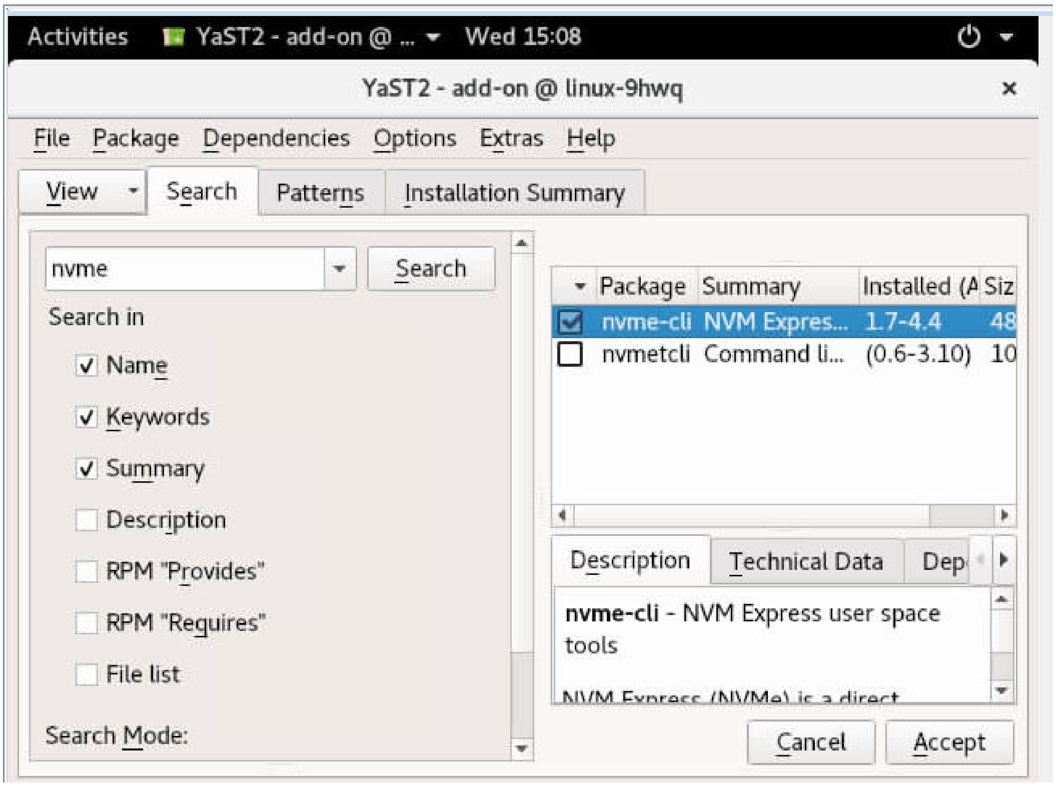
Run the following command to check whether the tool has been installed:

```
[root@localhost ~]# rpm -qa | grep nvme
nvme-cli-1.6-2.el8.x86_64
```

If no result is returned, the tool is not installed. You can use **yast** or **yum** to install the tool. Before the installation, you must configure the source by following instructions in the configuration guide of the specific OS.

In SUSE, use the **yast** command to install the tool. The following is an example.

Figure 4-8 Installing NVMe-CLI on SUSE



nvme

Search in

☑ Name

☑ Keywords

☑ Summary

□ Description

□ RPM "Provides"

□ RPM "Requires"

□ File list

Search Mode:

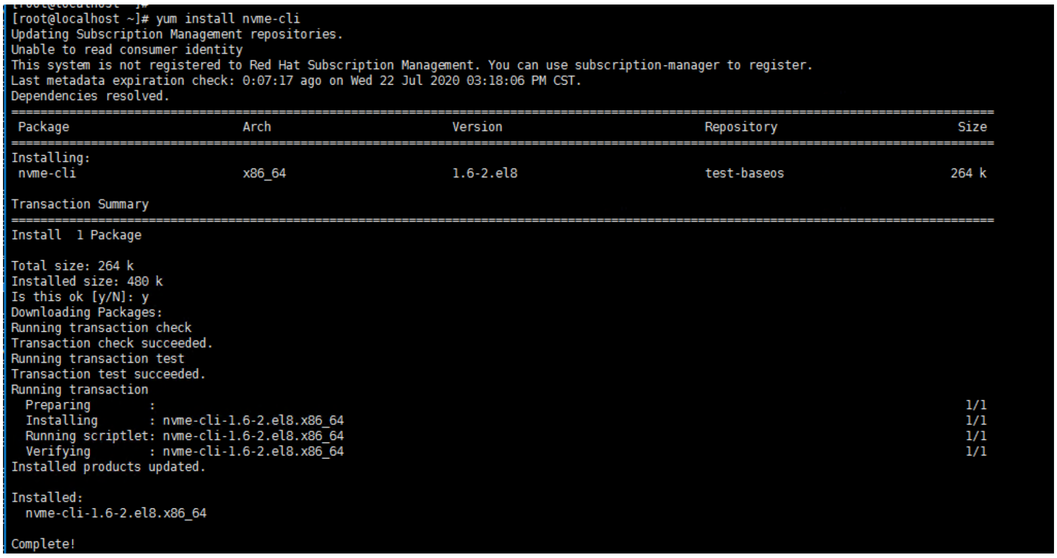
	Package	Summary	Installed (A	Siz
☑	nvme-cli	NVM Expres...	1.7-4.4	48
□	nvmetcli	Command li...	(0.6-3.10)	10

nvme-cli - NVM Express user space tools

NVM Express (NVMe) is a direct

In Red Hat, use the **yum** command to install the tool (you must connect to the Red Hat official website or configure a local yum source). The following is an example.

Figure 4-9 Installing NVMe-CLI on Red Hat



```
[root@localhost ~]# yum install nvme-cli
Updating Subscription Management repositories.
Unable to read consumer identity
This system is not registered to Red Hat Subscription Management. You can use subscription-manager to register.
Last metadata expiration check: 0:07:17 ago on Wed 22 Jul 2020 03:18:06 PM CST.
Dependencies resolved.
================================================================================================================================
 Package                      Arch                         Version                            Repository                   Size
================================================================================================================================
Installing:
 nvme-cli                     x86_64                       1.6-2.el8                          test-baseos                 264 k

Transaction Summary
================================================================================================================================
Install  1 Package

Total size: 264 k
Installed size: 480 k
Is this ok [y/N]: y
Downloading Packages:
Running transaction check
Transaction check succeeded.
Running transaction test
Transaction test succeeded.
Running transaction
  Preparing        :                                                                                                        1/1
  Installing       : nvme-cli-1.6-2.el8.x86_64                                                                              1/1
  Running scriptlet: nvme-cli-1.6-2.el8.x86_64                                                                              1/1
  Verifying        : nvme-cli-1.6-2.el8.x86_64                                                                              1/1
Installed products updated.

Installed:
  nvme-cli-1.6-2.el8.x86_64

Complete!
```

NOTE

For details on how to use the **yast** and **yum** commands, refer to the documents of the OS.

4.4 Installing OS Patches and Upgrading the HBA or NIC Driver/Firmware

In specific combinations of OSs and HBAs or NICs, there is a possibility that links cannot be established or services cannot be provisioned when NVMe is used to connect to storage systems. These problems are usually caused by defects in the OS kernel or HBA/NIC driver/firmware. You can install the kernel patches or upgrade the HBA/NIC driver/firmware to solve the problems.

Refer to the compatibility list for the OS patches that should be installed or HBA/NIC drivers/firmware that should be upgraded. For details on how to query the compatibility list, see **2.3 Interoperability Query**.

NOTICE

- If you plan to install or have installed the OceanStor NOF INI, use the default Mellanox NIC driver integrated in the OS. Do not upgrade the Mellanox NIC driver.
- For details on how to install the OS patches and upgrade the HBA/NIC driver/firmware, see the official documents of the OS or HBA/NIC vendor.

4.5 Installing and Configuring OceanStor NOF Enabler

NVMe has just started to develop. Its stability needs further improvement and the driver vulnerabilities must be gradually fixed. To address these issues, Huawei develops the OceanStor NOF Enabler software to fix the NVMe protocol vulnerabilities in operating systems and improve the robustness of the NVMe driver. In addition, the OceanStor NOF Enabler can automatically establish host connections and quickly identify faults and perform switchovers.

The OceanStor NOF Enabler software package contains the NOF INI and NOF Director.

- The OceanStor NOF INI is a plug-in to improve the robustness of the NVMe driver. It fixes the vulnerabilities discovered by the kernel community and those detected in Huawei's tests. It provides the same functions as the native NVMe driver, but is more reliable than the native NVMe driver.
- The OceanStor NOF Director is a plug-in to improve link reliability. After being installed in the host OS, it simplifies service deployment on the conventional Ethernet (lossless RoCE network) and can quickly detect faults and perform service switchover.

Currently, the OceanStor NOF Enabler supports only NVMe over RoCE and does not support FC-NVMe. You are advised to install the NVMe NOF Enabler for NVMe over RoCE connections.

For details about the functions, installation and configuration methods, and restrictions of the software, see the *OceanStor NOF Enabler x.x.x Usage Guide*.

NOTICE

Currently, the OceanStor NOF Enabler can be used only in a few operating systems, and the NOF Director must be used with specific Huawei switches. For details, refer to **https://support.huawei.com/enterprise/en/centralized-storage/oceanstor-nof-enabler-pid-251608654**.

5 Configuring Connectivity

This chapter describes how to configure connectivity between storage systems and hosts. The configuration methods for FC-NVMe and NVMe over RoCE connections are different. Follow the correct configuration guide based on your network type.

5.1 Establishing FC-NVMe Connections

This section describes how to establish FC-NVMe connections between application servers and storage systems.

5.1.1 Host Configuration

Query the HBA WWNs. The following is an example.

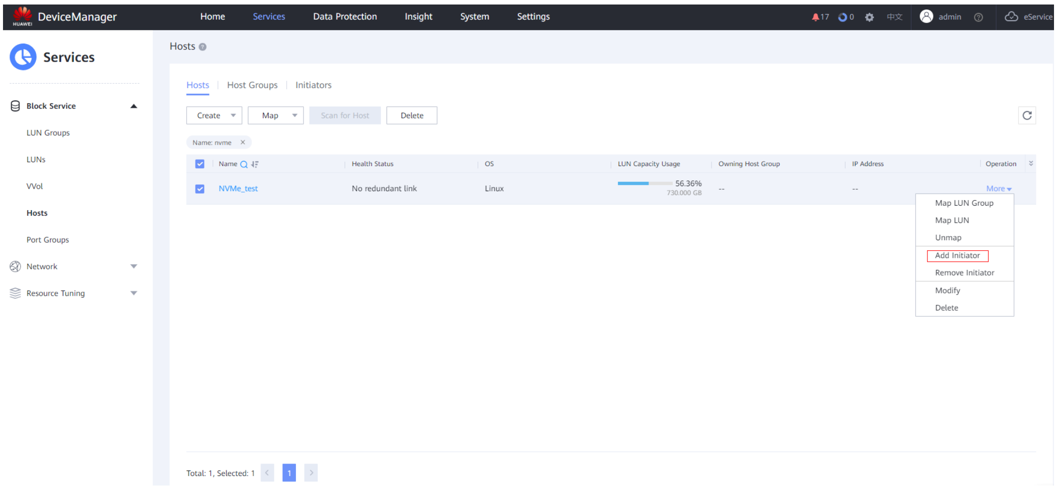
```
[root@localhost ~]# cat /sys/class/fc_host/host*/port_name
0x100000109b32a3c0
0x100000109b32a3c1
```

5.1.2 Storage System Configuration

This section details how to add initiators to the hosts on the storage system. For details on how to create hosts and mappings, see the *Basic Storage Service Configuration Guide* corresponding to your storage system.

Step 1 After configuring zones on the switches, log in to DeviceManager of the storage system and choose **Services** > **Hosts**. Select the desired host, click **More** on the right, and choose **Add Initiator**.

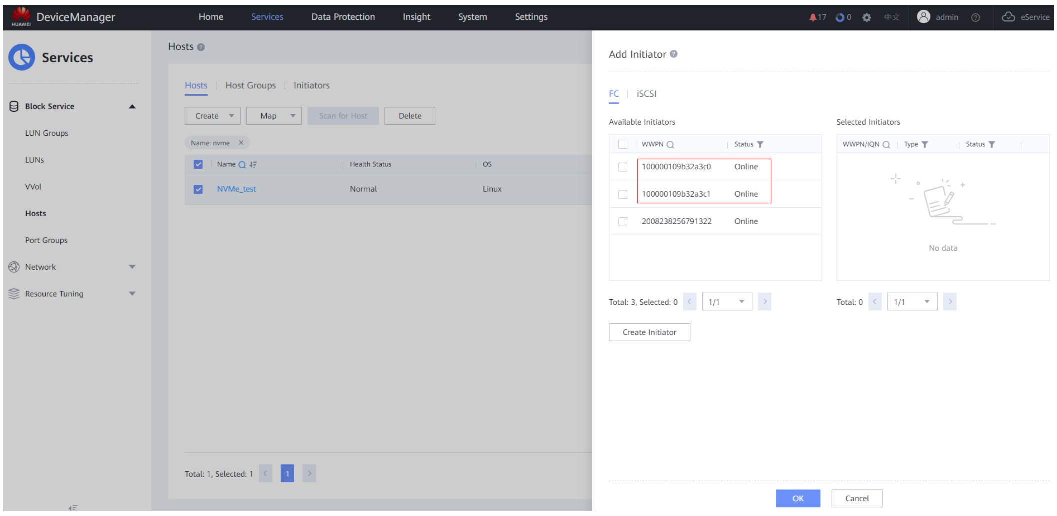
Figure 5-1 Adding an initiator



Name	Health Status	OS	LUN Capacity Usage	Owning Host Group	IP Address	Operation
NVMe_test	No redundant link	Linux	56.36% 730.000 GB	--	--	More

Step 2 Check whether the host initiators can be discovered.

Figure 5-2 Querying initiators

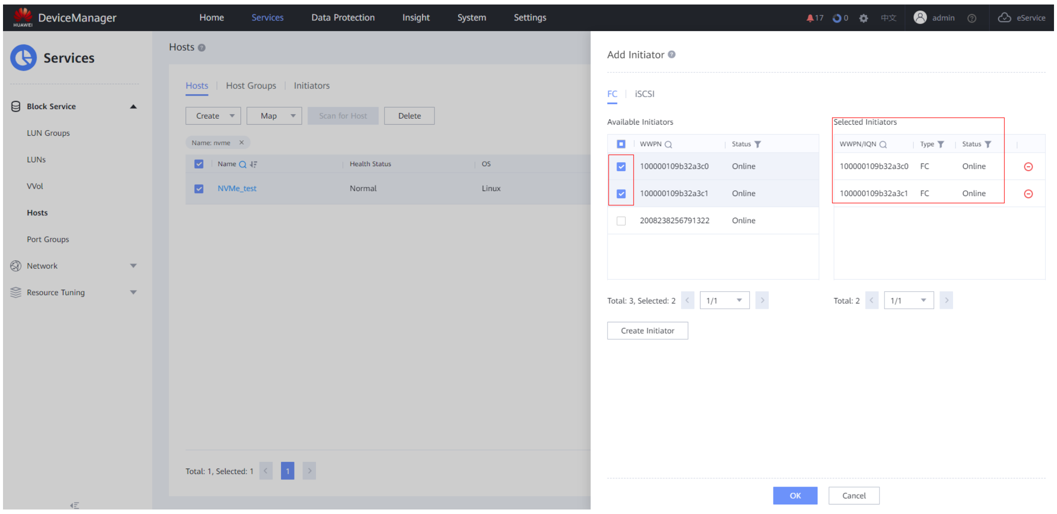


Name	Health Status	OS
NVMe_test	Normal	Linux

WWPN	Status
100000109b32a3c0	Online
100000109b32a3c1	Online
2008238256791322	Online

Step 3 Select the desired initiators and click **OK**.

Figure 5-3 Assigning initiators



Available Initiators

☑	WWPN	Status
☑	100000109b32a3c0	Online
☑	100000109b32a3c1	Online
□	2008238256791322	Online

Total: 3, Selected: 2 1/1

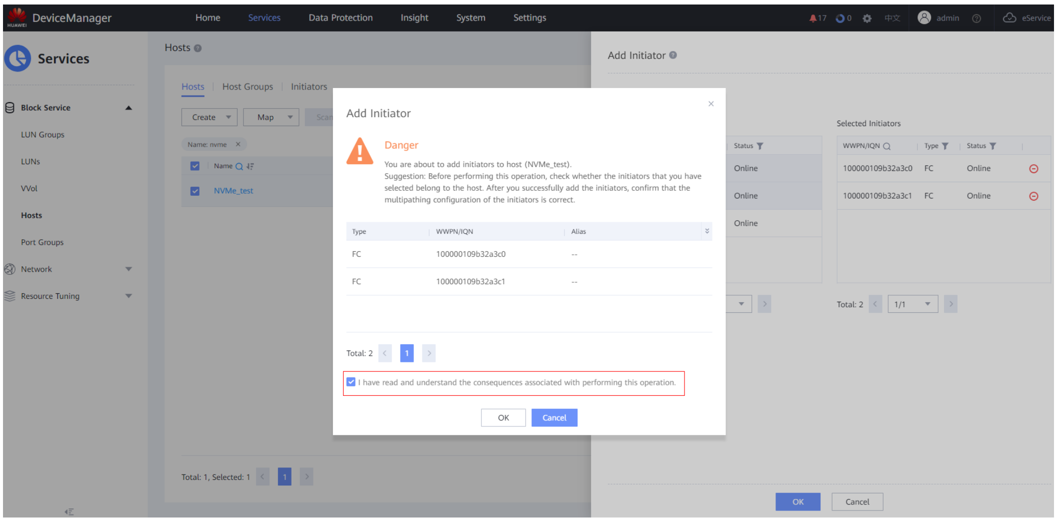
Selected Initiators

WWPN/IQN	Type	Status
100000109b32a3c0	FC	Online
100000109b32a3c1	FC	Online

Total: 2 1/1

Step 4 Confirm the information and click **OK**.

Figure 5-4 Confirming the operation



Add Initiator

Danger

You are about to add initiators to host (NVMe_test).
Suggestion: Before performing this operation, check whether the initiators that you have selected belong to the host. After you successfully add the initiators, confirm that the multipathing configuration of the initiators is correct.

Type	WWPN/IQN	Alias
FC	100000109b32a3c0	--
FC	100000109b32a3c1	--

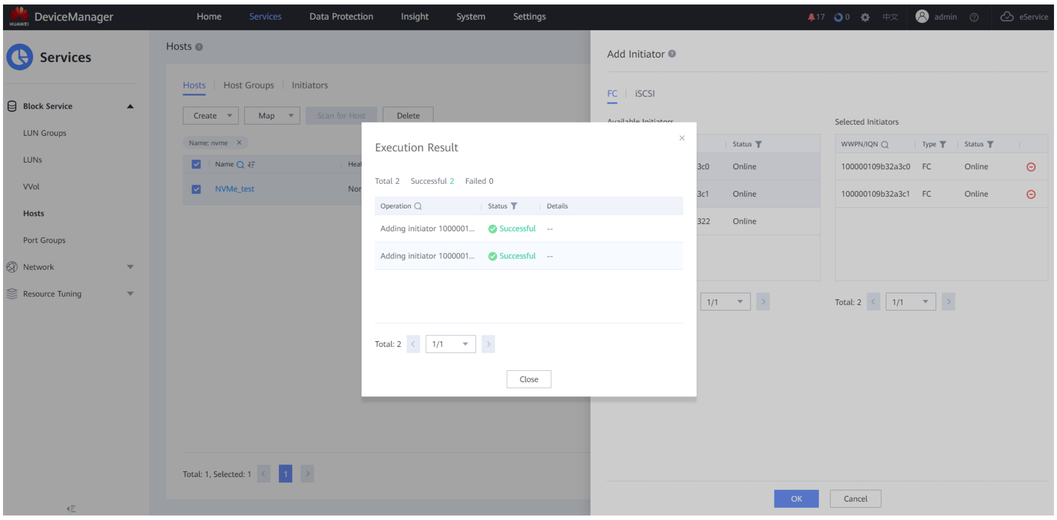
Total: 2 1

☑ I have read and understand the consequences associated with performing this operation.

OK Cancel

Step 5 After the system prompts that the initiators have been added successfully, click **Close**.

Figure 5-5 Initiators added successfully



Execution Result

Total 2 Successful 2 Failed 0

Operation	Status	Details
Adding initiator 1000001...	✓ Successful	--
Adding initiator 1000001...	✓ Successful	--

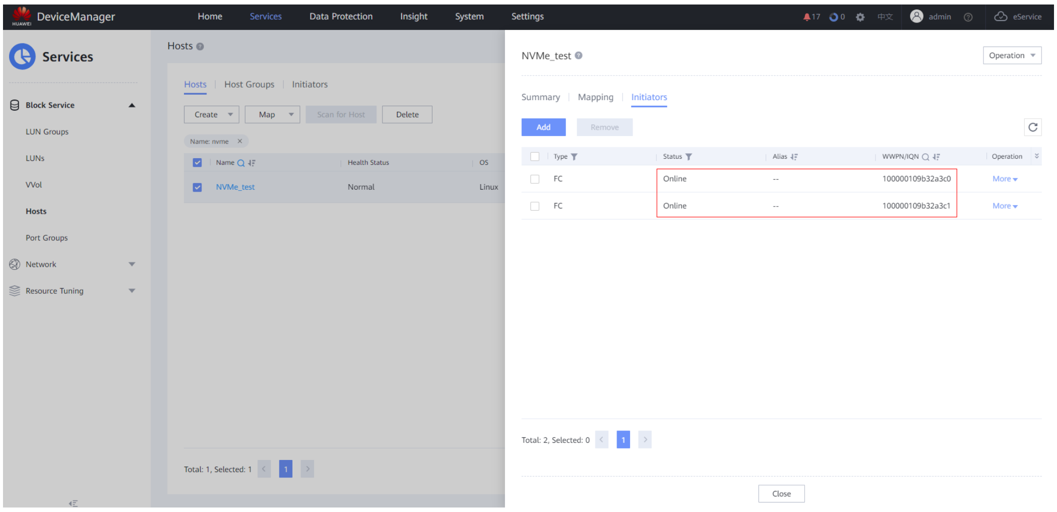
Total: 2 1/1

WWPN/IQN	Type	Status
100000109b32a3c0	FC	Online
100000109b32a3c1	FC	Online

Total: 2 1/1

Step 6 Click the host name and check the initiators. Ensure that the **Status** of the initiators is **Online**.

Figure 5-6 Checking the initiator status



Name	Health Status	OS
NVMe_test	Normal	Linux

Total: 1, Selected: 1

NVMe_test

Type	Status	Alias	WWPN/IQN	Operation
FC	Online	--	100000109b32a3c0	More
FC	Online	--	100000109b32a3c1	More

Total: 2, Selected: 0

----End

5.2 Establishing NVMe over RoCE Connections

This section describes how to establish NVMe over RoCE connections between application servers and storage systems.

5.2.1 Host Configuration

5.2.1.1 Configuring Network Information

On the host, configure the IP addresses, VLANs, and MTUs for the network ports to establish NVMe over RoCE connections with storage ports. The following provides examples.

5.2.1.1.1 Red Hat, CentOS, and Kylin Configuration

Step 1 Modify the configuration files for the ports on the RoCE NICs. The following uses the **enp129s0f0** NIC as an example.

```
[root@localhost ~]# cat /etc/sysconfig/network-scripts/ifcfg-enp129s0f0
TYPE=Ethernet
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=none
DEFROUTE=yes
IPV6INIT=no
IPV6_AUTOCONF=no
NAME=enp129s0f0
UUID=6974c94d-9cb1-4316-b0c7-865bea8f994a
DEVICE=enp129s0f0
ONBOOT=yes
MTU=5500
```

NOTE

- If automatic IP address configuration is required on the physical network port, ensure that DHCP has been correctly configured.
- The default MTU is 1500. For better performance, you are advised to set the MTU to 5500. Ensure that the MTUs of the host physical network port, host VLAN port, storage RoCE port, and storage VLAN are the same.
- If IPv4 addresses are used, IPv6 needs to be disabled by setting **IPV6INIT** to **no**. If IPv6 addresses are used, set **IPV6_AUTOCONF** to **no**.

Step 2 Because RoCE PFC is based on VLANs, you must create VLAN ports for the physical network ports on the host.Currently, Huawei storage supports manual configuration of RoCE PFC based on PCP and DSCP. In PCP mode, you must create VLAN ports for the physical network ports on the host. In DSCP mode, you can determine whether to create VLAN ports for the physical network ports on the host based on service requirements.

Network port **enp129s0f0** is used as an example. Create a VLAN port **enp129s0f0.55** (55 is the VLAN ID, which must be the same as the VLAN ID of the interconnected port on the storage system). The following is an example of the configuration file content:

```
[root@localhost ~]# cat /etc/sysconfig/network-scripts/ifcfg-enp129s0f0.55
BOOTPROTO=static
IPV4_FAILURE_FATAL=no
NAME=enp129s0f0.55
DEVICE=enp129s0f0.55
ONBOOT=yes
MTU=5500
VLAN=yes
IPADDR=192.168.5.5
NETMASK=255.255.255.0
VLAN_EGRESS_PRIORITY_MAP=0:3,1:3,2:3,3:3,4:3,5:3,6:3,7:3
```

NOTICE

- For all VLAN ports, the VLAN egress priority must be mapped to priority 3. Add **VLAN_EGRESS_PRIORITY_MAP=0:3,1:3,2:3,3:3,4:3,5:3,6:3,7:3** to the preceding configuration file.
- Repeat this operation on all host service ports to be used.
- The default MTU is 1500. For better performance, you are advised to set the MTU to 5500. Ensure that the MTUs of the host physical network port, host VLAN port, storage RoCE port, and storage VLAN are the same.
- If you use the **nmcli** command to create the VLAN NIC configuration file, set **IPV6INIT=no** and **IPV6_AUTOCONF=no**.

Step 3 Run the following command for the physical network port and VLAN port settings to take effect. Run the command on each physical network port and VLAN port you have modified.

```
[root@localhost /]# ifdown enp129s0f0
Connection 'enp129s0f0' successfully deactivated (D-Bus active path: /org/freedesktop/NetworkManager/
ActiveConnection/7)
[root@localhost /]# ifup enp129s0f0
Connection successfully activated (D-Bus active path: /org/freedesktop/NetworkManager/
ActiveConnection/8)
[root@localhost /]# ifdown enp129s0f0.55
Connection 'enp129s0f0.55' successfully deactivated (D-Bus active path: /org/freedesktop/NetworkManager/
ActiveConnection/6)
[root@localhost /]# ifup enp129s0f0.55
Connection successfully activated (D-Bus active path: /org/freedesktop/NetworkManager/
ActiveConnection/9)
```

NOTICE

Deactivating a network port may cause network exceptions. Ensure that no service is running before performing this operation.

Step 4 Check the IP address of the VLAN port to confirm that the configuration has taken effect.

```
[root@localhost ~]# ifconfig  enp129s0f0.55
enp129s0f0.55: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 5500
        inet 192.168.5.5  netmask 255.255.255.0  broadcast 192.168.5.255
        inet6 fe80::268a:7ff:fead:7a22  prefixlen 64  scopeid 0x20<link>
        ether 24:8a:07:ad:7a:22  txqueuelen 1000  (Ethernet)
        RX packets 0  bytes 0 (0.0 B)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 54  bytes 7628 (7.4 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

NOTICE

If the same VLAN is configured for different physical network ports on the host and IP addresses in the same network segment are configured for these VLAN ports, add the following content to the **/etc/sysctl.conf** file:

```
net.ipv4.conf.all.arp_ignore=1
```

Then run the **sysctl -w net.ipv4.conf.all.arp_ignore=1** command for the configuration to take effect.

Step 5 Verify that the VLAN egress priority has taken effect.

```
[root@localhost ~]# cat /proc/net/vlan/ens129s0f0.55
ens129s0f0.55  VID: 55 REORDER_HDR: 1  dev->priv_flags: 1021
         total frames received          132
          total bytes received         6582
      Broadcast/Multicast Rcvd            0

      total frames transmitted          102
       total bytes transmitted         6826
Device: ens129s0f0
INGRESS priority mappings: 0:0  1:0  2:0  3:0  4:0  5:0  6:0 7:0
 EGRESS priority mappings: 0:3 1:3 2:3 3:3 4:3 5:3 6:3 7:3
```

NOTICE

The storage system adapts to the VLAN priority on the host based on the connection setup request. For NVMe-oF connections that have been set up before the VLAN priority mapping is modified, set up the connections again after the VLAN priority mapping has taken effect.

----End

5.2.1.1.2 SUSE Configuration

Step 1 Use the **eth6** NIC as an example. Modify the configuration files for the ports on the RoCE NICs. You are advised to modify the content as follows:

```
[root@localhost ~]# cat /etc/sysconfig/network/ifcfg-eth6
BOOTPROTO='none'
BROADCAST=''
ETHTOOL_OPTIONS=''
IPADDR=''
MTU=5500
NETMASK=''
GATEWAY=''
NETWORK=''
REMOTE_IPADDR=''
STARTMODE='auto'
DHCLIENT_SET_DEFAULT_ROUTE='yes'
IPV6INIT=no
IPV6_AUTOCONF=no
```

NOTE

- If automatic IP address configuration is required on the physical network port, ensure that DHCP has been correctly configured.
- The default MTU is 1500. For better performance, you are advised to set the MTU to 5500. Ensure that the MTUs of the host physical network port, host VLAN port, storage RoCE port, and storage VLAN are the same.

Step 2 Currently, Huawei storage supports manual configuration of RoCE PFC based on PCP and DSCP. In PCP mode, you must create VLAN ports for the physical network ports on the host. In DSCP mode, you can determine whether to create VLAN ports for the physical network ports on the host based on service requirements. Network port **eth6** is used as an example:

1. Create **change-nvmeof-vlan-egress.sh** in **/etc/sysconfig/network/if-up.d** and add the following content:

```
#!/bin/sh
action=$1
interface=$2
```

```
logger "$0: Action: $action, interface: $interface."
if [ "x$action" == "xpre-up" ]; then
    logger "$0: Set $interface EGRESS priority mappings 0:3 1:3 2:3 3:3 4:3 5:3 6:3 7:3."
    ip link set $interface type vlan egress 0:3 1:3 2:3 3:3 4:3 5:3 6:3 7:3
fi
```

2. Run the following command to grant the execute permission on the **change-nvmeof-vlan-egress.sh** file:

```
[root@localhost /]# chmod a+x /etc/sysconfig/network/if-up.d/change-nvmeof-vlan-egress.sh
```

3. Create a VLAN port **eth6.55** (55 is the VLAN ID, which must be the same as the VLAN ID of the interconnected port on the storage system). The following is an example of the configuration file content:

```
[root@localhost ~]# cat /etc/sysconfig/network/ifcfg-eth6.55
BOOTPROTO='static'
BROADCAST=''
ETHTOOL_OPTIONS=''
IPADDR='192.168.5.5'
NETMASK='255.255.255.0'
MTU='5500'
GATEWAY=''
NETWORK=''
REMOTE_IPADDR=''
STARTMODE='auto'
DHCLIENT_SET_DEFAULT_ROUTE='yes'
ETHERDEVICE='eth6'
VLAN_ID='55'
PRE_UP_SCRIPT='wicked:/etc/sysconfig/network/if-up.d/change-nvmeof-vlan-egress.sh'
```

NOTE

- For all VLAN ports, the VLAN egress priority must be mapped to priority 3. Add **PRE_UP_SCRIPT='wicked:/etc/sysconfig/network/if-up.d/change-nvmeof-vlan-egress.sh'** to the preceding configuration file.
- Repeat this operation on all host service ports to be used.
- The default MTU is 1500. For better performance, you are advised to set the MTU to 5500. Ensure that the MTUs of the host physical network port, host VLAN port, storage RoCE port, and storage VLAN are the same.
- If you use the **nmcli** command to create the VLAN NIC configuration file, set **IPV6INIT=no** and **IPV6_AUTOCONF=no**.

Step 3 Run the following command for the physical network port and VLAN port settings to take effect.

```
[root@localhost ~]# systemctl restart network.service
```

NOTICE

Restarting the network service may cause network exceptions. Ensure that no service is running before performing this operation.

Step 4 Check the IP address of the VLAN port to confirm that the configuration has taken effect.

```
[root@localhost ~]# ifconfig  eth6.55
eth6.55: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 5500
        inet 192.168.5.5  netmask 255.255.255.0  broadcast 192.168.5.255
        inet6 fe80::268a:7ff:fead:7a22  prefixlen 64  scopeid 0x20<link>
        ether 24:8a:07:ad:7a:22  txqueuelen 1000  (Ethernet)
        RX packets 0  bytes 0 (0.0 B)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 54  bytes 7628 (7.4 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

NOTICE

If the same VLAN is configured for different physical network ports on the host and IP addresses in the same network segment are configured for these VLAN ports, add the following content to the **/etc/sysctl.conf** file:

```
net.ipv4.conf.all.arp_ignore=1
```

Then run the **sysctl -w net.ipv4.conf.all.arp_ignore=1** command for the configuration to take effect.

Step 5 Verify that the VLAN egress priority has taken effect.

```
[root@localhost ~]# cat /proc/net/vlan/eth6.55
eth6.55  VID: 55 REORDER_HDR: 1  dev->priv_flags: 1021
         total frames received          132
          total bytes received         6582
      Broadcast/Multicast Rcvd            0

      total frames transmitted          102
       total bytes transmitted         6826
Device: ens129s0f0
INGRESS priority mappings: 0:0  1:0  2:0  3:0  4:0  5:0  6:0 7:0
 EGRESS priority mappings: 0:3 1:3 2:3 3:3 4:3 5:3 6:3 7:3
```

NOTICE

The storage system adapts to the VLAN priority on the host based on the connection setup request. For NVMe-oF connections that have been set up before the VLAN priority mapping is modified, set up the connections again after the VLAN priority mapping has taken effect.

----End

5.2.1.2 Configuring Port PFC

Run the **mlnx_qos** command to configure PFC for all physical network ports used by NVMe over RoCE. Huawei storage supports priority 0 and priority 3 (priority 3 is recommended), and the corresponding parameter is **0,0,0,1,0,0,0,0**.

NOTE

By default, the OS does not support the **mlnx_qos** command. Perform the following operations:

1. Install Python on the OS. Python 3 is recommended.
2. Download the mlnx-tools software package from the GitHub website: https://github.com/Mellanox/mlnx-tools/releases/tag/v5.1.3. (The ibdev2netdev tool is missing in the software packages of versions later than 5.1.3. You are advised to download 5.1.3 and earlier versions.)
3. Upload the software package to the OS and decompress it. The following uses version 5.1.3 as an example
```
tar -xzvf mlnx-tools-5.1.3.tar.gz
```
4. Run the following command to install the utils tool of mlnx-tools (using Python 3 as an example):
```
cd mlnx-tools-5.1.3/ofed_scripts/utils/
python3 setup.py install
```
5. Run the following command to install the scripts tool of mlnx-tools:
```
cd mlnx-tools-5.1.3/ofed_scripts/
cp cma_roce_tos /usr/local/bin/ -v
cp ibdev2netdev /usr/local/bin/ -v
chmod a+x /usr/local/bin/cma_roce_tos
chmod a+x /usr/local/bin/ibdev2netdev
```

5.2.1.2.1 Red Hat, CentOS, and Kylin Configuration

Currently, Huawei storage supports manual configuration of PFC based on PCP and DSCP. The switch ports, host ports, and storage ports must use the same mode.

The following uses the **enp129s0f0** network port as an example to describe the configuration method for PCP:

```
[root@localhost ~]# mlnx_qos -i enp129s0f0 --pfc 0,0,0,1,0,0,0,0
DCBX mode: OS controlled
Priority trust state: pcp
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
         priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
         priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
         priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
         priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
         priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
         priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
         priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
         priority:  7
```

To permanently validate this configuration, modify the **/etc/rc.d/rc.local** file.

Add the PFC configuration of mlnx_qos at the end of **/etc/rc.d/rc.local** and save the modification (using network ports **enp129s0f0** and **enp129s0f1** as an example). The content after modification is as follows:

```
[root@localhost ~]# cat /etc/rc.d/rc.local
#!/bin/bash
# THIS FILE IS ADDED FOR COMPATIBILITY PURPOSES
#
# It is highly advisable to create own systemd services or udev rules
# to run scripts during boot instead of using this file.
#
# In contrast to previous versions due to parallel execution during boot
# this script will NOT be run after all other services.
#
# Please note that you must run 'chmod +x /etc/rc.d/rc.local' to ensure
# that this script will be executed during boot.

touch /var/lock/subsys/local

mlnx_qos -i enp129s0f0 --pfc 0,0,0,1,0,0,0,0
mlnx_qos -i enp129s0f1 --pfc 0,0,0,1,0,0,0,0
```

Modify the permission of **/etc/rc.d/rc.local**.

```
[root@localhost ~]# chmod +x /etc/rc.d/rc.local
```

The following uses the **enp129s0f0** network port as an example to describe the configuration method for DSCP:

Set the trust mode of the network port to **dscp**.

```
[root@localhost ~]# mlnx_qos -i enp129s0f0 --pfc 0,0,0,1,0,0,0,0 --trust dscp
DCBX mode: OS controlled
Priority trust state: dscp
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
         priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
         priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
         priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
         priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
         priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
         priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
         priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
         priority:  7
```

Query the network port name.

```
[root@localhost ~]# /usr/local/bin/ibdev2netdev
mlx5_0 port 1 ==> enp129s0f0 (Up)
mlx5_1 port 1 ==> enp129s0f1 (Up)
```

Set the DSCP value of the network port to **26** in the **mlnx-tools-5.1.3/ofed_scripts** directory.

```
[root@localhost ~]# /usr/local/bin/cma_roce_tos -d mlx5_0 -t 104
104
[root@localhost ~]# /usr/local/bin/cma_roce_tos -d mlx5_1 -t 104
104
```

To permanently validate this configuration, you must create the **/etc/NetworkManager/dispatcher.d/pre-up.d/set-nvmeof-qos** file. The following uses network ports **enp129s0f0** and **enp129s0f1** as an example:

```
[root@localhost ~]# cat /etc/NetworkManager/dispatcher.d/pre-up.d/set-nvmeof-qos
#!/bin/sh
# DSCP
netdevs_dscp=(enp129s0f0 enp129s0f1)


netdev=$1
action=$2

logger "$0: Action: $action, netdev: $netdev."
if [ "x${action}" != "xpre-up" ]; then
    exit 0
fi

# DSCP
for element in "${netdevs_dscp[@]}"; do
    if [[ "$element" =~ "${netdev}" ]]; then
        ibdev=$(/usr/local/bin/ibdev2netdev | /usr/bin/grep -w ${netdev} | /usr/bin/awk '{print $1}')
        /usr/local/bin/mlnx_qos -i ${netdev} --pfc 0,0,0,1,0,0,0,0 --trust=dscp
        /usr/local/bin/cma_roce_tos -d ${ibdev} -t 104
        logger "$0: Set ${netdev} pfc 0,0,0,1,0,0,0 trust dscp, set ${ibdev} tos 104."
        exit 0
    fi
done
```

Modify the permission of **/etc/NetworkManager/dispatcher.d/pre-up.d/set-nvmeof-qos**.

```
[root@localhost ~]# chmod +x /etc/NetworkManager/dispatcher.d/pre-up.d/set-nvmeof-qos
```

After the host is restarted, run the **mlnx_qos -i** command to check the PFC of the network ports.

```
[root@localhost ~]# mlnx_qos -i enp129s0f0
DCBX mode: OS controlled
# Priority trust state: dscp # DSCP
Priority trust state: pcp    # PCP
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
         priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
         priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
         priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
         priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
         priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
         priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
         priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
         priority:  7
```

5.2.1.2.2 SUSE Configuration

Currently, Huawei storage supports manual configuration of PFC based on PCP and DSCP.

The following uses the **eth6** network port as an example to describe the configuration method for PCP:

```
[root@localhost ~]# mlnx_qos -i eth6 --pfc 0,0,0,1,0,0,0,0
DCBX mode: OS controlled
```

```
Priority trust state: pcp
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
         priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
         priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
         priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
         priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
         priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
         priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
         priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
         priority:  7
```

To permanently validate this configuration, modify the **/etc/init.d/after.local** file.

Add the PFC configuration of mlnx_qos at the end of **/etc/init.d/after.local** and save the modification (using network ports **eth6** and **eth7** as an example). The content after modification is as follows:

```
localhost:~ # cat /etc/init.d/after.local
#! /bin/sh
mlnx_qos -i eth6 --pfc 0,0,0,1,0,0,0,0
mlnx_qos -i eth7 --pfc 0,0,0,1,0,0,0,0
```

Modify the permission of **/etc/init.d/after.local**.

```
localhost:~ # chmod +x /etc/init.d/after.local
```

The following uses the **eth6** network port as an example to describe the configuration method for DSCP:

Set the trust mode of the network port to **dscp**.

```
[root@localhost ~]# mlnx_qos -i eth6 --pfc 0,0,0,1,0,0,0,0 --trust dscp
DCBX mode: OS controlled
Priority trust state: dscp
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
         priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
         priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
         priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
         priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
         priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
         priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
         priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
         priority:  7
```

Query the network port name in the **mlnx-tools-5.1.3/ofed_scripts** directory.

```
[root@localhost ~]# ibdev2netdev
mlx5_0 port 1 ==> eth6 (Up)
mlx5_1 port 1 ==> eth7 (Up)
```

Set the DSCP value of the network port to **26** in the **mlnx-tools-5.1.3/ofed_scripts** directory.

```
[root@localhost ~]# ./cma_roce_tos -d mlx5_0 -t 104
104
[root@localhost ~]# ./cma_roce_tos -d mlx5_1 -t 104
104
```

To permanently validate this configuration, you must create the **/etc/sysconfig/network/if-up.d/set-nvmeof-qos-dscp.sh** file. The content is as follows:

```
localhost:~ # cat /etc/sysconfig/network/if-up.d/set-nvmeof-qos-dscp.sh
#!/bin/sh
action=$1
netdev=$2

logger "$0: Action: $action, netdev: $netdev."
if [ "x$action" == "xpre-up" ]; then
        ibdev=`ibdev2netdev | grep -w "$netdev" | awk '{print $1}'`

        mlnx_qos -i $netdev -f 0,0,0,1,0,0,0,0 --trust dscp
        cma_roce_tos -d $ibdev -t 104
        logger "$0: Set $netdev pfc 0,0,0,1,0,0,0 trust dscp, set $ibdev tos 104."
fi
```

Modify the permission of **/etc/sysconfig/network/if-up.d/set-nvmeof-qos-dscp.sh**.

```
localhost:~ # chmod +x /etc/sysconfig/network/if-up.d/set-nvmeof-qos-dscp.sh
```

Modify the configuration files for the ports on the RoCE NICs.

```
localhost:~ # cat /etc/sysconfig/network/ifcfg-eth6
......
PRE_UP_SCRIPT='wicked:/etc/sysconfig/network/if-up.d/set-nvmeof-qos-dscp.sh'
```

After the host is restarted, run the **mlnx_qos -i** command to check the PFC of the network ports.

```
[root@localhost ~]# mlnx_qos -i eth6
DCBX mode: OS controlled
# Priority trust state: dscp # DSCP
Priority trust state: pcp    # PCP
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
        priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
        priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
        priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
        priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
        priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
        priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
        priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
        priority:  7
```

5.2.1.3 Loading the Driver

Load the RDMA driver on the host. The following method loads the driver temporarily:

```
[root@localhost ~]# modprobe nvme-rdma
[root@localhost ~]# modprobe mlx5_ib
```

After loading the driver, run the **lsmod** command to verify that **nvme_rdma** and **mlx5_ib** have been loaded.

```
[root@localhost ~]# lsmod | grep rdma
nvme_rdma              32768  0
nvme_fabrics           24576  1 nvme_rdma
nvme_core              73728  2 nvme_rdma,nvme_fabrics
rdma_cm                69632  1 nvme_rdma
iw_cm                  53248  1 rdma_cm
ib_cm                  57344  1 rdma_cm
ib_core               282624  6 rdma_cm,nvme_rdma,iw_cm,ib_uverbs,mlx5_ib,ib_cm
[root@localhost ~]# lsmod | grep mlx5
mlx5_ib               344064  0
ib_uverbs             147456  2 rdma_ucm,mlx5_ib
ib_core               356352  13
rdma_cm,ib_ipoib,rpcrdma,ib_srpt,ib_srp,iw_cm,ib_iser,ib_umad,ib_isert,rdma_ucm,ib_uverbs,mlx5_ib,ib_cm
mlx5_core            1028096  1 mlx5_ib
mlxfw                  24576  1 mlx5_core
```

You can create and modify the driver loading file for the configuration to take effect permanently. The content is as follows:

```
[root@localhost ~]# cat /etc/modules-load.d/nvme-rdma.conf
nvme-rdma
mlx5_ib
```

5.2.1.4 Discovering and Connecting to a Target

If you have installed and enabled the OceanStor NOF Director, skip this operation. Otherwise, after the driver has been loaded, run the **nvme discover** and **nvme connect** commands to discover and connect to the target. The following example is a temporary configuration:

Step 1 Run the **nvme discover** command to discover the target. **192.168.5.6** is the service IP address configured for the storage port.

```
[root@localhost ~]# nvme discover -t rdma -a 192.168.5.6

Discovery Log Number of Records 1, Generation counter 2
=====Discovery Log Entry 0======
trtype:  rdma
adrfam:  ipv4
subtype: nvme subsystem
treq:    not specified
portid:  5
trsvcid: 4420
subnqn:  nqn.2020-02.huawei.nvme:nvm-subsystem-sn-2102353GSY10L4000003
traddr:  192.168.5.6
rdma_prtype: roce-v2
rdma_qptype: connected
rdma_cms:    rdma-cm
rdma_pkey: 0x0000
```

Step 2 Run the **nvme connect** command to connect to the target. In the command, the NQN following **-n** is the value of **subnqn** queried in the previous step.

```
[root@localhost ~]# nvme connect -t rdma -a 192.168.5.6 -n nqn.2020-02.huawei.nvme:nvm-subsystem-
sn-2102353GSY10L4000003
```

Step 3 Run the **nvme list-subsys** command to verify that the target is successfully connected. If there are multiple targets, repeat the preceding steps to connect each target.

```
[root@localhost ~]# nvme list-subsys
nvme-subsys0 - NQN=nqn.2020-02.huawei.nvme:nvm-subsystem-sn-2102353GSY10L4000003
\
 +- nvme0 rdma traddr=55.55.55.74 trsvcid=4420 live
```

----End

The preceding method is used to temporarily connect to the target. After the host is restarted, it does not automatically reconnect to the target. To enable automatic connection, perform the following steps:

Red Hat, CentOS, and Kylin Configuration

Step 1 Create **nvme_fabrics_persistent.service** in the system service directory **/etc/systemd/system**, which enables the host to automatically connect to the target after startup. The following is an example of the content:

```
[root@localhost ~]# cat /etc/systemd/system/nvme_fabrics_persistent.service
[Unit]
Description=NVMf auto discovery service
Requires=network.target
StartLimitInterval=320
StartLimitBurst=5
After=systemd-modules-load.service network.target network-online.target
[Service]
Type=idle
ExecStart=/usr/sbin/nvme connect-all
StandardOutput=journal
Restart=always
RestartSec=60
[Install]
WantedBy=multi-user.target timers.target
```

Step 2 Set the service to run at system startup.

```
[root@localhost ~]# systemctl enable nvme_fabrics_persistent.service
Created symlink from /etc/systemd/system/multi-user.target.wants/nvme_fabrics_persistent.service to /etc/
systemd/system/nvme_fabrics_persistent.service.
Created symlink from /etc/systemd/system/timers.target.wants/nvme_fabrics_persistent.service to /etc/
systemd/system/nvme_fabrics_persistent.service.
```

Step 3 Create the **discovery.conf** file in the **/etc/nvme** directory and add the target information.

```
[root@localhost ~]# cat /etc/nvme/discovery.conf
--transport rdma --traddr 192.168.5.6 --trsvcid 4420 --ctrl-loss-tmo 3600
--transport rdma --traddr 192.168.6.6 --trsvcid 4420 --ctrl-loss-tmo 3600
```

----End

SUSE Configuration

Step 1 Create **nvme_fabrics_persistent.service** in the system service directory **/etc/systemd/system**, which enables the host to automatically connect to the target after startup. The following is an example of the content:

```
[root@localhost ~]# cat /etc/systemd/system/nvme_fabrics_persistent.service
[Unit]
Description=NVMf auto discovery service
Requires=network.target
After=systemd-modules-load.service network.target network-online.target
[Service]
```

```
Type=idle
ExecStart=/usr/sbin/nvme connect-all
StandardOutput=journal
Restart=always
RestartSec=60
StartLimitInterval=320
StartLimitBurst=5
[Install]
WantedBy=multi-user.target timers.target
```

Step 2 Set the service to run at system startup.

```
[root@localhost ~]# systemctl enable nvme_fabrics_persistent.service
Created symlink from /etc/systemd/system/multi-user.target.wants/nvme_fabrics_persistent.service to /etc/
systemd/system/nvme_fabrics_persistent.service.
Created symlink from /etc/systemd/system/timers.target.wants/nvme_fabrics_persistent.service to /etc/
systemd/system/nvme_fabrics_persistent.service.
```

Step 3 Create the **discovery.conf** file in the **/etc/nvme** directory and add the target information.

```
[root@localhost ~]# cat /etc/nvme/discovery.conf
--transport rdma --traddr 192.168.5.6 --trsvcid 4420 --ctrl-loss-tmo 3600
--transport rdma --traddr 192.168.6.6 --trsvcid 4420 --ctrl-loss-tmo 3600
```

----End

NOTE

- If you have installed and enabled the OceanStor NOF Director on the host, and the switch and storage configurations meet the requirements for using the OceanStor NOF Director, the connections between the host and storage system can be automatically established. You do not need to manually discover and connect to the target.
- With the automatic connection function, the host automatically attempts to connect to the target every minute after the host is restarted, and retries for a maximum of five times. If the connection is not restored, run the **nvme connect-all** command to manually establish the connection.
- If the OceanStor NOF Enabler is not used, you may need to manually run the **nvme connect-all** command to establish connections during storage system reboot and rolling upgrade.
- For details about the preceding configuration, see the Mellanox official configuration document at **https://community.mellanox.com/s/article/howto-configure-persistent-nvme-over-fabrics-initiator**.

5.2.2 Storage System Configuration

Adding Initiators for the Host

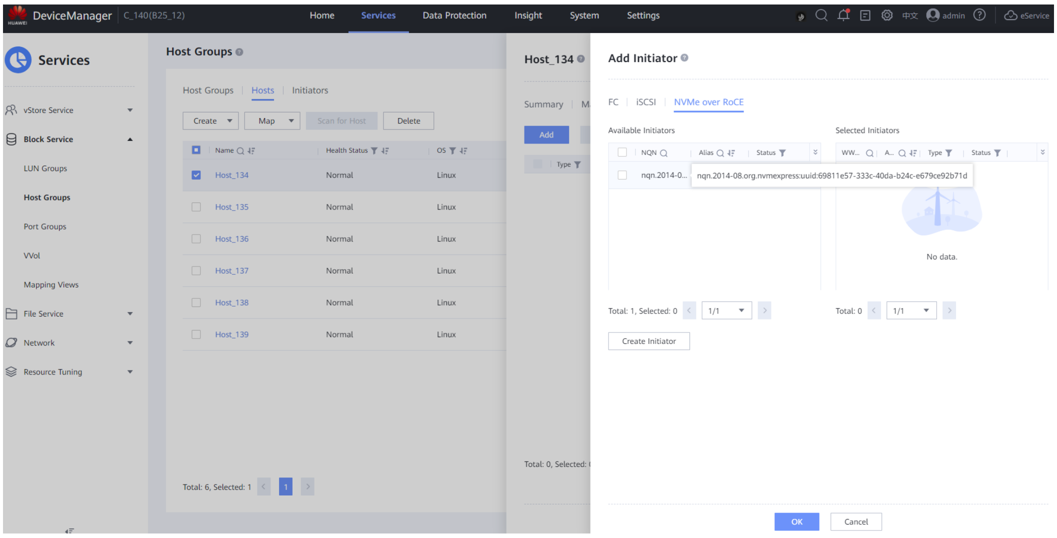
After the preceding configuration is complete, perform the following steps to add initiators to the hosts on the storage system:

Step 1 Query the initiator NQN on the host.

```
[root@localhost ~]# cat /etc/nvme/hostnqn
nqn.2014-08.org.nvmexpress:uuid:69811e57-333c-40da-b24c-e679ce92b71d
```

Step 2 Log in to DeviceManager of the storage system and choose **Services** > **Block Service** > **Host Groups** > **Hosts**. Select the desired host, click **More** on the right, and choose **Add Initiator**. On the **NVMe over RoCE** tab, select the initiator NQN queried in **Step 1** and click **OK**.

Figure 5-7 Adding an initiator



NOTICE

If multiple Linux hosts have the same host NQN, the initiator of only one host can be added to the storage device. Currently, nvme-cli-1.9-5-el8 of CentOS 8.2 has this issue. You are advised to upgrade the NVMe-CLI tool to a later version or perform the following steps to change the host NQNs:

1. Run the **nvme gen-hostnqn** command to generate a new host NQN.
```
[root@localhost ~]# nvme gen-hostnqn
nqn.2014-08.org.nvmexpress:uuid:899c9cae-5807-4d2e-88be-50739cde8f4e
```
2. Use the new host NQN to overwrite the original in **/etc/nvme/hostnqn**.
```
[root@localhost ~]# vi /etc/nvme/hostnqn

[root@localhost ~]# cat /etc/nvme/hostnqn
nqn.2014-08.org.nvmexpress:uuid:899c9cae-5807-4d2e-88be-50739cde8f4e
```
3. Restart the host for the new host NQN to take effect. Then add the updated NVMe over RoCE initiator to the Linux host on the storage device.

----End

5.3 Scanning LUNs on the Host

After establishing connections, scan LUNs on the host.

Run the **nvme list** command to query the scanned NVMe disks.

```
[root@localhost dev]# nvme list
Node             SN                   Model                                    Namespace Usage
Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- --------------------------
---------------- --------
/dev/nvme0n1     2102352TVE10K6000002 Huawei-XSG1                              1          21.49  GB / 107.37
GB    512   B +  0 B   1000001
/dev/nvme0n2     2102352TVE10K6000002 Huawei-XSG1                              2          10.74  GB /  53.69
GB    512   B +  0 B   1000001
/dev/nvme1n1     2102352TVE10K6000002 Huawei-XSG1                              1          21.49  GB / 107.37
GB    512   B +  0 B   1000001
/dev/nvme1n2     2102352TVE10K6000002 Huawei-XSG1                              2          10.74  GB /  53.69
GB    512   B +  0 B   1000001
```

If the LUN mapping or capacity is changed, perform the following operations to update the NVMe disk information:

Step 1 Go to the **/dev** directory and query the NVMe port numbers.

```
[root@localhost ~]# cd /dev
[root@localhost dev]# ls | grep nvme
nvme0
nvme1
nvme-fabrics
```

In this example, the NVMe port numbers are **nvme0** and **nvme1**.

Step 2 Run the **nvme ns-rescan /dev/nvme***X* command. *X* is the port number queried in step 1.

```
[root@localhost dev]# nvme ns-rescan /dev/nvme0
[root@localhost dev]# nvme ns-rescan /dev/nvme1
```

Step 3 Run the **nvme list** command again to query the updated NVMe disks.

```
[root@localhost dev]# nvme list
Node             SN                   Model                                    Namespace Usage
Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- --------------------------
---------------- --------
/dev/nvme0n1     2102352TVE10K6000002 Huawei-XSG1                              1          21.49  GB / 107.37
GB    512   B +  0 B   1000001
/dev/nvme0n2     2102352TVE10K6000002 Huawei-XSG1                              2          10.74  GB /  53.69
GB    512   B +  0 B   1000001
/dev/nvme1n1     2102352TVE10K6000002 Huawei-XSG1                              1          21.49  GB / 107.37
GB    512   B +  0 B   1000001
/dev/nvme1n2     2102352TVE10K6000002 Huawei-XSG1                              2          10.74  GB /  53.69
GB    512   B +  0 B   1000001
```

----End

6 Configuring Multipathing

This chapter describes the multipathing configurations on the hosts and storage systems.

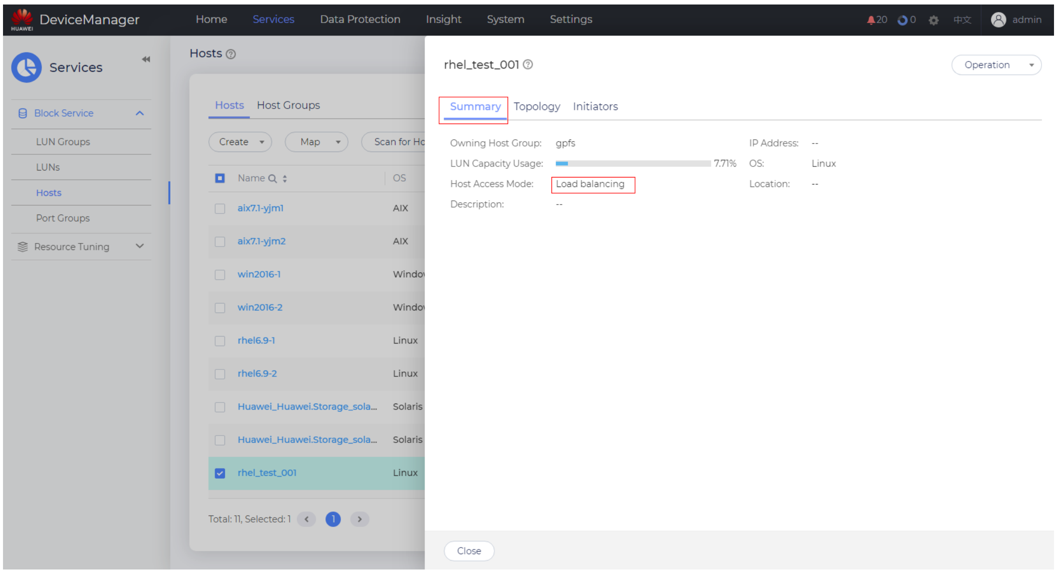
6.1 Non-HyperMetro Scenarios

6.1.1 UltraPath

6.1.1.1 Storage System Configuration

If UltraPath is used in non-HyperMetro scenarios, retain the default host and initiator settings. By default, **Host Access Mode** is **Load balancing**. You can click the host name and check the settings on the **Summary** tab page.

Figure 6-1 Checking storage configurations



Name	OS
aix7.1-yjm1	AIX
aix7.1-yjm2	AIX
win2016-1	Windo
win2016-2	Windo
rhel6.9-1	Linux
rhel6.9-2	Linux
Huawei_Huawei.Storage_sola...	Solaris
Huawei_Huawei.Storage_sola...	Solaris
rhel_test_001	Linux

Total: 11, Selected: 1

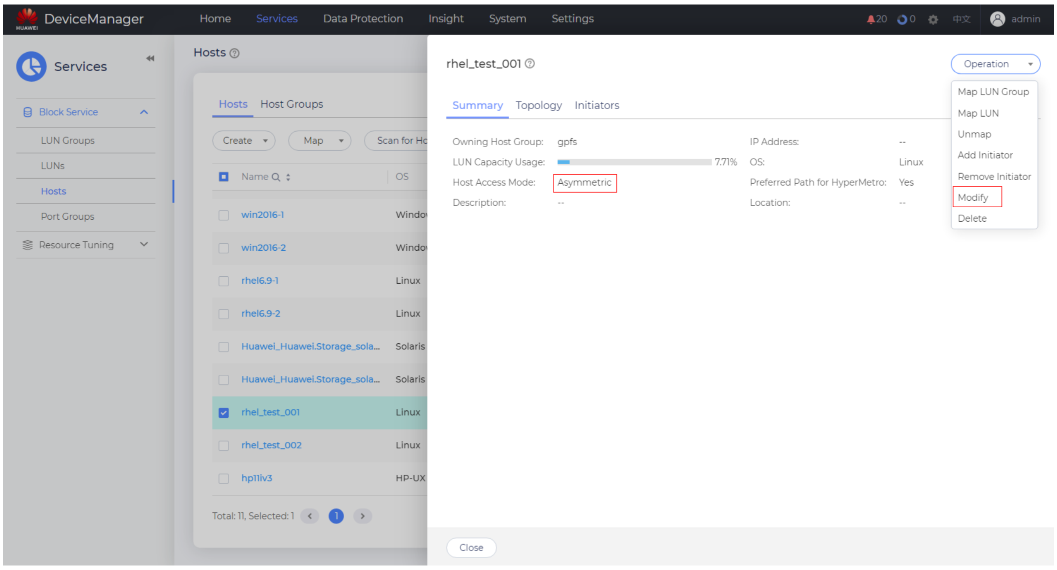
rhel_test_001

Owning Host Group: gpfs
LUN Capacity Usage: 7.71%
Host Access Mode: Load balancing
Description: --
IP Address: --
OS: Linux
Location: --

If **Host Access Mode** is not **Load balancing**, perform the following steps to change it:

Step 1 Click the host name and choose **Operation** > **Modify**.

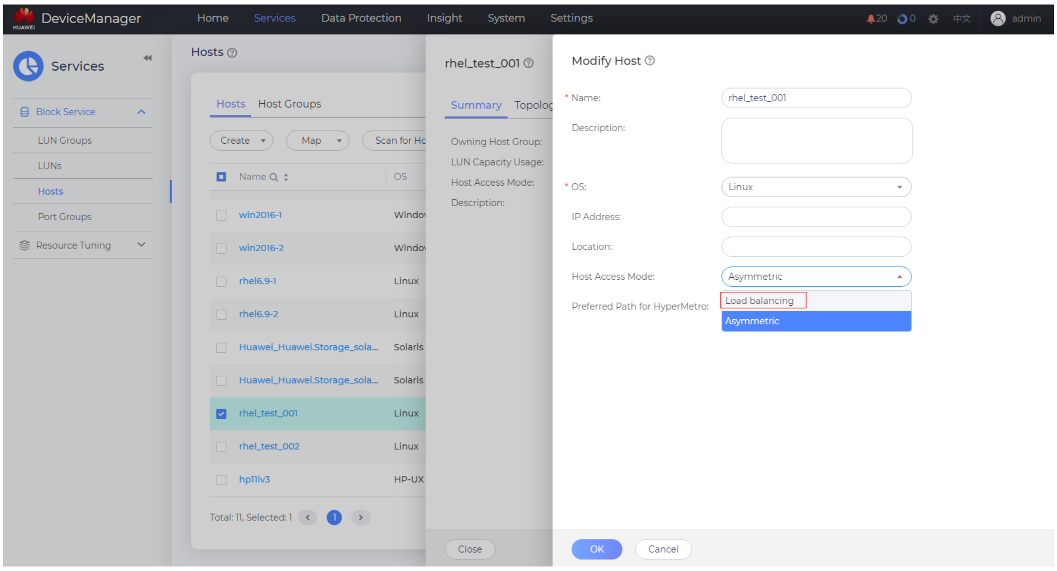
Figure 6-2 Modifying the host properties



Owning Host Group: gpfs
LUN Capacity Usage: 7.71%
Host Access Mode: Asymmetric
Description: --
IP Address: --
OS: Linux
Preferred Path for HyperMetro: Yes
Location: --

Step 2 Set **Host Access Mode** to **Load balancing** and click **OK**.

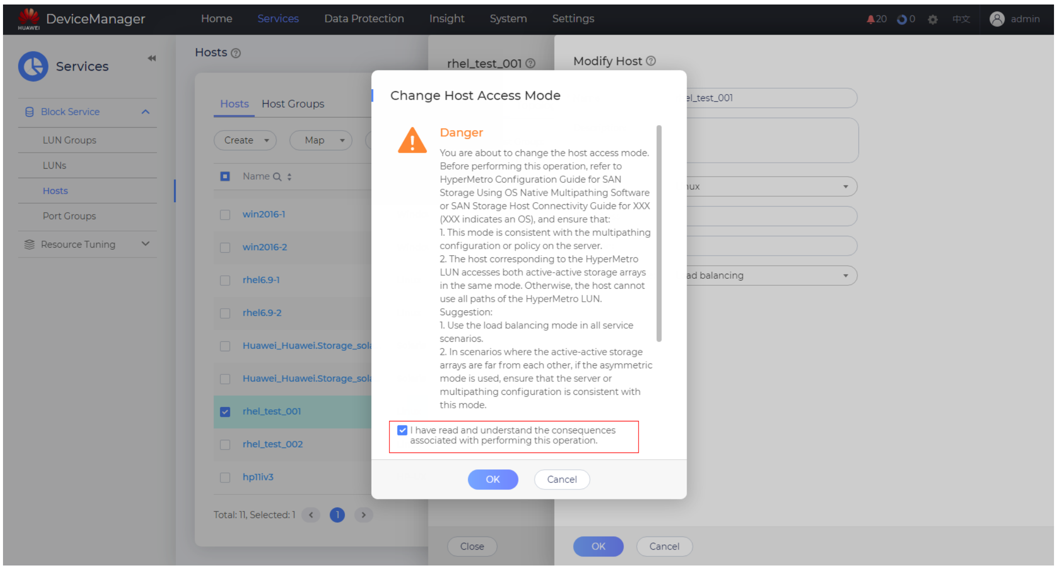
Figure 6-3 Modifying the host access mode



* Name: rhel_test_001
Description:
* OS: Linux
IP Address:
Location:
Host Access Mode: Asymmetric
Preferred Path for HyperMetro:

Step 3 Confirm the information and click **OK**.

Figure 6-4 Confirming the operation



Change Host Access Mode

Danger

You are about to change the host access mode. Before performing this operation, refer to HyperMetro Configuration Guide for SAN Storage Using OS Native Multipathing Software or SAN Storage Host Connectivity Guide for XXX (XXX indicates an OS), and ensure that:
1. This mode is consistent with the multipathing configuration or policy on the server.
2. The host corresponding to the HyperMetro LUN accesses both active-active storage arrays in the same mode. Otherwise, the host cannot use all paths of the HyperMetro LUN.
Suggestion:
1. Use the load balancing mode in all service scenarios.
2. In scenarios where the active-active storage arrays are far from each other, if the asymmetric mode is used, ensure that the server or multipathing configuration is consistent with this mode.

☑ I have read and understand the consequences associated with performing this operation.

OK Cancel

----End

NOTICE

- For details about the Linux versions, see the **Huawei Storage Interoperability Navigator**.
- If a LUN has been mapped to a host, you must restart the host for the configuration to take effect after you modify **Host Access Mode**. If you configure the host for the first time, restart is not needed.

6.1.1.2 Host Configuration

Refer to the *OceanStor UltraPath-NVMe xx for Linux User Guide* for the installation method and precautions of UltraPath.

NOTICE

UltraPath 31.0.RC1 and later versions support the NVMe functions in Linux. Select a user guide that matches your UltraPath version. The link is as follows:

https://support.huawei.com/enterprise/en/cloud-storage/ultrapath-pid-8576127?category=operation-maintenance&subcategory=user-guide

6.1.1.3 Verification

Run the **upadmin_plus show vlun** command to verify that the configuration has taken effect. The following is an example:

```
[root@localhost ~]# upadmin_plus show vlun
-------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------
 Vlun ID     Disk      Name                      Lun WWN              Status  Capacity  Ctrl(Own/Work)     Array
```

```
Name    Dev Lun ID  No. of Paths(Available/Total)
   0    ultrapathabh  lun_50GB_0000   71005ebff50065cfc469f09900000038  Normal  50.00GB      --/--
Huawei.Storage     56                4/4
   1    ultrapathabi  lun_50GB_0001   71005ebff50065cfc469f09900000039  Normal  50.00GB      --/--
Huawei.Storage     57                4/4
-------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------
```

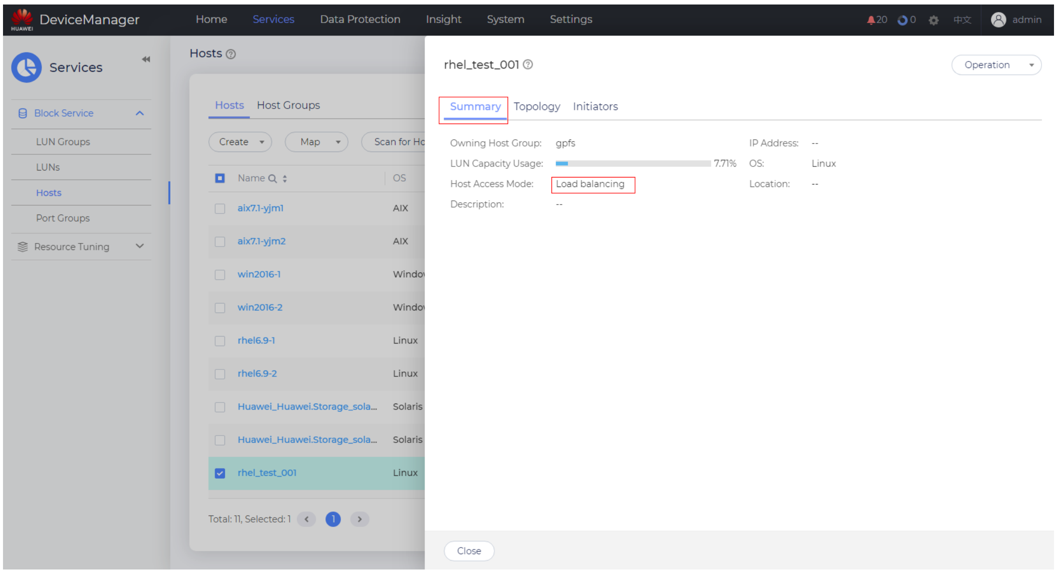
In the command output, the two mapped LUNs have been taken over by the native multipathing software of the system and all LUNs are in the normal state, indicating that the configuration has taken effect.

6.1.2 OS Native Device Mapper

6.1.2.1 Storage System Configuration

If you use UltraPath in non-HyperMetro scenarios, retain the default host and initiator settings. By default, the **Host Access Mode** is **Load balancing**. You can click the host name and check the settings on the **Summary** tab page.

Figure 6-5 Checking storage configurations



Name	OS
aix7.1-yjm1	AIX
aix7.1-yjm2	AIX
win2016-1	Windo
win2016-2	Windo
rhel6.9-1	Linux
rhel6.9-2	Linux
Huawei_Huawei.Storage_sola...	Solaris
Huawei_Huawei.Storage_sola...	Solaris
☑ rhel_test_001	Linux

Total: 11, Selected: 1

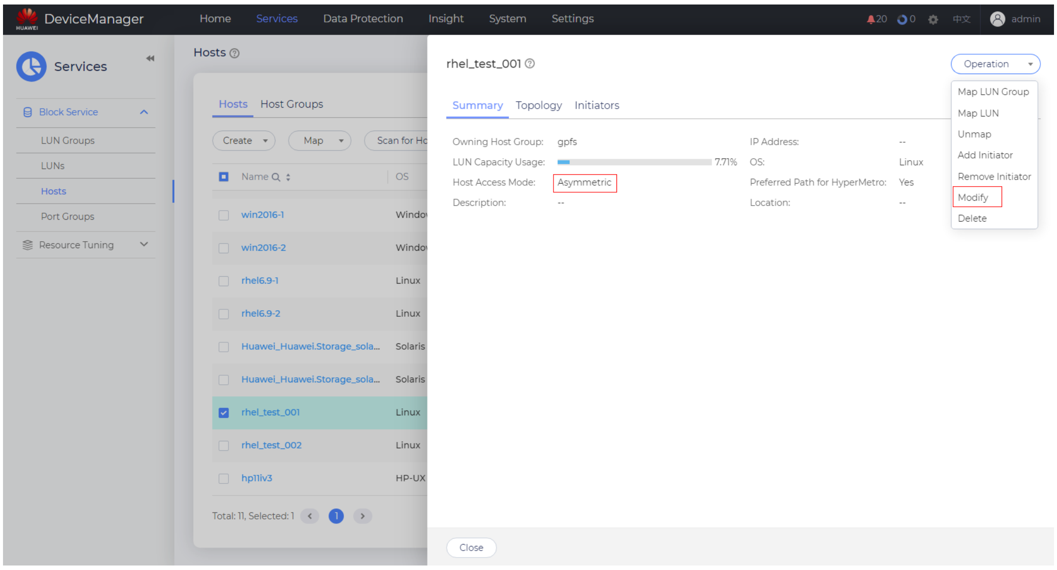
rhel_test_001

Owning Host Group: gpfs
LUN Capacity Usage: 7.71%
Host Access Mode: Load balancing
Description: --
IP Address: --
OS: Linux
Location: --

If the **Host Access Mode** is not **Load balancing**, perform the following steps to change it:

Step 1 Click the host name and choose **Operation** > **Modify**.

Figure 6-6 Modifying the host properties



Hosts

Name	OS
win2016-1	Windo
win2016-2	Windo
rhel6.9-1	Linux
rhel6.9-2	Linux
Huawei_Huawei.Storage_sola...	Solaris
Huawei_Huawei.Storage_sola...	Solaris
☑ rhel_test_001	Linux
rhel_test_002	Linux
hp11iv3	HP-UX

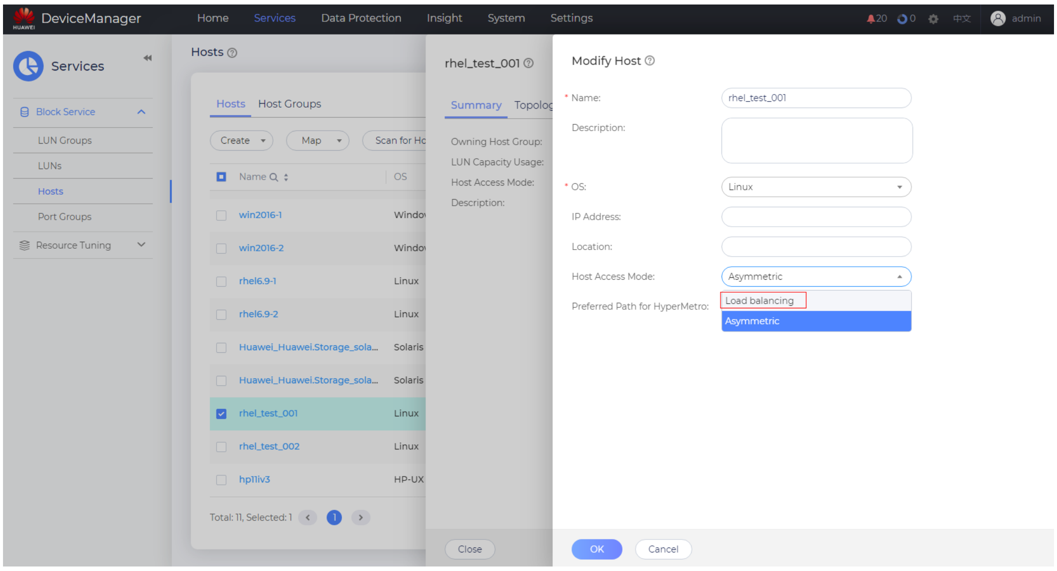
Total: 11, Selected: 1

rhel_test_001

Owning Host Group: gpfs
LUN Capacity Usage: 7.71%
Host Access Mode: Asymmetric
Description: --
IP Address: --
OS: Linux
Preferred Path for HyperMetro: Yes
Location: --

Step 2 Set **Host Access Mode** to **Load balancing** and click **OK**.

Figure 6-7 Modifying the host access mode

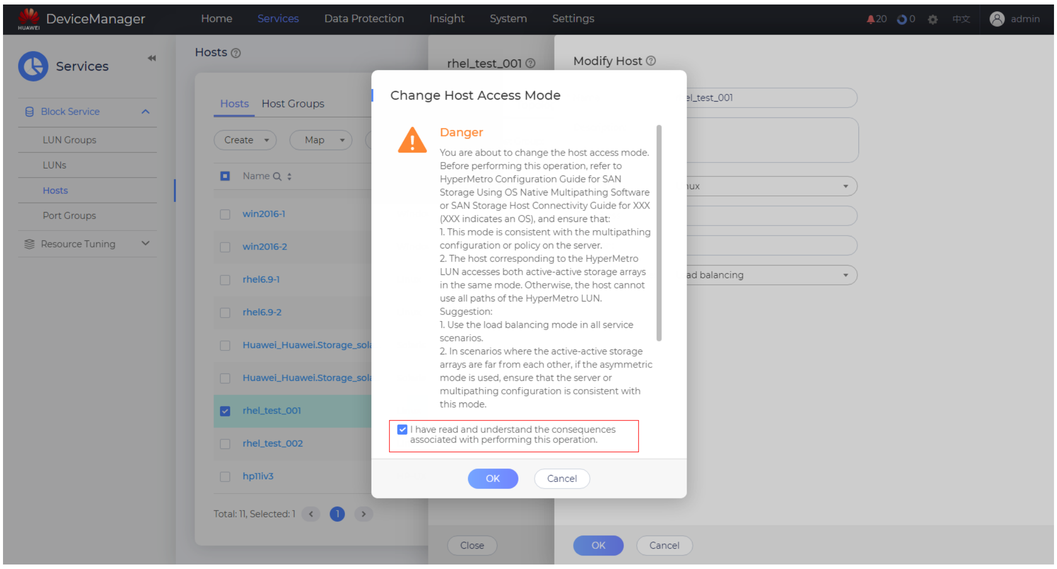


Modify Host

* Name: rhel_test_001
Description:
* OS: Linux
IP Address:
Location:
Host Access Mode: Asymmetric
Preferred Path for HyperMetro:

Step 3 Confirm the alarm information and click **OK**.

Figure 6-8 Confirming the operation



Change Host Access Mode

Danger

You are about to change the host access mode. Before performing this operation, refer to HyperMetro Configuration Guide for SAN Storage Using OS Native Multipathing Software or SAN Storage Host Connectivity Guide for XXX (XXX indicates an OS), and ensure that:
1. This mode is consistent with the multipathing configuration or policy on the server.
2. The host corresponding to the HyperMetro LUN accesses both active-active storage arrays in the same mode. Otherwise, the host cannot use all paths of the HyperMetro LUN.
Suggestion:
1. Use the load balancing mode in all service scenarios.
2. In scenarios where the active-active storage arrays are far from each other, if the asymmetric mode is used, ensure that the server or multipathing configuration is consistent with this mode.

☑ I have read and understand the consequences associated with performing this operation.

OK Cancel

----End

NOTICE

1. For details about supported Linux versions, see **Huawei Storage Interoperability Navigator**.

2. If a LUN has been mapped to a host, you must restart the host for the configuration to take effect after you modify the **Host Access Mode**. If you configure the host for the first time, restart is not needed.

6.1.2.2 Host Configuration

Installing Multipathing Software

Generally, the built-in multipathing software package of the Linux operating system is an RPM package starting with **device-mapper-multipath**. You can run the following command to check whether the corresponding software package has been installed:

```
[root@localhost ~]# rpm -qa | grep multipath
device-mapper-multipath-0.7.8-7.el8.x86_64
device-mapper-multipath-libs-0.7.8-7.el8.x86_64
```

If the query result is empty, the software package is not installed. You can obtain the software package from the system image and run the **rpm** command to install the software package, or use an operating system tool (such as YUM or YaST) to install the software package.

Disabling Native NVMe Multipath

When the native multipathing software of the OS is used, you must disable Native NVMe Multipath. You can run the following command to check the Native NVMe Multipath status:

```
linux-freh:~ # cat /sys/module/nvme_core/parameters/multipath
Y
```

If **No such file or directory** or **N** is displayed in the command output, you do not need to disable Native NVMe Multipath.

If **Y** is displayed in the command output, Native NVMe Multipath is enabled. Perform the following operations to disable it:

Step 1 Add **nvme_core.multipath=N** to the **GRUB_CMDLINE_LINUX_DEFAULT** entry in the **/etc/default/grub** file to disable Native NVMe Multipath.

```
[root@localhost ~]# cat /etc/default/grub | grep multipath
GRUB_CMDLINE_LINUX_DEFAULT="splash=silent resume=/dev/disk/by-path/pci-0000:00:10.0-scsi-0:0:0:0-
part4 mitigations=auto quiet nvme_core.multipath=N crashkernel=180M,high"
```

Step 2 If the **/sys/firmware/efi** directory exists in the OS, the OS uses the EFI boot mode. In this case, run the **grub2-mkconfig -o /boot/efi/EFI/centos/grub.cfg** command to generate a new grub configuration file. (**centos** indicates the OS vendor and varies depending on the OS.) If the traditional boot mode is used, run the **grub2-mkconfig -o /boot/grub2/grub.cfg** command to regenerate the grub configuration file.

Step 3 Run the **reboot** command to restart the host.

Step 4 Run the **cat /sys/module/nvme_core/parameters/multipath** command to check whether the modification takes effect.

```
[root@localhost ~]# cat /sys/module/nvme_core/parameters/multipath
N
```

----End

Configuring the Multipath Configuration File

DM-Multipath's most important configuration file is **/etc/multipath.conf**.

Some operating systems have such a file by default. You can copy the **multipath.conf** or **multipath.conf.synthetic** file to the **/etc** directory to obtain the template. If the file does not exist, manually create the **/etc/multipath.conf** file.

```
[root@localhost ~]# cp /usr/share/doc/device-mapper-multipath-0.4.9/multipath.conf /etc/multipath.conf
```

Redhat/CentOS 8.x:

Edit the multipath configuration file **/etc/multipath.conf**. You are advised to add the following content for this version:

```
defaults {
                polling_interval            1
                user_friendly_names          yes
}

blacklist_exceptions {
        property "(ID_WWN|SCSI_IDENT_.*|ID_SERIAL|DEVTYPE)"
        devnode "nvme*"
}

devices {
    device {
                vendor                      "NVME"
                product                     "Huawei-XSG1"
                uid_attribute                "ID_WWN"
                no_path_retry                 12
                rr_min_io_rq                  1
                prio                        "const"
                path_grouping_policy        multibus
                path_checker                  "directio"
                failback                    immediate
                fast_io_fail_tmo             0
                retain_attached_hw_handler  "no"
}
}
```

⚠ CAUTION

If the host is in a cluster, you are advised to set **no_path_retry** to **2**.

SuSE 15 SPx:

Edit the multipath configuration file **/etc/multipath.conf**. You are advised to add the following content tfor this version:

```
defaults {
            polling_interval            1
            user_friendly_names          yes
            enable_foreign              nvme
}

blacklist_exceptions {
        property "(ID_WWN|SCSI_IDENT_.*|ID_SERIAL|DEVTYPE)"
        devnode "nvme*"
}

devices {
    device {
            vendor                      "NVME"
            product                     "Huawei-XSG1"
            uid_attribute                "ID_WWN"
            no_path_retry                 12
            rr_min_io                    100
            path_grouping_policy         multibus
            path_checker                  "directio"
            prio                        "const"
            failback                    immediate
            retain_attached_hw_handler  "no"
}
}
```

⚠ CAUTION

1. If the host is in a cluster, you are advised to set **no_path_retry** to **2**.

2. For the SUSE operating system, you are advised to add the wwid of the local disk of the operating system to the blacklist field to prevent system startup exceptions caused by system disk takeover by the multipathing software. For details about the configuration method, see the following link:

https://documentation.suse.com/sles/12-SP4/html/SLES-all/cha-multipath.html#sec-multipath-blacklist

3. The **enable_foreign** parameter in the **defaults** field is introduced in SUSE 15 SP2. In earlier versions, this parameter does not need to be configured.

Starting the Multipathing Software

Run the following command to start the multipathing process:

```
systemctl start multipathd.service
```

Restarting the Multipathing Software

If **/etc/multipath.conf** is modified after the multipathing service has been started, you must restart or reload the multipathing service for the modification to take effect. The following is an example:

```
systemctl restart multipathd.service
systemctl reload multipathd.service
```

Setting the Multipathing Software to Run at System Startup

Run the following command to set the multipathing software to run at system startup:

```
systemctl enable multipathd.service
```

6.1.2.3 Verification

Run the **multipath -ll** command to verify that the configuration has taken effect. The following is an example:

```
[root@localhost dev]# multipath -ll
mpathe (eui.710037421701f65216212c2200000026) dm-14 NVME,Huawei-XSG1
size=50G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='service-time 0' prio=1 status=active
  |- 0:0:2:2 nvme0n2 259:1 active ready running
  `- 1:0:2:2 nvme1n2 259:3 active ready running
mpathd (eui.710037421701f64516212c7400000025) dm-15 NVME,Huawei-XSG1
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='service-time 0' prio=1 status=active
  |- 0:0:1:1 nvme0n1 259:0 active ready running
  `- 1:0:1:1 nvme1n1 259:2 active ready running
```

In the command output, the two mapped LUNs have been taken over by the native multipathing software of the system and all paths are active, indicating that the configuration has taken effect.

6.1.3 NVMe Native Multipathing Software

6.1.3.1 Storage System Configuration

If Native NVMe Multipath is used, ensure that the **Host Access Mode** on the storage system is **Load balancing** and no more configuration on the storage side is required. For details, see **6.1.2.1.**

6.1.3.2 Host Configuration

Checking the Multipath Status

Run the following command to check the NVMe Native multipathing status:

```
localhost:~ # cat /sys/module/nvme_core/parameters/multipath
Y
```

If the query result is **Y**, Native NVMe Multipath has been enabled and no additional configuration is required.

If the query result is **N**, you must enable NVMe Native multipathing. For details on how to enable NVMe Native multipathing, see **6.1.2.2**. You only need to change **nvme_core.multipath=N** to **nvme_core.multipath=Y**.

Querying the Multipathing Policy

Run the following command to query the multipathing policy:

```
localhost:~ # ls /sys/class/nvme-subsystem/ | grep subsys
nvme-subsys0
localhost:~ # cat /sys/class/nvme-subsystem/nvme-subsys0/iopolicy
numa
```

The default multipathing policy is **numa**. In this policy, the multipathing software delivers I/Os over only one path. You must modify the policy for load balancing.

Modifying the Multipathing Policy

To temporarily change the multipathing policy to **round-robin**, run the following command:

```
localhost:~ # echo round-robin > /sys/class/nvme-subsystem/nvme-subsys0/iopolicy
localhost:~ # cat /sys/class/nvme-subsystem/nvme-subsys0/iopolicy
round-robin
```

To permanently change the policy, perform the following steps:

Step 1 Create a service in the system service directory **/etc/systemd/system** and add the following content. The service name can be self-defined, for example, **nvme_aa_round-robin.service**.

```
localhost:~ # cat /etc/systemd/system/nvme_aa_round-robin.service
[Unit]
Description=Add active/active support to native NVMe multipath
After=systemd-modules-load.service
[Service]
Type=oneshot
ExecStart=/bin/sh -c "echo round-robin > /sys/class/nvme-subsystem/nvme-subsys0/iopolicy"
[Install]
WantedBy=default.target
```

⚠ CAUTION

1. In the preceding information, **nvme-subsys0** needs to be changed based on the real environment, especially when there are more than one subsys.

Step 2 Verify the service.

```
localhost:~ # cat /sys/class/nvme-subsystem/nvme-subsys0/iopolicy
numa
localhost:~ #
localhost:~ # systemctl start nvme_aa_round-robin.service
localhost:~ # cat /sys/class/nvme-subsystem/nvme-subsys0/iopolicy
round-robin
```

Step 3 Create a timer in the system directory **/etc/systemd/system** and add the following content. The service name can be self-defined, for example, **nvme_aa_round-robin.timer**.

```
localhost:~ # cat /etc/systemd/system/nvme_aa_round-robin.timer
[Unit]
Description=Add active/active support to native NVMe multipath
[Timer]
OnUnitActiveSec=20
Unit=nvme_aa_round-robin.service
[Install]
WantedBy=multi-user.target timers.target
```

Step 4 Set the service to run at system startup.

```
localhost:~ # systemctl enable nvme_aa_round-robin.service
localhost:~ # systemctl enable nvme_aa_round-robin.timer
```

----End

6.1.3.3 Verification

Run the **nvme list** command to query disks. When Native NVMe Multipath is used, each namespace has only one drive letter, regardless of the number of its physical paths.

```
localhost:~ # nvme list
Node             SN                   Model                                    Namespace Usage
Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- --------------------------
---------------- --------
/dev/nvme0n1     2102354XBA10N9100003 Huawei-XSG1                              1           2.15  GB /
107.37  GB    512   B +  0 B   1000001
/dev/nvme0n2     2102354XBA10N9100003 Huawei-XSG1                              2           2.15  GB /
107.37  GB    512   B +  0 B   1000001
```

Run the **multipath -ll** command to check the path status. In the following example, two LUNs are mapped to the host, which correspond to two namespaces on the host. Each namespace has two physical paths, both in the live state.

```
localhost:~ # multipath -ll
eui.71005ebff5021fd6c469f0f900000039 [nvme]:nvme0n1 NVMe,Huawei-XSG1,1000001
size=209715200 features='n/a' hwhandler='ANA' wp=rw
|-+- policy='n/a' prio=50 status=optimized
| `- 0:0:1   nvme0c0n1 0:0 n/a   optimized live
`-+- policy='n/a' prio=50 status=optimized
  `- 0:1:1   nvme0c1n1 0:0 n/a   optimized live
eui.71005ebff5021fd6c469f0f90000003a [nvme]:nvme0n2 NVMe,Huawei-XSG1,1000001
size=209715200 features='n/a' hwhandler='ANA' wp=rw
```

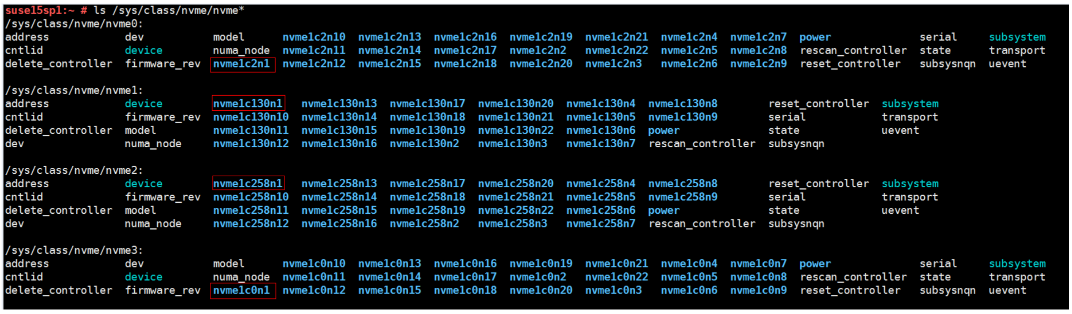
```
|-+- policy='n/a' prio=50 status=optimized
| `- 0:0:1   nvme0c0n1 0:0 n/a   optimized live
`-+- policy='n/a' prio=50 status=optimized
 `- 0:1:1   nvme0c1n1 0:0 n/a   optimized live
```

NOTE

1. In some new OS versions (such as SUSE 12 SP5, 15 SP2, or CentOS 8.3 and later versions), external dynamic libraries are disabled by default, including **libforeign-nvme.so**. If the **libforeign-nvme.so** dynamic library is not loaded, the device parameters generated by nvme-multipath cannot be queried. As a result, no command output is displayed after the **multipath -ll** command is executed. To solve the problem, edit **multipath.conf** and add the **enable_foreign** parameter to the **defaults** field. The following is an example:

```
localhost:~ # cat /etc/multipath.conf
 defaults {
     user_friendly_names yes
     find_multipaths yes
     enable_foreign "nvme"
 }
```

2. In the latest version, the output of the **multipath -ll** command displays the same logical path for different namespaces due to a defect of the multipath service. You can run the **ls /sys/class/nvme/nvme*** command to query all logical paths. In the following example, **nvme1c2n1**, **nvme1c130n1**, **nvme1c258n1**, and **nvme1c0n1** are the four logical paths of namespace **nvme1n1**.



```
suse15sp1:~ # ls /sys/class/nvme/nvme*
/sys/class/nvme/nvme0:
address           dev           model        nvme1c2n10   nvme1c2n13   nvme1c2n16   nvme1c2n19   nvme1c2n21   nvme1c2n4   nvme1c2n7   power              serial      subsystem
cntlid            device        numa_node    nvme1c2n11   nvme1c2n14   nvme1c2n17   nvme1c2n2    nvme1c2n22   nvme1c2n5   nvme1c2n8   rescan_controller  state       transport
delete_controller firmware_rev  nvme1c2n1    nvme1c2n12   nvme1c2n15   nvme1c2n18   nvme1c2n20   nvme1c2n3    nvme1c2n6   nvme1c2n9   reset_controller   subsysnqn   uevent

/sys/class/nvme/nvme1:
address           device        nvme1c130n1   nvme1c130n13  nvme1c130n17  nvme1c130n20  nvme1c130n4  nvme1c130n8        reset_controller  subsystem
cntlid            firmware_rev  nvme1c130n10  nvme1c130n14  nvme1c130n18  nvme1c130n21  nvme1c130n5  nvme1c130n9        serial            transport
delete_controller model         nvme1c130n11  nvme1c130n15  nvme1c130n19  nvme1c130n22  nvme1c130n6  power              state             uevent
dev               numa_node     nvme1c130n12  nvme1c130n16  nvme1c130n2   nvme1c130n3   nvme1c130n7  rescan_controller  subsysnqn

/sys/class/nvme/nvme2:
address           device        nvme1c258n1   nvme1c258n13  nvme1c258n17  nvme1c258n20  nvme1c258n4  nvme1c258n8        reset_controller  subsystem
cntlid            firmware_rev  nvme1c258n10  nvme1c258n14  nvme1c258n18  nvme1c258n21  nvme1c258n5  nvme1c258n9        serial            transport
delete_controller model         nvme1c258n11  nvme1c258n15  nvme1c258n19  nvme1c258n22  nvme1c258n6  power              state             uevent
dev               numa_node     nvme1c258n12  nvme1c258n16  nvme1c258n2   nvme1c258n3   nvme1c258n7  rescan_controller  subsysnqn

/sys/class/nvme/nvme3:
address           dev           model        nvme1c0n10   nvme1c0n13   nvme1c0n16   nvme1c0n19   nvme1c0n21   nvme1c0n4   nvme1c0n7   power              serial      subsystem
cntlid            device        numa_node    nvme1c0n11   nvme1c0n14   nvme1c0n17   nvme1c0n2    nvme1c0n22   nvme1c0n5   nvme1c0n8   rescan_controller  state       transport
delete_controller firmware_rev  nvme1c0n1    nvme1c0n12   nvme1c0n15   nvme1c0n18   nvme1c0n20   nvme1c0n3    nvme1c0n6   nvme1c0n9   reset_controller   subsysnqn   uevent
```

6.2 HyperMetro Scenarios

This section describes the multipathing software configurations on the hosts and storage systems. For details about how to configure HyperMetro services, see the *HyperMetro Feature Guide for Block*.

Storage System Configuration

Table 6-1 provides the configurations of **Host Access Mode** and **Preferred Path for HyperMetro**.

Table 6-1 Storage configurations for interconnection with Red Hat, CentOS, Kylin, and SUSE application servers

HyperMetro Working Mode	Storage System	Host Access Mode	Preferred Path for HyperMetro	Description
Load balancing mode	Local storage	Load balancing	N/A	The host uses all paths of a disk with equal priority.

HyperMetro Working Mode	Storage System	Host Access Mode	Preferred Path for HyperMetro	Description
	Remote storage	Load balancing	N/A	
Local preferred mode	Local storage	Asymmetric	Yes	The host considers the paths from the local storage system as preferred paths, and those from the remote storage system as non-preferred paths.
	Remote storage	Asymmetric	No	

NOTICE

- For details about the Linux versions, see the **Huawei Storage Interoperability Navigator**.
- If a LUN has been mapped to a host, you must restart the host for the configuration to take effect after you modify **Host Access Mode** or **Preferred Path for HyperMetro**. If you map the LUN for the first time, restart is not needed.

Host Configuration

Install, configure, and use UltraPath by following instructions in the *OceanStor UltraPath for Linux User Guide*.

NOTE

To obtain the document, log in to Huawei's technical support website (**https://support.huawei.com/enterprise/**), enter **UltraPath** in the search box, and select the associated path to the documentation page. Then find and download the desired document. To download software, click the **Software Download** tab and find the desired software.

Step 1 Set the HyperMetro working mode.

You can set the HyperMetro working mode using either of the following methods:

Method 1: Run the **upadmin_plus set hypermetro workingmode=auto** command to configure UltraPath to automatically adapt the HyperMetro working mode. This setting enables UltraPath to periodically query the host access mode configured on HyperMetro storage systems and adapt its HyperMetro working mode according to the host access mode.

Method 2: Run the following command to set UltraPath to work in a fixed HyperMetro working mode:

Command	Example
set hypermetro workingmode=*[priority | balance]* **primary_array_id=***ID*	**upadmin_plus set hypermetro workingmode=priority primary_array_id=0**

The following table describes the parameters in the command.

Parameter	Description	Default Value
workingmode	HyperMetro working mode. • **priority**: local preferred mode • **balance**: load balancing mode	priority **priority** is recommended. **balance** is applicable when two active-active data centers are in the same equipment room or on the same floor.
primary_array_id	ID of the preferred storage system. The ID is allocated by UltraPath. Select the storage system that resides in the same data center as the application host. Run the **upadmin show array** command to obtain the storage system ID. **NOTE** • In **priority** mode, the value of the parameter indicates the storage system to which I/Os are preferentially delivered. • In **balance** mode, the value of the parameter indicates the storage system where the first slice section resides.	None **NOTE** Mapping relationship between application hosts and storage systems: • Storage system A is the preferred system for all application hosts in data center A. • Storage system B is the preferred system for all application hosts in data center B.

NOTICE

If you set UltraPath to automatically adapt the HyperMetro working mode, ensure that the host access mode on the storage system is consistent with that on the physical network.

Step 2 Configure the load balancing policy.

If HyperMetro works in load balancing mode, you can run the **upadmin_plus set hypermetro loadbalancemode=***[split-size | round-robin]* command to configure the load balancing policy. The following table describes the parameters in the command.

Parameter	Description	Default Value
loadbalancemode	Load balancing policy for HyperMetro systems. ● **split-size**: slicing mode across storage systems. In this mode, UltraPath delivers I/Os to a specific storage system based on the start addresses of I/Os, slice size, and preferred storage system. For example, if the slice size is 128 MB, the I/Os whose start addresses range from 0 to 128 MB (excluding 128 MB) are preferentially delivered to the preferred storage system and the I/Os whose start addresses range from 128 MB to 256 MB (excluding 256 MB) are delivered to the non-preferred storage system. The default slice size is 128 MB. You can run the **upadm set hypermetro split_size** command to change it. ● **round-robin**: round-robin mode across storage systems. In this mode, UltraPath selects two storage systems in turn to deliver I/Os.	split-size

----End

6.2.1 UltraPath

6.2.1.1 Storage System Configuration

If UltraPath is used in HyperMetro scenarios, retain the default settings of the initiator and configure **Host Access Mode** and **Preferred Path for HyperMetro** as required. Table 6-1 lists the detailed settings.

Table 6-2 Storage configurations for interconnection with a Linux system

OS	Host Configuration on the Storage System				Configuration Description
	HyperMetro Mode	Storage	Host Access Mode	Preferred Path for HyperMetro	

Redhat/ CentOS/ Kylin/Suse	Load balancing mode	Local storage system	Balanced	N/A	The host uses all paths of a disk with equal priority.
		Remote storage system	Balanced	N/A	
	Local preferred	Local storage system	Asymmetric	Yes	The host considers the paths from the local storage system as preferred paths, and those from the remote storage system as non-preferred paths.
		Remote storage system	Asymmetric	No	

NOTICE

1. For details about the supported Linux versions, see the Huawei Storage Interoperability Navigator.
2. If a LUN has been mapped to a host, you must restart the host for the configuration to take effect after you modify the **Host Access Mode** or **Preferred Path for HyperMetro**. If you configure the host for the first time, restart is not needed.

6.2.1.2 Host Configuration

Install, configure, and use UltraPath by following instructions in the *OceanStor UltraPath-NVMe xx for Linux User Guide*.

NOTE

To obtain the document, log in to Huawei's technical support website (https://support.huawei.com/enterprise/), enter **UltraPath** in the search box, and select the associated path to the documentation page. Then find and download the desired document. To download software, click the **Software Download** tab and find the desired software.

Step 1 Set the HyperMetro working mode.

You can set the HyperMetro working mode using either of the following methods:

Method 1: Set UltraPath to automatically configure its HyperMetro working mode. You can run the **upadmin_plus set hypermetro workingmode=auto** command to allow UltraPath to periodically identify the host access mode on the storage systems and adjust its HyperMetro working mode accordingly.

Method 2: Run the following command to set UltraPath to work in a fixed HyperMetro working mode:

Command	Example
set hypermetro workingmode=*[priority | balance]* **primary_array_id=***ID*	**upadmin_plus set hypermetro workingmode=priority primary_array_id=0**

The following table describes the parameters in the command.

Parameter	Description	Default Value
workingmode	HyperMetro working mode ● **priority**: local preferred mode ● **balance**: load balancing mode	priority **priority** is recommended. **balance** is applicable when two active-active data centers are in the same equipment room or on the same floor.
primary_array_id	ID of the preferred storage system. The ID is allocated by UltraPath. Select the storage system that resides in the same data center as the application host. Run the **upadm show array** command to obtain the storage system ID. Description ● In **priority** mode, the value of the parameter indicates the storage system to which I/Os are preferentially delivered. ● In **balance** mode, the value of the parameter indicates the storage system where the first slice section resides.	None Description Mapping relationship between application hosts and storage systems: ● Storage system A is the preferred system for all application hosts in data center A. ● Storage system B is the preferred system for all application hosts in data center B.

⚠ DANGER

If you set UltraPath to automatically adapt the HyperMetro working mode, ensure that host access mode on the storage system are consistent with those on the physical network.

Step 2 Configure the load balancing policy.

If HyperMetro works in load balancing mode, you can run the **upadmin_plus set hypermetro loadbalancemode=***[split-size* | *round-robin]* command to configure the load balancing policy. The following table describes the parameters in the command.

Parameter	Description	Default Value
loadbalancemode	Load balancing mode for HyperMetro systems. ● **split-size**: slicing mode across storage systems. ● In this mode, UltraPath delivers I/Os to a specific storage system based on the start addresses of I/Os, slice size, and preferred storage system. For example, if the slice size is 128 MB, the I/Os whose start addresses range from 0 to 128 MB (excluding 128 MB) are preferentially delivered to the preferred storage system and the I/Os whose start addresses range from 128 MB to 256 MB (excluding 256 MB) are delivered to the non-preferred storage system. The default slice size is 128 MB. You can run the **upadmin set hypermetro split_size** command to change it. ● **round-robin**: round-robin mode across storage systems. ● In this mode, UltraPath selects two storage systems in turn to deliver I/Os.	split-size

----End

6.2.1.3 Verification

Run the **upadmin_plus show vlun type=hypermetro** command to verify that the configuration has taken effect. The following is an example:

```
[root@localhost ~]# upadmin_plus show vlun type=hypermetro
-----------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------
```

```
 Vlun ID     Disk          Name                  Lun WWN              Status  Capacity  Ctrl(Own/Work)   Array
Name      Dev Lun ID  No. of Paths(Available/Total)
   0     ultrapathak  50linuxLun   7100b5d6710047d024a52cda0000003e  Normal  50.00GB       --/--
Huawei.Storage1      62                8/8
   0     ultrapathak  50linuxLun   7100b5d6710047d024a52cda0000003e  Normal  50.00GB       --/--
Huawei.Storage2      62                8/8
   1     ultrapathal  100linuxLun  7100b5d6710047d324a52cbe0000003f  Normal  100.00GB      --/--
Huawei.Storage1      63                8/8
   2     ultrapathal  100linuxLun  7100b5d6710047d324a52cbe0000003f  Normal  100.00GB      --/--
Huawei.Storage2      63                8/8
----------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------
```

In the command output, the two mapped LUNs have been taken over by the native multipathing software of the system and all LUNs are in the normal state, indicating that the configuration has taken effect.

6.2.2 OS Native Device Mapper

6.2.2.1 Storage System Configuration

If the OS native multipathing software Device Mapper is used, retain the default settings of the initiator and configure **Host Access Mode** and **Preferred Path for HyperMetro** as required. Table 6-2 lists the detailed settings.

Table 6-3 Storage configurations for interconnection with a Linux system

OS	Host Configuration on the Storage System				Configuration Description
	HyperMetro Mode	Storage	Host Access Mode	Preferred Path for HyperMetro	
Redhat/CentOS/Kylin/Suse	Load balancing mode	Local storage system	Balanced	N/A	The host uses all paths of a disk with equal priority.
		Remote storage system	Balanced	N/A	
	Local preferred	Local storage system	Asymmetric	Yes	The host considers the paths from the local storage system as preferred paths, and those from the remote storage system as non-preferred paths.
		Remote storage system	Asymmetric	No	

⚠ CAUTION

1. For details about the supported Linux versions, see the Huawei Storage Interoperability Navigator.
2. If a LUN has been mapped to a host, you must restart the host for the configuration to take effect after you modify the **Host Access Mode** or **Preferred Path for HyperMetro**. If you configure the host for the first time, restart is not needed.

6.2.2.2 Host Configuration

Installing Multipathing Software

Generally, the built-in multipathing software package of the Linux operating system is an RPM package starting with **device-mapper-multipath**. You can run the following command to check whether the corresponding software package has been installed:

```
[root@localhost ~]# rpm -qa | grep multipath
device-mapper-multipath-0.7.8-7.el8.x86_64
device-mapper-multipath-libs-0.7.8-7.el8.x86_64
```

If the query result is empty, the software package is not installed. You can obtain the software package from the system image and run the **rpm** command to install the software package, or use an operating system tool (such as YUM or YaST) to install the software package.

Disabling Native NVMe Multipath

When the native multipathing software of the OS is used, you must disable Native NVMe Multipath. You can run the following command to check the Native NVMe Multipath status:

```
[root@localhost ~]# cat /sys/module/nvme_core/parameters/multipath
Y
```

If **No such file or directory** or **N** is displayed in the command output, you do not need to disable Native NVMe Multipath.

If **Y** is displayed in the command output, Native NVMe Multipath is enabled. Perform the following operations to disable it:

Step 1 Add **nvme_core.multipath=N** to the **GRUB_CMDLINE_LINUX_DEFAULT** entry in the **/etc/default/grub** file to disable Native NVMe Multipath.

```
[root@localhost ~]# cat /etc/default/grub | grep multipath
GRUB_CMDLINE_LINUX_DEFAULT="splash=silent resume=/dev/disk/by-path/pci-0000:00:10.0-scsi-0:0:0:0-part4 mitigations=auto quiet nvme_core.multipath=Ncrashkernel=180M,high"
```

Step 2 If the **/sys/firmware/efi** directory exists in the OS, the OS uses the EFI boot mode. In this case, run the **grub2-mkconfig -o /boot/efi/EFI/centos/grub.cfg** command to generate a new grub configuration file. (**centos** indicates the OS vendor and varies depending on the OS.) If the traditional boot mode is used, run the **grub2-mkconfig -o /boot/grub2/grub.cfg** command to regenerate the grub configuration file.

Step 3 Run the **reboot** command to restart the host.

Step 4 Run the **cat /sys/module/nvme_core/parameters/multipath** command to check whether the modification takes effect.

```
[root@localhost ~]# cat /sys/module/nvme_core/parameters/multipath
N
```

----End

Configuring the Multipath Configuration File

DM-Multipath's most important configuration file is **/etc/multipath.conf**.

Some operating systems have such a file by default. You can copy the **multipath.conf** or **multipath.conf.synthetic** file to the **/etc** directory to obtain the template. If the file does not exist, manually create the **/etc/multipath.conf** file.

```
[root@localhost ~]# cp /usr/share/doc/device-mapper-multipath-0.4.9/multipath.conf /etc/multipath.conf
```

Redhat/CentOS 8.x:

If the **Host Access Mode** is set to **Load balancing** on the storage system, adding the following content to the **/etc/multipath.conf** file is recommended:

```
defaults {
                polling_interval            1
                user_friendly_names          yes
}

blacklist_exceptions {
        property "(ID_WWN|SCSI_IDENT_.*|ID_SERIAL|DEVTYPE)"
        devnode "nvme*"
}

devices {
      device {
                vendor                       "NVME"
                product                      "Huawei-XSG1"
                uid_attribute                 "ID_WWN"
                no_path_retry                  12
                rr_min_io_rq                  1
                prio                         "const"
                path_grouping_policy        multibus
                path_checker                  "directio"
                failback                    immediate
                fast_io_fail_tmo            0
                retain_attached_hw_handler  "no"
}
}
```

If the **Host Access Mode** is set to **Asymmetric** on the storage system, adding the following content to the **/etc/multipath.conf** file is recommended:

```
defaults {
                polling_interval            1
                user_friendly_names          yes
}

blacklist_exceptions {
        property "(ID_WWN|SCSI_IDENT_.*|ID_SERIAL|DEVTYPE)"
        devnode "nvme*"
}

devices {
      device {
```

```
                vendor                      "NVME"
                product                     "Huawei-XSG1"
                uid_attribute               "ID_WWN"
                no_path_retry               12
                rr_min_io_rq                1
                prio                        "alua"
                path_grouping_policy        group_by_prio
                path_checker                "directio"
                failback                    immediate
                fast_io_fail_tmo            0
                retain_attached_hw_handler  "no"
}
}
```

Starting the Multipathing Software

Run the following command to start the multipathing process:

```
systemctl start multipathd.service
```

Restarting the Multipathing Software

If **/etc/multipath.conf** is modified after the multipathing service has been started, you must restart or reload the multipathing service for the modification to take effect. The following is an example:

```
systemctl restart multipathd.service
systemctl reload multipathd.service
```

Setting the Multipathing Software to Run at System Startup

Run the following command to set the multipathing software to run at system startup:

```
systemctl enable multipathd.service
```

6.2.2.3 Verification

Verifying the Load Balancing Mode

Run the **multipath -ll** command to verify that the configuration has taken effect. In load balancing mode, all paths are active. The following is an example:

```
[root@localhost ~]# multipath -ll
mpathd (eui.7100b5d6710031a624a52c5400000000) dm-3 NVME,Huawei-XSG1
size=50G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='service-time 0' prio=50 status=active
  |- 9:787:1:21  nvme9n1  259:7  active ready running
  |- 3:261:1:21  nvme3n1  259:4  active ready running
  |- 0:781:1:21  nvme0n1  259:0  active ready running
  |- 2:263:1:21  nvme2n1  259:1  active ready running
  |- 11:259:1:21 nvme11n1 259:3  active ready running
  |- 1:785:1:21  nvme1n1  259:6  active ready running
  |- 8:783:1:21  nvme8n1  259:2  active ready running
  |- 10:257:1:21 nvme10n1 259:5  active ready running
  |- 6:257:3:21  nvme6n3  259:44 active ready running
  |- 13:3:3:21   nvme13n3 259:48 active ready running
  |- 12:1:3:21   nvme12n3 259:50 active ready running
  |- 7:261:3:21  nvme7n3  259:53 active ready running
  |- 14:263:3:21 nvme14n3 259:51 active ready running
  |- 4:5:3:21    nvme4n3  259:55 active ready running
  |- 5:7:3:21    nvme5n3  259:58 active ready running
  `- 15:259:3:21 nvme15n3 259:54 active ready running
```

```
mpathe (eui.710037421701f64516212c7400000025) dm-4 NVME,Huawei-XSG1
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='service-time 0' prio=50 status=active
  |- 8:783:2:22  nvme8n2  259:10 active ready running
  |- 3:261:2:22  nvme3n2  259:11 active ready running
  |- 10:257:2:22 nvme10n2 259:12 active ready running
  |- 0:781:2:22  nvme0n2  259:9  active ready running
  |- 11:259:2:22 nvme11n2 259:8  active ready running
  |- 2:263:2:22  nvme2n2  259:13 active ready running
  |- 9:787:2:22  nvme9n2  259:15 active ready running
  |- 1:785:2:22  nvme1n2  259:19 active ready running
  |- 12:1:4:22   nvme12n4 259:57 active ready running
  |- 13:3:4:22   nvme13n4 259:56 active ready running
  |- 14:263:4:22 nvme14n4 259:59 active ready running
  |- 15:259:4:22 nvme15n4 259:60 active ready running
  |- 6:257:4:22  nvme6n4  259:52 active ready running
  |- 7:261:4:22  nvme7n4  259:61 active ready running
  |- 5:7:4:22    nvme5n4  259:63 active ready running
  `- 4:5:4:22    nvme4n4  259:62 active ready running
```

Verifying the Local Preferred Mode

Run the **multipath -ll** command to verify that the configuration has taken effect. In local preferred mode, **status=active** corresponds to the preferred paths on the local storage system, and **status=enabled** corresponds to the non-preferred paths on the remote storage system. The following command output indicates that the configuration has taken effect. Generally, the **prio** value of preferred paths is **50** and that of non-preferred paths is **10** in Linux systems. The following is an example.

```
[root@localhost ~]# multipath -ll
mpathb (eui.7100b5d67100656f24a52cb900000014) dm-3 NVME,Huawei-XSG1
size=50G features='1 queue_if_no_path' hwhandler='0' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 9:779:1:21  nvme9n1  259:4  active ready running
| |- 11:527:1:21 nvme11n1 259:3  active ready running
| |- 0:1:1:21    nvme0n1  259:5  active ready running
| |- 8:9:1:21    nvme8n1  259:1  active ready running
| |- 2:5:1:21    nvme2n1  259:0  active ready running
| |- 3:519:1:21  nvme3n1  259:2  active ready running
| |- 10:13:1:21  nvme10n1 259:6  active ready running
| `- 1:771:1:21  nvme1n1  259:7  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 5:3:3:21    nvme5n3  259:45 active ready running
  |- 13:7:3:21   nvme13n3 259:50 active ready running
  |- 6:257:3:21  nvme6n3  259:47 active ready running
  |- 12:5:3:21   nvme12n3 259:44 active ready running
  |- 4:1:3:21    nvme4n3  259:53 active ready running
  |- 14:261:3:21 nvme14n3 259:58 active ready running
  |- 7:259:3:21  nvme7n3  259:59 active ready running
  `- 15:263:3:21 nvme15n3 259:54 active ready running
mpathc (eui.7100b5d67100657224a52ce000000015) dm-4 NVME,Huawei-XSG1
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 11:527:2:22 nvme11n2 259:9  active ready running
| |- 2:5:2:22    nvme2n2  259:10 active ready running
| |- 0:1:2:22    nvme0n2  259:13 active ready running
| |- 10:13:2:22  nvme10n2 259:18 active ready running
| |- 3:519:2:22  nvme3n2  259:8  active ready running
| |- 1:771:2:22  nvme1n2  259:16 active ready running
| |- 9:779:2:22  nvme9n2  259:15 active ready running
| `- 8:9:2:22    nvme8n2  259:11 active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 12:5:4:22   nvme12n4 259:52 active ready running
  |- 5:3:4:22    nvme5n4  259:56 active ready running
  |- 6:257:4:22  nvme6n4  259:55 active ready running
  |- 13:7:4:22   nvme13n4 259:57 active ready running
```

```
 |- 15:263:4:22 nvme15n4 259:60 active ready running
 |- 14:261:4:22 nvme14n4 259:62 active ready running
 |- 7:259:4:22  nvme7n4  259:61 active ready running
 `- 4:1:4:22    nvme4n4  259:63 active ready running
```

7 FAQs

7.1 The Host Runs RHEL 8.0 or SLES 15 SP1 and Uses the QLE2742 HBA. After the Storage Port Protocol Is Changed to FC-NVMe, Execution of the nvme list Command Is Suspended

Symptom

After the protocol of the storage port is changed from FC-SCSI to FC-NVMe, the link is up and the host initiator is online on the storage system. However, when the **nvme list** command is run on the host to query LUN information, the command is suspended. The **multipath -ll** command output shows that the path is failed.

Solution

1. Set **fast_io_fail_tmo** to **0**.

```
echo 0 > /sys/class/fc_remote_ports/rport*/fast_io_fail_tmo
```

NOTE

This parameter automatically changes to the default value (5s) after the port protocol is changed back to FC-SCSI. You must modify this parameter again if you use FC-NVMe again.

2. Intermittently disconnect the host port or restart the host.

   To intermittently disconnect the host port, run the following command:

```
echo 1 > /sys/class/fc_host/host13/issue_lip
```

7.2 The Host Runs RHEL 8.0 or SLES 15 SP1 and Uses the QLE2742 HBA. After the Storage Port Protocol Is Changed to FC-NVMe, There Is a Possibility That Establishing the Connection Fails

Symptom

After the protocol of the storage port is changed from FC-SCSI to FC-NVMe, there is a possibility that the connection fails to be set up and the initiator is offline.

Solution

Reload the QLogic driver or restart the host when no other services are running on the host.

To reload the QLogic driver, run the following command:

```
rmmod qla2xxx
modprobe qla2xxx
```

7.3 How Can I Modify LUN Mappings for NVMe-oF Services in Red Hat or SUSE?

Symptom

After a LUN is replaced (the new LUN uses the same host LUN ID as the original LUN), the information about the new LUN cannot be updated.

Solution

To prevent this problem, you must follow the correct sequence when changing the LUN mapping.

Step 1 Before removing or changing LUN mapping on the storage system, stop all services running on the disk mapped by the LUN.

Step 2 On DeviceManager, unmap the LUN.

Step 3 Run the **upRescan_plus** command to rescan for disks.

Step 4 On DeviceManager, map a new LUN to the host.

Step 5 Run the **upRescan_plus** command to rescan for disks.

Step 6 Contact the administrator to restart the services.

----End

7.4 How Can I Modify the Granularity for Reclaiming Thin LUNs on a Host?

Currently, the maximum space reclamation granularity of Huawei storage systems is 64 MB. You must set the reclamation granularity on the host to less than or equal to 64 MB. The following uses 64 MB as an example.

UltraPath

UltraPath automatically sets the space reclamation granularity of all virtual disks to 64 MB.

NVMe Native Multipathing

If the NVMe native multipathing software is used and the NOF INI software of OceanStor NOF Enabler is installed on the host, NOF INI automatically sets the space reclamation granularity of all physical disks to 64 MB. Manual change is not needed.

If NOF INI is not installed on the host, you must manually change the space reclamation granularity for all physical disks. (The change loses efficacy after the host is restarted.) For example, to change the space reclamation granularity of virtual disk **nvme0n1** to 64 MB, which includes two physical disks **nvme0c256n1** and **nvme0c642n1**, run the following commands:

```
[root@localhost ~]# echo 67108864 > /sys/class/block/nvme0c256n1/queue/discard_max_bytes
[root@localhost ~]# echo 67108864 > /sys/class/block/nvme0c642n1/queue/discard_max_bytes
[root@localhost ~]# cat /sys/class/block/nvme0c256n1/queue/discard_max_bytes
67108864
[root@localhost ~]# cat /sys/class/block/nvme0c642n1/queue/discard_max_bytes
67108864
```

OS Native Multipathing Software

If the OS native multipathing software is used, you must manually change the space reclamation granularity for all virtual disks generated by DM-Multipath, no matter whether NOF INI of OceanStor NOF Enabler is installed. (The change loses efficacy after the host is restarted.) For example, to change the space reclamation granularity of virtual disk **dm-1** to 64 MB, run the following commands:

```
[root@localhost ~]# echo 67108864 > /sys/class/block/dm-1/queue/discard_max_bytes
[root@localhost ~]# cat /sys/class/block/dm-1/queue/discard_max_bytes
67108864
```

7.5 How Can I Manually Disconnect the NVMe Link?

If you want to disconnect all NVMe links during routine maintenance, run the following command:

```
[root@localhost ~]# nvme disconnect-all
```

7.6 How Can I Configure PFC on Cisco Nexus 9000 Series Switches?

For details about configuring PFC on Cisco Nexus 9000 series switches, see "Configuring Priority Flow Control" on Cisco's official website. The website link is as follows: https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/qos/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_Quality_of_Service_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_Quality_of_Service_Configuration_Guide_7x_chapter_01011.html

8 Acronyms and Abbreviations

C	
CLI	Command Line Interface
D	
DM-Multipath	Device Mapper-Multipath
E	
Ext2	Second Extended File System
Ext3	Third Extended File System
Ext4	Fourth Extended File System
F	
FC	Fibre Channel
H	
HBA	Host Bus Adapter
L	
LUN	Logical Unit Number
LV	Logic Volume
LVM	Logical Volume Manager
M	
MB	MByte
N	
NVMe	Non-Volatile Memory Express

R	
RAID	Redundant Array of Independent Disks
RHEL	Red Hat Enterprise Linux
RDMA	Remote Direct Memory Access
RoCE	RDMA over Converged Ethernet
S	
SAN	Storage Area Network
P	
PE	Physical Extent
PV	Physical Volume
V	
VG	Volume Group
W	
WWN	World Wide Name

9 Appendix A Volume Management

The most widely applied volume management software in Linux hosts is the built-in Logical Volume Manager (LVM).

This chapter details the LVM.

9.1 Overview

LVM can combine several disks (physical volumes) into a volume group and divide the volume group into logical volumes (LVM partitions).

LVM provides the following functions:

- Creating logical volumes across multiple disks
- Creating logical volumes on one disk
- Expanding and compressing logical volumes on demand

9.2 LVM Installation

By default, LVM is installed together with the host operating system. LVM requires no extra configuration.

9.3 Modifying the LVM Configuration File

When OS native multipathing software is used, LVM is configured using the multipathing drive letter such as **/dev/mapper/mpathX** instead of the NVMe drive letter such as **/dev/nvmeXnX**. To prevent exceptions in volume management functions caused by drive letter conflicts, add the following content to the **filter** field in the LVM configuration file **/etc/lvm/lvm.conf**:

```
filter = ["a|/dev/mapper/mpath.*|","a|/dev/sda.*|", "r|.*|"]
```

The preceding configuration enables LVM to accept and preferentially use the multipathing drive letters, and ignore other drive letters. **sda** is the drive letter of the system's local disk, which can be modified based on actual configurations.

When Huawei UltraPath is used, configure the **filter** field by following instructions in the UltraPath-NVMe user guide specific to your product version.

9.4 Common Configuration Commands

NOTICE

The drive letters, volume group names, and logical volume names in the following commands are examples. Change them according to the actual situation.

Creating a Physical Volume

Step 1 Run the **pvcreate** command to create physical volumes.

```
[root@root ~]# pvcreate /dev/mapper/mpatha
  Physical volume "/dev/mapper/mpatha" successfully created
[root@root ~]# pvcreate /dev/mapper/mpathb
  Physical volume "/dev/mapper/mpathb" successfully created
```

Step 2 Run the **pvdisplay -v** command to verify the physical volume creation.

```
[root@root ~]# pvdisplay -v
    Scanning for physical volume names
    Wiping cache of LVM-capable devices
  --- Physical volume ---
  PV Name               /dev/sda2
  VG Name               VolGroup00
  PV Size               557.65 GB / not usable 21.17 MB
  Allocatable           yes (but full)
  PE Size (KByte)       32768
  Total PE              17844
  Free PE               0
  Allocated PE          17844
  PV UUID               KyucjQ-9zte-1Zyr-0sZ0-Xxzt-HVjZ-2vQp8B

  "/dev/mapper/mpatha" is a new physical volume of "1.53 GB"
  --- NEW Physical volume ---
  PV Name               /dev/mapper/mpatha
  VG Name
  PV Size               1.53 GB
  Allocatable           NO
  PE Size (KByte)       0
  Total PE              0
  Free PE               0
  Allocated PE          0
  PV UUID               g60zN0-3sYn-qPbd-7y0M-dGfZ-hVs7-763Ywo

  "/dev/mapper/mpathb" is a new physical volume of "1.53 GB"
  --- NEW Physical volume ---
  PV Name               /dev/mapper/mpathb
  VG Name
  PV Size               1.53 GB
  Allocatable           NO
  PE Size (KByte)       0
  Total PE              0
  Free PE               0
```

```
 Allocated PE          0
 PV UUID               5UhmY2-fS4p-gdCo-OOgZ-nOa9-AV3H-LkvrNc
```

----End

Changing the Size of a Physical Volume

Run the **pvresize** command to change the size of a physical volume. The command syntax is as follows:

pvresize –setphysicalvolumesize *capacity size* (unit: m or g) *device name*

In the following example, the size of a physical volume is changed from 1.53 GB to 300 MB.

```
[root@root ~]# pvscan
 PV /dev/sda2   VG VolGroup00      lvm2 [557.62 GB / 0    free]
 PV /dev/mapper/mpatha             lvm2 [1.53 GB]
 PV /dev/mapper/mpathb             lvm2 [1.53 GB]
 Total: 3 [560.69 GB] / in use: 1 [557.62 GB] / in no VG: 2 [3.06 GB]
 [root@root ~]# pvresize --setphysicalvolumesize 300 /dev/mapper/mpatha
 Physical volume "/dev/mapper/mpatha" changed
 1 physical volume(s) resized / 0 physical volume(s) not resized
[root@root ~]# pvscan
 PV /dev/sda2   VG VolGroup00      lvm2 [557.62 GB / 0    free]
 PV /dev/mapper/mpatha             lvm2 [300.00 MB]
 PV /dev/mapper/mpathb             lvm2 [1.53 GB]
 Total: 3 [559.45 GB] / in use: 1 [557.62 GB] / in no VG: 2 [1.83 GB]
```

Creating a Volume Group

Run the **vgcreate** command to create a volume group:

```
[root@root ~]# vgcreate vg0 /dev/mapper/mpatha /dev/mapper/mpathb
 Volume group "vg0" successfully created
```

Expanding a Volume Group

Run the following command to expand a volume group:

```
vgextend vgname pvname
```

The following is an example:

```
[root@root ~]# vgdisplay -v /dev/vg0
   Using volume group(s) on command line
   Finding volume group "vg0"
 --- Volume group ---
 VG Name               vg0
 System ID
 Format                lvm2
 Metadata Areas        2
 Metadata Sequence No  1
 VG Access             read/write
 VG Status             resizable
 MAX LV                0
 Cur LV                0
 Open LV               0
 Max PV                0
 Cur PV                2
 Act PV                2
 VG Size               1.82 GB
 PE Size               4.00 MB
 Total PE              466
 Alloc PE / Size       0 / 0
```

```
  Free  PE / Size       466 / 1.82 GB
  VG UUID               ARkbdL-9ID6-5HCy-DSQG-Aj5z-dQap-9VkM5X
  --- Physical volumes ---
  PV Name               /dev/mapper/mpatha
  PV UUID               g60zN0-3sYn-qPbd-7y0M-dGfZ-hVs7-763Ywo
  PV Status             allocatable
  Total PE / Free PE    74 / 74
  PV Name               /dev/mapper/mpathb
  PV UUID               5UhmY2-fS4p-gdCo-OOgZ-nOa9-AV3H-LkvrNc
  PV Status             allocatable
  Total PE / Free PE    392 / 392
[root@root ~]# vgextend /dev/vg0 /dev/mapper/mpathc
  Volume group "vg0" successfully extended
[root@root ~]# vgdisplay -v /dev/vg0
    Using volume group(s) on command line
    Finding volume group "vg0"
  --- Volume group ---
  VG Name               vg0
  System ID
  Format                lvm2
  Metadata Areas        3
  Metadata Sequence No  2
  VG Access             read/write
  VG Status             resizable
  MAX LV                0
  Cur LV                0
  Open LV               0
  Max PV                0
  Cur PV                3
  Act PV                3
  VG Size               3.35 GB
  PE Size               4.00 MB
  Total PE              858
  Alloc PE / Size       0 / 0
  Free  PE / Size       858 / 3.35 GB
  VG UUID               ARkbdL-9ID6-5HCy-DSQG-Aj5z-dQap-9VkM5X
  --- Physical volumes ---
  PV Name               /dev/mapper/mpatha
  PV UUID               g60zN0-3sYn-qPbd-7y0M-dGfZ-hVs7-763Ywo
  PV Status             allocatable
  Total PE / Free PE    74 / 74
  PV Name               /dev/mapper/mpathb
  PV UUID               5UhmY2-fS4p-gdCo-OOgZ-nOa9-AV3H-LkvrNc
  PV Status             allocatable
  Total PE / Free PE    392 / 392
  PV Name               /dev/mapper/mpathc
  PV UUID               iF5Att-fVIj-9dOy-5055-rJlq-pOrS-aW8g2P
  PV Status             allocatable
  Total PE / Free PE    392 / 392
```

In this example, volume group **/dev/vg0** originally contains physical volumes **/dev/mapper/mpatha** and **/dev/mapper/mpathb**. After the command is run, **/dev/mapper/mpathc** is added to this volume group.

Creating a Logical Volume

Step 1 Run the **lvcreate** command to create a logical volume. The following is an example:

```
[root@root ~]# lvcreate -L 10m -n lv0 vg0
  Rounding up size to full physical extent 12.00 MB
  Logical volume "lv0" created
```

The parameters in this output are described as follows:

- -L: indicates the size of a logical volume, expressed in MB.
- -n: indicates the name of a logical volume.

Step 2 View and confirm that the information about the newly created logical volume is correct.

```
[root@root ~]# vgdisplay -v vg0
    Using volume group(s) on command line
    Finding volume group "vg0"
  --- Volume group ---
  VG Name               vg0
  System ID
  Format                lvm2
  Metadata Areas        3
  Metadata Sequence No  3
  VG Access             read/write
  VG Status             resizable
  MAX LV                0
  Cur LV                1
  Open LV               0
  Max PV                0
  Cur PV                3
  Act PV                3
  VG Size               3.35 GB
  PE Size               4.00 MB
  Total PE              858
  Alloc PE / Size       3 / 12.00 MB
  Free  PE / Size       855 / 3.34 GB
  VG UUID               ARkbdL-9ID6-5HCy-DSQG-Aj5z-dQap-9VkM5X

  --- Logical volume ---
  LV Name                /dev/vg0/lv0
  VG Name                vg0
  LV UUID                H6uskM-6clf-NVh2-KMiO-1Gk2-0iBz-nXOav2
  LV Write Access        read/write
  LV Status              available
  # open                 0
  LV Size                12.00 MB
  Current LE             3
  Segments               1
  Allocation             inherit
  Read ahead sectors     auto
  - currently set to     256
  Block device           253:2

  --- Physical volumes ---
  PV Name               /dev/mapper/mpatha
  PV UUID               g60zN0-3sYn-qPbd-7y0M-dGfZ-hVs7-763Ywo
  PV Status             allocatable
  Total PE / Free PE    74 / 74

  PV Name               /dev/mapper/mpathb
  PV UUID               5UhmY2-fS4p-gdCo-OOgZ-nOa9-AV3H-LkvrNc
  PV Status             allocatable
  Total PE / Free PE    392 / 389

  PV Name               /dev/mapper/mpathc
  PV UUID               iF5Att-fVIj-9dOy-5055-rJlq-pOrS-aW8g2P
  PV Status             allocatable
  Total PE / Free PE    392 / 392
[root@root ~]# lvdisplay -v /dev/vg0/lv0
    Using logical volume(s) on command line
  --- Logical volume ---
  LV Name                /dev/vg0/lv0
  VG Name                vg0
  LV UUID                H6uskM-6clf-NVh2-KMiO-1Gk2-0iBz-nXOav2
  LV Write Access        read/write
  LV Status              available
  # open                 0
  LV Size                12.00 MB
  Current LE             3
  Segments               1
  Allocation             inherit
```

```
 Read ahead sectors     auto
 - currently set to     256
 Block device           253:2
```

----End

Creating a File System

Step 1 Run the **mkfs.xx** command to create a file system. In the following example, an ext3 file system is created.

```
[root@root ~]# mkfs.ext3 /dev/vg0/rlv0
mke2fs 1.39 (29-May-2006)
Filesystem label=
OS type: Linux
Block size=1024 (log=0)
Fragment size=1024 (log=0)
3072 inodes, 12288 blocks
614 blocks (5.00%) reserved for the super user
First data block=1
Maximum filesystem blocks=12582912
2 block groups
8192 blocks per group, 8192 fragments per group
1536 inodes per group
Superblock backups stored on blocks:
    8193

Writing inode tables: done
Creating journal (1024 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 20 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

Step 2 Create a mount point and mount the logical volume.

```
[root@root ~]# mkdir /test/mnt1
[root@root ~]# mount /dev/vg0/lv0 /test/mnt1/
Display the mounting information.
[root@root ~]# df -l
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/mapper/VolGroup00-LogVol00
                     548527904 3105828 517108888 1% /
/dev/sda1               101086     15667     80200  17% /boot
tmpfs                  8137904         0 8137904 0% /dev/shm
/dev/mapper/vg0-lv0      11895      1138     10143  11% /test/mnt1
```

The output shows that the logical volume is mounted correctly and can be used for subsequent data read and write.

Step 3 (Optional) You can run the following command to unmount the logical volume:

```
[root@root ~]# umount /dev/vg0/lv0
[root@root ~]# df -l
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/mapper/VolGroup00-LogVol00
                     548527904 3105828 517108888 1% /
/dev/sda1               101086     15667     80200  17% /boot
tmpfs                  8137904         0 8137904 0% /dev/shm
```

----End

Expanding a Logical Volume

Run the **lvextend** command to expand a logical volume. The command syntax is as follows:

```
lvextend  -L +Target capacity  Logical volume path
```

The following is an example:

```
[root@root ~]# lvscan
  ACTIVE            '/dev/vg0/lv0' [12.00 MB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol00' [540.03 GB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol01' [17.59 GB] inherit
[root@root ~]# pvscan
  PV /dev/mapper/mpatha    VG vg0          lvm2 [296.00 MB / 296.00 MB free]
  PV /dev/mapper/mpathb    VG vg0          lvm2 [1.53 GB / 1.52 GB free]
  PV /dev/mapper/mpathc    VG vg0          lvm2 [1.53 GB / 1.53 GB free]
  PV /dev/sda2             VG VolGroup00   lvm2 [557.62 GB / 0    free]
  Total: 4 [560.98 GB] / in use: 4 [560.98 GB] / in no VG: 0 [0   ]
[root@root ~]# lvextend -L +100m /dev/vg0/lv0
  Extending logical volume lv0 to 112.00 MB
  Logical volume lv0 successfully resized
[root@root ~]# lvscan
  ACTIVE            '/dev/vg0/lv0' [112.00 MB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol00' [540.03 GB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol01' [17.59 GB] inherit
[root@root ~]# pvscan
  PV /dev/mapper/mpatha    VG vg0          lvm2 [296.00 MB / 296.00 MB free]
  PV /dev/mapper/mpathb    VG vg0          lvm2 [1.53 GB / 1.42 GB free]
  PV /dev/mapper/mpathc    VG vg0          lvm2 [1.53 GB / 1.53 GB free]
  PV /dev/sda2             VG VolGroup00   lvm2 [557.62 GB / 0    free]
  Total: 4 [560.98 GB] / in use: 4 [560.98 GB] / in no VG: 0 [0   ]
```

The output shows that the logical volume capacity is expanded.

Downsizing a Logical Volume

Run the **lvreduce** command to downsize a logical volume. The command syntax is as follows:

```
lvreduce  -L -Target capacity  Logical volume path
```

The following is an example:

```
[root@root ~]# lvscan
  ACTIVE            '/dev/vg0/lv0' [112.00 MB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol00' [540.03 GB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol01' [17.59 GB] inherit
[root@root ~]# pvscan
  PV /dev/mapper/mpatha    VG vg0          lvm2 [296.00 MB / 296.00 MB free]
  PV /dev/mapper/mpathb    VG vg0          lvm2 [1.53 GB / 1.42 GB free]
  PV /dev/mapper/mpathc    VG vg0          lvm2 [1.53 GB / 1.53 GB free]
  PV /dev/sda2             VG VolGroup00   lvm2 [557.62 GB / 0    free]
  Total: 4 [560.98 GB] / in use: 4 [560.98 GB] / in no VG: 0 [0   ]
[root@root ~]# lvreduce -L -100m /dev/vg0/lv0
  WARNING: Reducing active logical volume to 12.00 MB
  THIS MAY DESTROY YOUR DATA (filesystem etc.)
Do you really want to reduce lv0? [y/n]: y
  Reducing logical volume lv0 to 12.00 MB
  Logical volume lv0 successfully resized
[root@root ~]# lvscan
  ACTIVE            '/dev/vg0/lv0' [12.00 MB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol00' [540.03 GB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol01' [17.59 GB] inherit
[root@root ~]# pvscan
  PV /dev/mapper/mpatha    VG vg0          lvm2 [296.00 MB / 296.00 MB free]
  PV /dev/mapper/mpathb    VG vg0          lvm2 [1.53 GB / 1.52 GB free]
  PV /dev/mapper/mpathc    VG vg0          lvm2 [1.53 GB / 1.53 GB free]
  PV /dev/sda2             VG VolGroup00   lvm2 [557.62 GB / 0    free]
  Total: 4 [560.98 GB] / in use: 4 [560.98 GB] / in no VG: 0 [0   ]
```

The output shows that the logical volume capacity is reduced.

Activating a Volume Group

Run the following command to activate a volume group:

```
vgchange -a y Volume group name
```

The following is an example:

```
[root@root ~]# vgchange -a y /dev/vg0
  1 logical volume(s) in volume group "vg0" now active
```

Deactivating a Volume Group

Run the following command to deactivate a volume group:

```
vgchange –a n y Volume group name
```

The following is an example:

```
[root@root ~]# vgchange -a n /dev/vg0
  0 logical volume(s) in volume group "vg0" now active
```

Exporting a Volume Group

A volume group needs to be imported or exported in clusters, data backup, or recovery.

Run the following command to export a volume group:

```
vgexport Volume group name
```

The following is an example:

```
[root@root ~]# vgexport vg0
  Volume group "vg0" successfully exported
[root@root ~]# pvscan
  PV /dev/mapper/mpatha    is in exported VG vg0 [296.00 MB / 296.00 MB free]
  PV /dev/mapper/mpathb    is in exported VG vg0 [1.53 GB / 1.52 GB free]
  PV /dev/mapper/mpathc    is in exported VG vg0 [1.53 GB / 1.53 GB free]
  PV /dev/sda2   VG VolGroup00   lvm2 [557.62 GB / 0    free]
  Total: 4 [560.98 GB] / in use: 4 [560.98 GB] / in no VG: 0 [0   ]
```

Importing a Volume Group

Run the following command to import a volume group:

```
vgimport Volume group name
```

The following is an example (importing a volume group on a local computer):

```
[root@root ~]# vgimport vg0
  Volume group "vg0" successfully imported
[root@root ~]# pvscan
  PV /dev/mapper/mpatha    VG vg0          lvm2 [296.00 MB / 296.00 MB free]
  PV /dev/mapper/mpathb    VG vg0          lvm2 [1.53 GB / 1.52 GB free]
  PV /dev/mapper/mpathc    VG vg0          lvm2 [1.53 GB / 1.53 GB free]
  PV /dev/sda2             VG VolGroup00   lvm2 [557.62 GB / 0    free]
  Total: 4 [560.98 GB] / in use: 4 [560.98 GB] / in no VG: 0 [0   ]
```

Deleting a Logical Volume

Run the following command to delete a logical volume:

```
lvremove Logical volume name
```

The following is an example:

```
[root@root ~]# lvscan
  inactive          '/dev/vg0/lv0' [12.00 MB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol00' [540.03 GB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol01' [17.59 GB] inherit
[root@root ~]# lvremove /dev/vg0/lv0
  Logical volume "lv0" successfully removed
[root@root ~]# lvscan
  ACTIVE            '/dev/VolGroup00/LogVol00' [540.03 GB] inherit
  ACTIVE            '/dev/VolGroup00/LogVol01' [17.59 GB] inherit
```

Deleting a Volume Group

Run the following command to delete a volume group:

```
vgremove Volume group name
```

Perform the following steps:

Step 1 Ensure that all logical volumes are deleted from the volume group.

```
[root@root ~]# vgdisplay -v /dev/vg0
    Using volume group(s) on command line
    Finding volume group "vg0"
  --- Volume group ---
  VG Name               vg0
  System ID
  Format                lvm2
  Metadata Areas        3
  Metadata Sequence No  8
  VG Access             read/write
  VG Status             resizable
  MAX LV                0
  Cur LV                0
  Open LV               0
  Max PV                0
  Cur PV                3
  Act PV                3
  VG Size               3.35 GB
  PE Size               4.00 MB
  Total PE              858
  Alloc PE / Size       0 / 0
  Free  PE / Size       858 / 3.35 GB
  VG UUID               ARkbdL-9ID6-5HCy-DSQG-Aj5z-dQap-9VkM5X

  --- Physical volumes ---
  PV Name               /dev/mapper/mpatha
  PV UUID               g60zN0-3sYn-qPbd-7y0M-dGfZ-hVs7-763Ywo
  PV Status             allocatable
  Total PE / Free PE    74 / 74

  PV Name               /dev/mapper/mpathb
  PV UUID               5UhmY2-fS4p-gdCo-OOgZ-nOa9-AV3H-LkvrNc
  PV Status             allocatable
  Total PE / Free PE    392 / 392

  PV Name               /dev/mapper/mpathc
  PV UUID               iF5Att-fVIj-9dOy-5055-rJlq-pOrS-aW8g2P
  PV Status             allocatable
  Total PE / Free PE    392 / 392
```

Step 2 Delete the volume group.

```
[root@root ~]# vgremove /dev/vg0
  Volume group "vg0" successfully removed
[root@root ~]# vgdisplay -v /dev/vg0
    Using volume group(s) on command line
    Finding volume group "vg0"
  Wiping cache of LVM-capable devices
 Volume group "vg0" not found
```

----End

Deleting a Physical Volume

Run the following command to delete a physical volume:

```
Pvremove Physical volume name
```

The following is an example:

```
[root@root ~]# pvremove /dev/mapper/mpatha
  Labels on physical volume "/dev/mapper/mpatha" successfully wiped
[root@root ~]# pvremove /dev/mapper/mpathb
  Labels on physical volume "/dev/mapper/mpathb" successfully wiped
[root@root ~]# pvremove /dev/mapper/mpathc
  Labels on physical volume "/dev/mapper/mpathc" successfully wiped
```

10 Appendix B High Availability Technology

Currently, only clusters without reservation are supported, such as Oracle RAC. For details, see the compatibility list.