



Cisco UCS C-Series Integrated Management Controller Configuration Guide for RDMA over Converged Ethernet (RoCEv2), Release 4.3

First Published: 2023-12-08

Americas Headquarters

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
<http://www.cisco.com>
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883



CONTENTS

PREFACE

Preface	v
Audience	v
Conventions	v
Related Cisco UCS Documentation	vii

CHAPTER 1

RDMA Over Converged Ethernet (RoCE) version 2	1
Introduction	1

CHAPTER 2

Configuring SMB Direct with RoCEv2 in Windows	3
Guidelines for Using SMB Direct with RoCEv2	3
Windows Requirements	5
Configuring vNIC Properties in Mode 1	5
Configuring RoCEv2 Mode 1 on the Host System	6
Configuring vNIC Properties in Mode 2	7
Configuring RoCEv2 Mode 2 on the Host System	9
Verifying the Configurations on the Host	11
Removing RoCEv2 on vNIC Interface Using Cisco IMC GUI	13

CHAPTER 3

Configuring NVMe Over Fabrics (NVMeoF) with RoCEv2 in Linux	15
Guidelines for using NVMe over Fabrics (NVMeoF) with RoCEv2 on Linux	15
Linux Requirements	16
Configuring RoCEv2 for NVMeoF using Cisco IMC GUI	17
Enabling an SRIOV BIOS Policy	17
Configuring RoCEv2 for NVMeoF on the Host System	18
Installing Cisco enic and enic_rdma Drivers	19
Discovering the NVMe Target	20

Setting Up Device Mapper Multipath 21
 Deleting RoCEv2 Interface Using Cisco IMC CLI 22

CHAPTER 4

Configuring NVMe on RoCEv2 with ESXi 23
 Guidelines for using RoCEv2 Protocol in the Native ENIC driver on ESXi 23
 ESXi nENIC RDMA Requirements 23
 Installing NENIC Driver 24
 Configuring and Enabling RoCEv2 on Cisco IMC 25
 Creating and Configuring the ESXi Adapter Policy in Cisco IMC 25
 ESXi NVMe RDMA Host Side Configuration 26
 NENIC RDMA Functionality 26
 Create Network Connectivity Switches 27
 Creating VMHBA Ports in ESXi 29
 Displaying vmnic and vmrdma Interfaces 31
 NVMe Fabrics and Namespace Discovery 32
 Deleting the ESXi RoCEv2 Interface Using Cisco IMC 33

CHAPTER 5

Using the Cisco IMC CLI to Configure the RoCEv2 Interface 35
 Configuring RoCEv2 Interface Using Cisco IMC CLI 35
 Deleting RoCEv2 Interface Using Cisco IMC CLI 37

CHAPTER 6

Known Issues in RoCEv2 39
 Known Limitations and Behavior 39



Preface

This preface includes the following sections:

- [Audience, on page v](#)
- [Conventions, on page v](#)
- [Related Cisco UCS Documentation, on page vii](#)

Audience

This guide is intended primarily for data center administrators with responsibilities and expertise in one or more of the following:

- Server administration
- Storage administration
- Network administration
- Network security

Conventions

Text Type	Indication
GUI elements	GUI elements such as tab titles, area names, and field labels appear in this font . Main titles such as window, dialog box, and wizard titles appear in this font .
Document titles	Document titles appear in <i>this font</i> .
TUI elements	In a Text-based User Interface, text the system displays appears in <i>this font</i> .
System output	Terminal sessions and information that the system displays appear in <i>this font</i> .
CLI commands	CLI command keywords appear in this font . Variables in a CLI command appear in <i>this font</i> .

Text Type	Indication
[]	Elements in square brackets are optional.
{x y z}	Required alternative keywords are grouped in braces and separated by vertical bars.
[x y z]	Optional alternative keywords are grouped in brackets and separated by vertical bars.
string	A nonquoted set of characters. Do not use quotation marks around the string or the string will include the quotation marks.
<>	Nonprinting characters such as passwords are in angle brackets.
[]	Default responses to system prompts are in square brackets.
!, #	An exclamation point (!) or a pound sign (#) at the beginning of a line of code indicates a comment line.



Note Means *reader take note*. Notes contain helpful suggestions or references to material not covered in the document.



Tip Means *the following information will help you solve a problem*. The tips information might not be troubleshooting or even an action, but could be useful information, similar to a Timesaver.



Timesaver Means *the described action saves time*. You can save time by performing the action described in the paragraph.



Caution Means *reader be careful*. In this situation, you might perform an action that could result in equipment damage or loss of data.



Warning IMPORTANT SAFETY INSTRUCTIONS

This warning symbol means danger. You are in a situation that could cause bodily injury. Before you work on any equipment, be aware of the hazards involved with electrical circuitry and be familiar with standard practices for preventing accidents. Use the statement number provided at the end of each warning to locate its translation in the translated safety warnings that accompanied this device.

SAVE THESE INSTRUCTIONS

Related Cisco UCS Documentation

Documentation Roadmaps

For a complete list of all B-Series documentation, see the *Cisco UCS B-Series Servers Documentation Roadmap* available at the following URL: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/overview/guide/UCS_roadmap.html

For a complete list of all C-Series documentation, see the *Cisco UCS C-Series Servers Documentation Roadmap* available at the following URL: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/overview/guide/ucs_rack_roadmap.html.

For information on supported firmware versions and supported UCS Manager versions for the rack servers that are integrated with the UCS Manager for management, refer to [Release Bundle Contents for Cisco UCS Software](#).



CHAPTER 1

RDMA Over Converged Ethernet (RoCE) version 2

- [Introduction, on page 1](#)

Introduction

RDMA Over Converged Ethernet (RoCEv2)

Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCEv2) allows direct memory access over the network. It does this by encapsulating an Infiniband (IB) transport packet over Ethernet. There are two RoCE versions: RoCEv1 and RoCEv2. RoCEv1 is an Ethernet link layer protocol and hence allows communication between any two hosts in the same Ethernet broadcast domain. RoCEv2 is an internet layer protocol, which means that RoCEv2 packets can be routed.

The RoCEv2 protocol exists on top of either the UDP/IPv4 or the UDP/IPv6 protocol. The UDP destination port number 4791 has been reserved for RoCEv2. Since RoCEv2 packets are routable, the RoCEv2 protocol is sometimes called Routable RoCE.

RoCEv2 is supported on the Windows, Linux and ESXi platforms.

This document provides information to configure RoCEv2 in Mode 1 and Mode 2 using Cisco Integrated Management Controller (Cisco IMC). This document does not provide detailed steps to configure vNIC properties. For detailed steps to configure vNIC properties, refer the [configuration guide](#) for your Cisco IMC release.

NVMe over Fibre

NVMe over Fabrics (NVMeoF) is a communication protocol that allows one computer to access NVMe namespaces available on another computer. The commands for discovering, connecting, and disconnecting a NVMeoF storage device are integrated into the nvme utility provided in Linux. The NVMeoF fabric that Cisco supports is RDMA over Converged Ethernet version 2 (RoCEv2). The eNIC RDMA driver works in conjunction with the eNIC driver, which must be loaded first when configuring NVMeoF.



CHAPTER 2

Configuring SMB Direct with RoCEv2 in Windows

- [Guidelines for Using SMB Direct with RoCEv2, on page 3](#)
- [Windows Requirements, on page 5](#)
- [Configuring vNIC Properties in Mode 1, on page 5](#)
- [Configuring RoCEv2 Mode 1 on the Host System, on page 6](#)
- [Configuring vNIC Properties in Mode 2, on page 7](#)
- [Configuring RoCEv2 Mode 2 on the Host System, on page 9](#)
- [Verifying the Configurations on the Host, on page 11](#)
- [Removing RoCEv2 on vNIC Interface Using Cisco IMC GUI, on page 13](#)

Guidelines for Using SMB Direct with RoCEv2

General Guidelines and Limitations

- Cisco IMC 4.1.x and later releases support Microsoft SMB Direct with RoCEv2 on Windows. Cisco recommends that you have all KB updates from Microsoft. See [Windows Requirements, on page 5](#).



Note RoCEv2 is not supported on Windows Server 2016.

- Cisco recommends you check [UCS Hardware and Software Compatibility](#) specific to your Cisco IMC release to determine support for Microsoft SMB Direct with RoCEv2 on Microsoft 2019.
- Microsoft SMB Direct with RoCEv2 is supported only with Cisco UCS VIC 14xx series adapters. RoCEv2 is not supported on UCS VIC 12xx Series and 13xx Series adapters.



Note RoCE v1 is not supported with Cisco UCS VIC 14xx adapters.

- RoCEv2 configuration is supported only between Cisco adapters. Interoperability between Cisco adapters and third party adapters is not supported.
- RoCEv2 supports two RoCEv2 enabled vNIC per adapter and four virtual ports per adapter interface, independent of SET switch configuration.

- RoCEv2 cannot be used on the same vNIC interface as NVGRE, NetFlow, and VMQ features.
- RoCEv2 enabled vNIC interfaces must have the no-drop QoS system class enabled in Cisco IMC.
- The RoCEv2 properties queue pairs setting must be a minimum of 4 Queue pairs.
- Maximum number of queue pairs per adapter is 2048.
- The maximum number of memory regions per RNIC interface is 131072.
- Cisco IMC does not support fabric failover for vNICs with RoCEv2 enabled.
- QOS no-drop class configuration needs to be configured correctly on upstream switches. For example: N9K
QOS configurations will vary between different upstream switches.
- Configuration of RoCEv2 on the Windows platform requires first configuring RoCEv2 Mode 1, then configuring RoCEv2 Mode 2. Modes 1 and 2 relate to the implementation of Network Direct Kernel Provider Interface (NDKPI): Mode 1 is native RDMA, and Mode 2 involves configuration for the virtual port with RDMA.

MTU Properties

- MTU in Windows is derived from the **Jumbo Packet** advanced property, rather than from the Cisco IMC configuration.
- In older versions of the VIC driver, the MTU was derived from Cisco IMC in standalone mode. This behavior changed for VIC 14xx series adapters, where MTU is controlled from the Windows OS **Jumbo Packet** advanced property. A value configured from Cisco IMC has no effect.
- The RoCEv2 MTU value is always power-of-two and the maximum limit is 4096.
- RoCEv2 MTU is derived from the Ethernet MTU.
- RoCEv2 MTU is the highest power-of-two that is less than the Ethernet MTU. For example:
 - If the Ethernet value is 1500, then the RoCEv2 MTU value is 1024.
 - If the Ethernet value is 4096, then the RoCEv2 MTU value is 4096.
 - If the Ethernet value is 9000, then the RoCEv2 MTU value is 4096.

RoCEv2 Modes of Operation

Cisco IMC provides two modes of RoCEv2 configuration depending on the release:

1. From Cisco IMC Release 4.1(1c) onwards, RoCEv2 can be configured with Mode 1 and Mode 2.

Mode 1 uses the existing RoCEv2 properties with Virtual Machine Queue (VMQ).

Mode 2 introduces additional feature to configure Multi-Queue RoCEv2 properties.

RoCEv2 enabled vNICs for Mode2 operation require that the **Trust Host CoS** is enabled.

RoCEv2 Mode1 and Mode2 are mutually exclusive: RoCEv2 Mode1 must be enabled to operate RoCEv2 Mode2.

2. In Cisco IMC releases prior to 4.1(1c), only mode 1 is supported and could be configured from VMQ RoCE properties.

Downgrade Limitations

Cisco recommends you remove the RoCEv2 configuration before downgrading to any non-supported RoCEv2 release. If the configuration is not removed or disabled, downgrade may fail.

Windows Requirements

Configuration and use of RDMA over Converged Ethernet for RoCEv2 in Windows Server requires the following:

- Windows Server 2019 or Windows Server 2022 with latest Microsoft updates.
- VIC Driver version 5.4.0.x or later
- UCS M5 C-Series servers with VIC 1400 Series adapters: only Cisco UCS VIC 1400 Series or 15000 series adapters are supported.

Configuring vNIC Properties in Mode 1

Follow this procedure to configure vNIC Properties using the VMQ RoCEv2 properties.

Before you begin

Ensure that you are familiar with Cisco IMC GUI interface.

SUMMARY STEPS

1. In the **Navigation** pane, click the **Networking** menu.
2. In the **Adapter Card** pane, click the **vNICs** tab.
3. In the **vNICs** pane, select the vNIC (either the default eth0 or eth1, or any other newly created vNIC).
4. Configure the vNIC properties as desired. See the [configuration guide](#) for detailed procedures. In addition to configuring RoCEv2 in Mode 1, perform the remaining steps.
5. In the **vNIC Properties** pane, under the **Ethernet Interrupt** area, update the following fields:
6. In the vNIC Properties, under the **RoCE Properties** area, update the following fields:

DETAILED STEPS

-
- Step 1** In the **Navigation** pane, click the **Networking** menu.
 - Step 2** In the **Adapter Card** pane, click the **vNICs** tab.
 - Step 3** In the **vNICs** pane, select the vNIC (either the default eth0 or eth1, or any other newly created vNIC).
 - Step 4** Configure the vNIC properties as desired. See the [configuration guide](#) for detailed procedures. In addition to configuring RoCEv2 in Mode 1, perform the remaining steps.
 - Step 5** In the **vNIC Properties** pane, under the **Ethernet Interrupt** area, update the following fields:

Field	Description
Interrupt Count field	Set Interrupt count as Logical Processors times 2 + 4.

Step 6 In the vNIC Properties, under the **RoCE Properties** area, update the following fields:

Field	Description
RoCE check box	Check the RoCE check box to enable the RoCE Properties.
Queue Pairs field	The number of Queue pairs per adapter. Enter an integer between 1 to 2048. We recommend that the value be an integer power of 2. The recommended value is 256.
Memory Regions field	The number of memory regions per adapter. Enter an integer between 1 to 524288. We recommend that the value be an integer power of 2. The recommended value is 131072.
Resource Groups field	The number of resource groups per adapter. Enter an integer between 1 to 128. We recommend that the value be an integer power of 2. The recommended value is 2.
Class of Service drop-down list	NO Drop QOS COS to be specified. This same value should be configured at the up link switch. Default No Drop QOS COS is 5.

What to do next

Perform the host verification to ensure that the Mode 1 is configured correctly. See [Verifying the Configurations on the Host, on page 11](#).

Configuring RoCEv2 Mode 1 on the Host System

Perform this procedure to configure connection between smb-client and smb-server on two host interfaces. For each of these servers, smb-client and smb-server, configure the RoCEv2-enabled vNIC.

Before you begin

Configure RoCEv2 for Mode 1 from Cisco IMC. See [Configuring vNIC Properties in Mode 1, on page 5](#).

- Step 1** In the Windows host, go to the **Device Manager** and select the appropriate Cisco VIC Internet Interface.
- Step 2** Select the **Advanced** tab and verify that the **Network Direct Functionality** property is **Enabled**. If not, enable it and click **OK**. Perform this step for both the smb-server and smb-client vNICs.
- Step 3** Select **Tools > Computer Management > Device Manager > Network Adapter** and select **VIC Network Adapter > Properties > Advanced > Network Direct Functionality**.
- Step 4** Verify that RoCEv2 is enabled on the host operating system using PowerShell.
- Execute the **Get-NetOffloadGlobalSetting** command to verify that **NetworkDirect** is enabled:

```
PS C:\Users\Administrator> Get-NetOffloadGlobalSetting

ReceiveSideScaling           : Enabled
ReceiveSegmentCoalescing    : Enabled
Chimney                      : Disabled
TaskOffload                  : Enabled
NetworkDirect                : Enabled
NetworkDirectAcrossIPSubnets : Blocked
PacketCoalescingFilter      : Disabled
```

Step 5 Bring up Powershell and execute the **SmbClientNetworkInterface** command.

Step 6 Enter **enable - netadapterrdma [-name] ["Ethernetname"]**

Step 7 Verify the overall RoCEv2 Mode 1 configuration at the host:

- Use the Powershell command **netstat -xan** to verify the listeners in both the smb-client and smb-server Windows host; listeners are shown in the command output.
- Go to the smb-client server fileshare and start an I/O operation.
- Go to the performance monitor and check that it displays the RDMA activity.

Step 8 In the Powershell command window, check the connection entries with the **netstat -xan** output command to ensure they are displayed.

Step 9 By default, SMB Direct of Microsoft establishes two RDMA connections per RDMA Interface. You can change the number of RDMA connections per RDMA interface to one or any number of connections.

To increase the number of RDMA connections to 4, execute the following command in PowerShell:

```
PS C:\Users\Administrator> Set-ItemProperty -Path `
"HKLM:\SYSTEM\CurrentControlSet\Services\LanmanWorkstation\Parameters"
ConnectionCountPerRdmaNetworkInterface -Type DWORD -Value 4 -Force
```

What to do next

Configure RoCEv2 Mode 2. See [Configuring vNIC Properties in Mode 2, on page 7](#).

Configuring vNIC Properties in Mode 2

Follow this procedure to configure vNIC Properties in Mode 2. You can perform this procedure using Cisco IMC release 4.1(1c) or higher.

Before you begin

- Ensure that you are familiar with Cisco IMC GUI interface.
- Ensure that you are using Cisco IMC release 4.1(1c) or higher.

SUMMARY STEPS

- In the **Navigation** pane, click the **Networking** menu.
- In the **Adapter Card** pane, click the **vNICs** tab.
- In the **vNICs** pane, select the vNIC (either the default eth0 or eth1, or any other newly created vNIC).
- Configure the vNIC properties as desired. See the [configuration guide](#) for detailed procedures. In addition to configuring RoCEv2 in Mode 1, perform the remaining steps.

5. In the **vNIC Properties** pane, under the **General** area, update the following :
6. In the **vNIC Properties** pane, under the **Ethernet Interrupt** area, update the following fields:
7. In the **vNIC Properties** pane, under the **Multi Queue** area, update the following fields:
8. In the **vNIC Properties** pane, under the **RoCE Properties** area, update the following fields:

DETAILED STEPS

- Step 1** In the **Navigation** pane, click the **Networking** menu.
- Step 2** In the **Adapter Card** pane, click the **vNICs** tab.
- Step 3** In the **vNICs** pane, select the vNIC (either the default eth0 or eth1, or any other newly created vNIC).
- Step 4** Configure the vNIC properties as desired. See the [configuration guide](#) for detailed procedures. In addition to configuring RoCEv2 in Mode 1, perform the remaining steps.
- Step 5** In the **vNIC Properties** pane, under the **General** area, update the following :

Field	Description
Trust Host CoS check box	Check the Trust Host CoS check.
Enable VMQ check box	Check the Enable VMQ check. Note Uncheck the RoCE check to disable RoCE properties before enabling VMQ .
Enable Multi Queue check box	Check the Enable Multi Queue check.
No. of Sub vNICs field	Enter the number of sub vNICs. Default value is 64.

- Step 6** In the **vNIC Properties** pane, under the **Ethernet Interrupt** area, update the following fields:

Field	Description
Interrupt Count field	Set Interrupt count as Logical Processors times 2 + 4.

- Step 7** In the **vNIC Properties** pane, under the **Multi Queue** area, update the following fields:

Field	Description
RoCE check box	Check the RoCE check box to enable the RoCE Properties.
Queue Pairs field	The number of Queue pairs per adapter. Enter an integer between 1 and 2048. We recommend that the value be an integer power of 2. Recommended value is 256.
Memory Regions field	The number of memory regions per adapter. Enter an integer between 1 and 524288. We recommend that the value be an integer power of 2. The recommended value is 65536.
Resource Groups field	The number of resource groups per adapter. Enter an integer between 1 and 128. We recommend that the value be an integer power of 2. The recommended value is 2.

Field	Description
Class of Service drop-down list	NO Drop QOS COS to be specified. This same value should be configured at the up link switch. Default No Drop QOS COS is 5.
Receive Queue Count field	The number of receive queue count per adapter. Enter an integer between 1 and 1000.
Transmit Queue Count field	The number of transmit queue count per adapter. Enter an integer between 1 and 1000.
Completion Queue Count field	The number of completed queue count per adapter. Enter an integer between 1 and 2000.

Step 8 In the vNIC Properties pane, under the **RoCE Properties** area, update the following fields:

Field	Description
RoCE check box	Check the RoCE check box to enable the RoCE Properties.
Queue Pairs field	The Number of Queue pairs per adapter. Enter an integer between 1 to 2048. We recommend that the value be an integer power of 2. Recommended value is 256.
Memory Regions field	The number of memory regions per adapter. Enter an integer between 1 to 524288. We recommend that the value be an integer power of 2. The recommended value is 131072.
Resource Groups field	The Number of resource groups per adapter. Enter an integer between 1 to 128. We recommend that the value be an integer power of 2. Recommended value is 2.
Class of Service drop-down list	NO Drop QOS COS to be specified. This same value should be configured at the up link switch. Default No Drop QOS COS is 5.

What to do next

Perform the host verification to ensure that the Mode 2 is configured correctly. See [Verifying the Configurations on the Host, on page 11](#).

Configuring RoCEv2 Mode 2 on the Host System

Before you begin

1. Configure and confirm the connection for RoCEv2 Mode 2 for both Cisco IMC and the host.
2. Configure RoCEv2 Mode 2 connection for Cisco IMC.

3. Enable Hyper-V at the Windows host server.

Step 1 Go to the Hyper-V switch manager.

Step 2 Create a new Virtual Network Switch (vSwitch) for the RoCEv2-enabled Ethernet interface.

- a) Choose **External Network** and select **VIC Ethernet Interface 2** and **Allow management operating system to share this network adapter**.
- b) Click **OK** to create the virtual switch.

Step 3 Bring up the Powershell interface.

Step 4 Configure the non-default vPort and enable RDMA with the following Powershell commands:

```
add-vmNetworkAdapter -switchname vswitch -name vp1 -managementOS
enable-netAdapterRdma -name "vEthernet (vp1)"
```

- a) Configure set-switch using the following Powershell commands.

```
new-vmswitch -name setswitch -netAdapterName "Ethernet x" -enableEmbeddedTeam $true
```

This creates the switch. Use the following command to display the interfaces:

```
get-netadapterrdma
add-vmNetworkAdapter -switchname setswitch -name svp1
```

You can see the new vPort when you again enter:

```
get-netadapterrdma
```

- b) Add a vPort.

```
add-vmNetworkAdapter -switchname setswitch -name svp1
```

You will see the new vport when you again enter

```
get-netadapterrdma
```

- c) Enable the RDMA on the vport:

```
enable-netAdapterRdma -name "vEthernet (svp1)"
```

Step 5 Configure IPv4 addresses for the vPorts.

Step 6 Create a share in smb-server and map the share in the smb-client.

- a) For smb-client and smb-server in the host system, configure the RoCEv2-enabled vNIC as described above.
- b) Configure the IPV4 addresses on the RDMA enabled vport in both servers, using the same IP subnet and same unique vLAN for both.
- c) Create a share in smb-server and map the share in the smb-client.

Step 7 Verify the Mode 2 configuration.

- a) Use the Powershell command **netstat -xan** to display the listeners and their associated IP addresses.
- b) Start any RDMA I/O in the file share in smb-client.
- c) Issue the **netstat -xan** command again and check for the connection entries to verify they are displayed.

Verifying the Configurations on the Host

Once the configurations are done, you should perform the following:

- Host verification of Mode 1 and Mode 2 configurations
- Host verification for RDMA capable ports
- Verification of RDMA capable ports using **Advanced Property**
- V port assignment on each PF

SUMMARY STEPS

1. NIC driver creates Kernel Socket Listeners on each RDMA capable ports in Mode 1 and V ports in Mode 2 to accept incoming remote RDMA requests.
2. Host verification for RDMA capable ports at host.
3. Netstat-xan output shows established connections in addition to Listeners. If output shows only listeners with traffic, it indicates traffic is passing only on TCP path. If connections are created on PF or vPorts, traffic is passing on RDMA Path.
4. Verification of RDMA capable port using **Advanced Property**. According to the driver, **Network Direct functionality** to be enabled on RDMA Capable VNIC.
5. Verify V Port assignment on each PF.

DETAILED STEPS

- Step 1** NIC driver creates Kernel Socket Listeners on each RDMA capable ports in Mode 1 and V ports in Mode 2 to accept incoming remote RDMA requests.

Example:

```
Ps C:\Users\Administrator . ADMINSTRATOR9 NETSTAT.EXE Xan
active NetworkDirect Connections, Listeners, SharedEndpo int s
Mode    IFIndex Type Local Address Foreign Address PID
Kernel  75 Listener 50.6.5.33:445 NA 0
Kernel  19 Listener 58.6.5.34:445 NA 0
Kernel  38 Listener 59.6.5.35:445 NA 0
Kernel  89 Listener 58.6.5.36:445 NA 0
Kernel  37 Listener 59.6.5.37:445 NA 0
Kernel  23 Listener 59.6.5.38:445 NA 0
Kernel  42 Listener 5e.6.5.39:445 NA 0
Kernel  40 Listener 59.6.5.40:445 NA 0
Kernel  61 Listener 58.6.5.41:445 NA 0
Kernel  79 Listener 58.6.5.42:445 NA 0
Kernel  2 Listener 59.6.5.43:445 NA 0
Kernel  88 Listener 5.5.5.44:445 NA 0
Kernel  11 Listener 59.6.5.45:445 NA 0
Kernel  9 Listener 58.6.5.46:445 NA 0
Kernel  82 Listener 59.6.5.47:445 NA 0
Kernel  83 Listener 58.6.5.48:445 NA 0
Kernel  73 Listener 58.6.5.49:445 NA 0
Kernel  71 Listener 50.6.5.50:445 NA 0
Kernel  se Listener 50.6.5.51:445 NA 0
Kernel  8 Listener 58.6.5.52:445 NA 0
```

Verifying the Configurations on the Host

```
Kernel 5 Listener 50.6.5.53:445 NA 0
Kernel 68 Listener 58.6.5.54:445 NA 0
Kernel 76 Listener 58.6.5.55:445 NA 0
Kernel 34 Listener 50.6.5.56:445 NA 0
```

Step 2 Host verification for RDMA capable ports at host.**Example:**

```
PS C:\Users\administrator> Get-NetAdapterRdma
```

Name	InterfaceDescription	Enabled	PFC	ETS
eth2	Cisco VIC Ethernet Interface #3	True	False	False
eth1	Cisco VIC Ethernet Interface #2	True	False	False
eth0	Cisco VIC Ethernet Interface	False	False	False

Step 3 Netstat-xan output shows established connections in addition to Listeners. If output shows only listeners with traffic, it indicates traffic is passing only on TCP path. If connections are created on PF or vPorts, traffic is passing on RDMA Path.**Example:**

```
PS C:\Users\administrator> netstat -xan
```

Active NetworkDirect Connections, Listeners, SharedEndpoints

Mode	IfIndex	Type	Local Address	Foreign Address	PID
Kernel	3	Connection	50.28.1.19:445	50.28.1.14:9408	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.14:9664	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.84:12480	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.84:13504	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.105:15808	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.97:20672	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.111:10432	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.111:11968	0
Kernel	3	Connection	50.28.1.19:445	50.28.1.111:12736	0
Kernel	3	Connection	50.28.1.19:1472	50.28.1.14:445	0

Step 4 Verification of RDMA capable port using **Advanced Property**. According to the driver, **Network Direct functionality** to be enabled on RDMA Capable VNIC.**Step 5** Verify V Port assignment on each PF.**Example:**

```
PS C:\Users\Administrator> Get-NetAdapterVPort
```

Name	ID	MacAddress	VID	ProcMask	FID	State	ITR	QPairs
Eth3-605-RDMA	0			0:0	PF	Activated	Unknown	1
Eth3-605-RDMA	1	00-15-5D-ED-EE-36		0:2	PF	Activated	Adaptive	1
Eth3-605-RDMA	2	00-15-5D-ED-EE-2A		0:0	PF	Activated	Adaptive	1
Eth3-605-RDMA	3	00-15-5D-ED-EE-35		0:0	PF	Activated	Adaptive	1
Eth3-605-RDMA	4	00-15-5D-ED-EE-2D		0:0	PF	Activated	Adaptive	1
Eth3-605-RDMA	5	00-15-5D-ED-EE-31		0:0	PF	Activated	Adaptive	1
Eth5-605-RDMA	0			0:0	PF	Activated	Unknown	1
Eth5-605-RDMA	1	00-15-5D-ED-EE-33		0:8	PF	Activated	Adaptive	1
Eth5-605-RDMA	2	00-15-5D-ED-EE-2B		0:0	PF	Activated	Adaptive	1
Eth5-605-RDMA	3	00-15-5D-ED-EE-29		0:0	PF	Activated	Adaptive	1
Eth5-605-RDMA	4	00-15-5D-ED-EE-30		0:0	PF	Activated	Adaptive	1
Eth5-605-RDMA	5	00-15-5D-ED-EE-2C		0:0	PF	Activated	Adaptive	1

Removing RoCEv2 on vNIC Interface Using Cisco IMC GUI

You must perform this task to remove RoCEv2 on the vNIC interface.

-
- Step 1** In the **Navigation** pane, click **Networking**.
 - Step 2** Expand **Networking** and select the adapter from which you want to remove RoCEv2 configuration.
 - Step 3** Select **vNICs** tab.
 - Step 4** Select the vNIC from which you want to remove RoCEv2 configuration.
 - Step 5** Expand **RoCE Properties** tab and uncheck the **RoCE** check box.
 - Step 6** Click **Save Changes**.
 - Step 7** Reboot the server for the above changes to take effect.
-



CHAPTER 3

Configuring NVMe Over Fabrics (NVMeoF) with RoCEv2 in Linux

- [Guidelines for using NVMe over Fabrics \(NVMeoF\) with RoCEv2 on Linux](#), on page 15
- [Linux Requirements](#), on page 16
- [Configuring RoCEv2 for NVMeoF using Cisco IMC GUI](#), on page 17
- [Enabling an SRIOV BIOS Policy](#), on page 17
- [Configuring RoCEv2 for NVMeoF on the Host System](#), on page 18
- [Installing Cisco enic and enic_rdma Drivers](#), on page 19
- [Discovering the NVMe Target](#), on page 20
- [Setting Up Device Mapper Multipath](#), on page 21
- [Deleting RoCEv2 Interface Using Cisco IMC CLI](#), on page 22

Guidelines for using NVMe over Fabrics (NVMeoF) with RoCEv2 on Linux

General Guidelines and Limitations

- Cisco recommends that you check [UCS Hardware and Software Compatibility](#) specific to your Cisco IMC release to determine support for NVMeoF. NVMeoF is supported on Cisco UCS C-Series M5 and later servers.
- NVMeoF with RoCEv2 is supported only with the Cisco UCS VIC 14xx series adapters. NVMeoF is not supported on Cisco UCS VIC 12xx or 13xx series adapters.
- When creating RoCEv2 interfaces, use Cisco IMC provided Linux-NVMe-RoCE adapter policy.
- When configuring RoCEv2 interfaces, use both the enic and enic_rdma binary drivers downloaded from cisco.com and install the matched set of enic and enic_rdma drivers. Attempting to use the binary enic_rdma driver downloaded from cisco.com with an inbox enic driver does not work.
- Only two RoCEv2 enabled vNICs per adapter are supported.
- Booting from an NVMeoF namespace is not supported.
- Layer 3 routing is not supported.

- RoCEv2 does not support bonding.
- Saving a crashdump to an NVMeoF namespace during a system crash is not supported.
- NVMeoF cannot be used with usNIC, VxLAN, VMQ, VMMQ, NVGRE, and DPDK features.
- The QoSno drop class configuration must be properly configured on upstream switches such as Cisco Nexus 9000 series switches. QoS configurations vary between different upstream switches.
- Set MTU size correctly on the VLANs and QoS policy on upstream switches.
- Spanning Tree Protocol (STP) may cause temporary loss of network connectivity when a failover or failback event occurs. To prevent this issue from occurring, disable STP on uplink switches.

Interrupts

- Linux RoCEv2 interface supports only MSIx interrupt mode. Cisco recommends that you avoid changing interrupt mode when the interface is configured with RoCEv2 properties.
- The minimum interrupt count for using RoCEv2 with Linux is 8.

Downgrade Limitations

Cisco recommends that you remove the RoCEv2 configuration before downgrading to any non-supported RoCEv2 release.

Linux Requirements

Configuration and use of RoCEv2 in Linux requires the following:

- Red Hat Enterprise Linux:
 - Red Hat Enterprise Linux 7.6 with Z-Kernel 3.10.0-957.27.2
 - Redhat Enterprise Linux 7.7 with Linux Z-kernel-3.10.0-1062.9.1 and above
 - Redhat Enterprise Linux 7.8, 7.9, and 8.2



Note Additional Linux distributions will be supported in later releases.

- InfiniBand kernel API module `ib_core`
- Cisco IMC Release 4.2(2x) or later
- VIC firmware 5.1(1x) or later
- UCS C-Series M5 servers with Cisco UCS VIC 14xx series and 15xxx series adapters
- eNIC driver version 4.0.0.6-802-21 or later provided with the 4.1(1x) release package
- `enic_rdma` driver version 1.0.0.6-802-21 or later provided with the 4.1(1x) release package
- A storage array that supports NVMeoF connection

Configuring RoCEv2 for NVMeoF using Cisco IMC GUI

- Step 1** In the **Navigation** pane, click **Networking**.
- Step 2** Expand **Networking** and click on the adapter to configure RoCEv2 vNIC.
- Step 3** Select the **vNICs** tab.
- Step 4** Perform one the following:
- Click **Add vNIC** to create a new vNIC and modify the properties as mentioned in next step.
OR
 - From the left pane, select an existing vNIC and modify the properties as mentioned in next step.
- Step 5** Expand RoCE Properties.
- Step 6** Select RoCE checkbox.
- Step 7** Modify the following vNIC properties:

Property	Field	Value
Ethernet Interrupt	Interrupt count field	256
Ethernet Receive Queue	Count field	1
	Ring Size field	512
Ethernet Transmit Queue	Count field	1
	Ring Size field	256
Completion Queue	Count field	2
RoCE Properties	Queue Pairs field	1024
	Memory Regions field	131072
	Resource Groups field	8
	Class of Service drop-down list	5

- Step 8** Click **Save Changes**.
- Step 9** Select **Reboot** when prompted.

Enabling an SRIOV BIOS Policy

Use these steps to configure the server with RoCEv2 vNIC to enable the SRIOV BIOS policy before enabling the IOMMU driver in the Linux kernel.

-
- Step 1** In the **Navigation** pane, click **Compute**.
 - Step 2** Expand **BIOS > Configure BIOS > I/O**.
 - Step 3** Select **Intel VT for direct IO** to **Enabled**.
 - Step 4** Click **Save**.
 - Step 5** Reboot the host for the changes to take effect.
-

Configuring RoCEv2 for NVMeoF on the Host System

Before you begin

Configure the server with RoCEv2 vNIC and the SRIOV-enabled BIOS policy.

-
- Step 1** Open the `/etc/default/grub` file for editing.
 - Step 2** Add `intel_iommu=on` at the end of the line in `GRUB_CMDLINE_LINUX` as shown in the following example:


```
sample /etc/default/grub configuration file after adding intel_iommu=on:
# cat /etc/default/grub
GRUB_TIMEOUT=5
GRUB_DISTRIBUTOR="$(sed 's, release .*$,,g' /etc/system-release)"
GRUB_DEFAULT=saved
GRUB_DISABLE_SUBMENU=true
GRUB_TERMINAL_OUTPUT="console"
GRUB_CMDLINE_LINUX="crashkernel=auto rd.lvm.lv=rhel/root rd.lvm.lv=rhel/swap biosdevname=1 rhgb quiet
intel_iommu=on"
GRUB_DISABLE_RECOVERY="true"
```
 - Step 3** Save the file.
 - Step 4** Run the following command to generate a new `grub.cfg` file:
 - For Legacy boot:


```
# grub2-mkconfig -o /boot/grub2/grub.cfg
```
 - For UEFI boot:


```
# grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
```
 - Step 5** Reboot the server for the changes to take effect after enabling IOMMU.
 - Step 6** Use the following to check the output file and verify that the server is booted with the `intel_iommu=on` option:

```
cat /proc/cmdline | grep iommu
```

Note its inclusion at the end of the output.

Example:

```
[root@localhost basic-setup]# cat /proc/cmdline | grep iommu
BOOT_IMAGE=vmlinux-3.10.0-957.27.2.el7.x86_64 root=/dev/mapper/rhel-root ro crashkernel=auto
rd.lvm.lv=rhel/root rd.lvm.lv=rhel/swap rhgb quiet intel_iommu=on LANG=en_US.UTF-8
```

What to do next

Download the enic and enic_rdma drivers.

Installing Cisco enic and enic_rdma Drivers

The enic_rdma driver requires enic driver. When installing enic and enic_rdma drivers, download and use the matched set of enic and enic_rdma drivers from [here](#). Do not attempt to use the binary enic_rdma driver downloaded from cisco.com with an inbox enic driver.

Before you begin

- RHEL 7.6
- Server updated with kernel version 3.10.0-957.27.2 or above
- InfiniBand kernel API module ib_core

Step 1 Run the following command to install the enic and enic_rdma rpm packages:

```
# rpm -ivh kmod-enic-<version>.x86_64.rpm kmod-enic_rdma-<version>.x86_64.rpm
```

The enic_rdma driver is now installed but not loaded in the running kernel.

Step 2 Reboot the server to load enic_rdma driver into the running kernel.

Step 3 Run the following command to verify the installation of enic_rdma driver and RoCEv2 interface:

```
# dmesg | grep enic_rdma
[ 4.025979] enic_rdma: Cisco VIC Ethernet NIC RDMA Driver, ver 1.0.0.6-802.21 init
[ 4.052792] enic 0000:62:00.1 eth1: enic_rdma: IPv4 RoCEv2 enabled
[ 4.081032] enic 0000:62:00.2 eth2: enic_rdma: IPv4 RoCEv2 enabled
```

Step 4 Run the following command to load the nvme-rdma kernel module:

```
# modprobe nvme-rdma
```

After the server reboots, nvme-rdma kernel module is unloaded. To load nvme-rdma kernel module on every server reboot, create nvme_rdma.conf file using:

```
# echo nvme_rdma > /etc/modules-load.d/nvme_rdma.conf
```

Note For more information about enic_rdma after installation, use the `rpm -q -l kmod-enic_rdma` command to extract the README file.

What to do next

Discover targets and connect to NVMe namespaces. If your system needs multipath access to the storage, see [Setting Up Device Mapper Multipath, on page 21](#).

Discovering the NVMe Target

Use this procedure to discover the NVMe target and connect NVMe namespaces.

Before you begin

- Ensure that you have **nvme-cli** version 1.6 or later.
- Configure the IP address on the RoCEv2 interface and make sure the interface can ping the target IP.

Step 1 Perform the following to create an `nvme` folder in `/etc`, and then manually generate `hostnqn`.

```
# mkdir /etc/nvme
# nvme gen-hostnqn > /etc/nvme/hostnqn
```

Step 2 Perform the following to create a `settos.sh` file and run the script to set priority flow control (PFC) in IB frames.

Note To avoid failure of sending NVMeoF traffic, you must create and run this script after every server reboot.

```
# cat settos.sh
#!/bin/bash
for f in `ls /sys/class/infiniband`;
do
    echo "setting TOS for IB interface:" $f
    mkdir -p /sys/kernel/config/rdma_cm/$f/ports/1
    echo 186 > /sys/kernel/config/rdma_cm/$f/ports/1/default_roce_tos
done
```

Step 3 Run the following command to discover the NVMe target:

```
nvme discover --transport=rdma --traddr=<IP address of transport target port>
```

Example:

To discover the target at 50.2.85.200:

```
# nvme discover --transport=rdma --traddr=50.2.85.200

Discovery Log Number of Records 1, Generation counter 2
====Discovery Log Entry 0====
trtype: rdma
adrfam: ipv4
subtype: nvme subsystem
treq: not required
portid: 3
trsvcid: 4420
subnqn: nqn.2010-06.com.purestorage:flasharray.9a703295ee2954e
traddr: 50.2.85.200
rdma_prtype: roce-v2
rdma_qptype: connected
rdma_cms: rdma-cm
rdma_pkey: 0x0000
```

Step 4 Run the following command to connect to the discovered NVMe target:

```
nvme connect --transport=rdma --traddr=<IP address of transport target port>> -n <subnqn value from
nvme discover>
```

Example:

To discover the target at 50.2.85.200 and the subnqn value found above:

```
# nvme connect --transport=rdma --traddr=50.2.85.200 -n
nqn.2010-06.com.purestorage:flasharray.9a703295ee2954e
```

Step 5 Use the `nvme list` command to verify the mapped namespaces:

```
# nvme list
Node                               SN                               Model                               Namespace Usage
Format                             FW Rev
-----
/dev/nvme0n1                       09A703295EE2954E               Pure Storage FlashArray           72656      4.29 GB / 4.29 GB
  512 B + 0 B                       99.9.9
/dev/nvme0n2                       09A703295EE2954E               Pure Storage FlashArray           72657      5.37 GB / 5.37 GB
  512 B + 0 B                       99.9.9
```

Setting Up Device Mapper Multipath

If your system is configured with Device Mapper Multipathing (DM Multipath), use this procedure to set up device mapper multipath.

Step 1 Install the `device-mapper-multipath` package.

Step 2 Perform the following to enable and start `multipathd`:

```
# mpathconf --enable --with_multipathd y
```

Step 3 Edit the `etc/multipath.conf` file to use the following values:

```
defaults {
    polling_interval          10
    path_selector             "queue-length 0"
    path_grouping_policy      multibus
    fast_io_fail_tmo         10
    no_path_retry             0
    features                  0
    dev_loss_tmo              60
    user_friendly_names       yes
}
```

Step 4 Perform the following to flush with the updated multipath device maps:

```
# multipath -F
```

Step 5 Perform the following to restart multipath service:

```
# systemctl restart multipathd.service
```

Step 6 Perform the following to rescan multipath devices:

```
# multipath -v2
```

Step 7 Perform the following to check the multipath status:

```
# multipath -ll
```

Deleting RoCEv2 Interface Using Cisco IMC CLI

SUMMARY STEPS

1. server # **scope chassis**
2. server/chassis # **scope adapter** *index_number*
3. server/chassis/adapter # **scope host-eth-if** *vNIC_name*
4. server/chassis/adapter/host-eth-if # **set rocev2 disabled**
5. server/chassis/adapter/host-eth-if *# **commit**

DETAILED STEPS

	Command or Action	Purpose
Step 1	server # scope chassis	Enters the chassis command mode.
Step 2	server/chassis # scope adapter <i>index_number</i>	Enters the command mode for the adapter card at the PCI slot number specified by <i>index_number</i> . Note Ensure that the server is powered on before you attempt to view or change adapter settings. To view the <i>index</i> of the adapters configured on your server, use the show adapter command.
Step 3	server/chassis/adapter # scope host-eth-if <i>vNIC_name</i>	Enters the command mode for the vNIC specified by <i>vNIC_name</i> .
Step 4	server/chassis/adapter/host-eth-if # set rocev2 disabled	Disables RoCE properties on the vNIC.
Step 5	server/chassis/adapter/host-eth-if *# commit	Commits the transaction to the system configuration. Note The changes take effect when the server is rebooted.

Example

```
server# scope chassis
server/chassis # scope adapter 1
server/chassis/adapter # scope host-eth-if vNIC_Test
server/chassis/adapter/host-eth-if # set rocev2 disabled
server/chassis/adapter/host-eth-if *# commit
```



CHAPTER 4

Configuring NVMe on RoCEv2 with ESXi

- [Guidelines for using RoCEv2 Protocol in the Native ENIC driver on ESXi, on page 23](#)
- [ESXi nENIC RDMA Requirements, on page 23](#)
- [Installing NENIC Driver, on page 24](#)
- [Configuring and Enabling RoCEv2 on Cisco IMC, on page 25](#)
- [ESXi NVMe RDMA Host Side Configuration, on page 26](#)

Guidelines for using RoCEv2 Protocol in the Native ENIC driver on ESXi

General Guidelines and Limitations:

- Cisco IMC release 4.2(3b) supports RoCEv2 only on ESXi 7.0 U3.
- Cisco recommends you check [UCS Hardware and Software Compatibility](#) specific to your Cisco IMC release to determine support for ESXi. RoCEv2 on ESXi is supported on UCS M6 C-Series servers with Cisco UCS VIC 15000 Series adapters.
- RoCEv2 on ESXi is not supported on UCS VIC 1400 Series adapters.
- RDMA on ESXi nENIC currently supports only ESXi NVMe that is part of the ESXi kernel. The current implementation does not support the ESXi user space RDMA application.
- Multiple mac addresses and multiple VLANs are supported only on VIC 15000 Series adapters.
- RoCEv2 supports maximum two RoCEv2 enabled interfaces per adapter.
- PvrDMA, VSAN over RDMA, and iSER are not supported.

Downgrade Limitations:

- Cisco recommends you remove the RoCEv2 configuration before downgrading to any non-supported RoCEv2 release.

ESXi nENIC RDMA Requirements

Configuration and use of RoCEv2 in ESXi requires the following:

- VMWare ESXi version 7.0 U3.
- Cisco IMC release 4.2.3 or later
- RoCEv2 is supported on Cisco UCS M6 C-Series servers with Cisco UCS VIC 15xxx adapters.
- nenic-2.0.4.0-1OEM.700.1.0.15843807.x86_64.vib provides both standard eNIC and RDMA support.
- A storage array that supports NVMeoF connection. Currently, tested and supported on Pure Storage with Cisco Nexus 9300 Series switches.

Downgrade Limitations:

- Cisco recommends you remove the RoCEv2 configuration before downgrading to any non-supported RoCEv2 release.

Installing NENIC Driver

The enic drivers, which contain the rdma driver, are available as a combined package. Download and use the enic driver on cisco.com.

These steps assume this is a new installation.



Note While this example uses the /tmp location, you can place the file anywhere that is accessible to the ESX console shell.

Step 1 Copy the enic VIB or offline bundle to the ESX server. The example below uses the Linux **scp** utility to copy the file from a local system to an ESX server located at 10.10.10.10: and uses the location /tmp.

```
scp nenic-2.0.4.0-1OEM.700.1.0.15843807.x86_64.vib root@10.10.10.10:/tmp
```

Step 2 Specifying the full path, issue the command shown below.

```
esxcli software vib install -v {VIBFILE}
```

or

```
esxcli software vib install -d {OFFLINE_BUNDLE}
```

Here is an example:

```
esxcli software vib install -v /tmp/nenic-2.0.4.0-1OEM.700.1.0.15843807.x86_64.vib
```

Note Depending on the certificate used to sign the VIB, you may need to change the host acceptance level. To do this, use the command: `esxcli software acceptance set --level=<level>`

Depending on the type of VIB being installed, you may need to put ESX into maintenance mode. This can be done through the VI Client, or by adding the `--maintenance-mode` option to the above `esxcli` command.

Upgrading NENIC Driver

a. To upgrade NENIC driver, enter the command:

```
esxcli software vib update -v {VIBFILE}
```


or

```
esxcli software vib update -d {OFFLINE_BUNDLE}
```

- b. Copy the enic VIB or offline bundle to the ESX server using Step 1 given above.

Configuring and Enabling RoCEv2 on Cisco IMC

Creating and Configuring the ESXi Adapter Policy in Cisco IMC

This procedure applies to configuring the ESXi adapter policy for RoCEv2.

Before you begin

Download and install the enic-nvme driver which supports RoCEv2.

- Step 1** In the **Navigation** pane, click the **Networking** menu.
- Step 2** Expand **Networking** and click on the adapter to configure RoCEv2 vNIC.
- Step 3** Select the vNICs tab.
- Step 4** Perform one the following:
 - Click **Add vNIC** to create a new vNIC and modify the properties as mentioned in next step.
 - From the left pane, select an existing vNIC and modify the properties as mentioned in next step.
- Step 5** Expand **General** pane.
 - a) On the **MAC address** dropdown, select the **Auto** checkbox or enter the desired address.
 - b) Select which VLAN you want use use from the drop-down list.
 - c) Click **OK**.
- Step 6** Expand **RoCE Properties**.
- Step 7** Select **RoCE** checkbox.
- Step 8** Modify the following vNIC properties:

Property	Field	Value
Ethernet Interrupt	Interrupt count field	256
	Coalescing Time field	125
	Interrupt Mode field	MSIx
	Coalescing Type field	MIN
Ethernet Receive Queue	Count field	1
	Ring Size field	512

Property	Field	Value
Ethernet Transmit Queue	Count field	1
	Ring Size field	256
Completion Queue	Count field	2
RoCE Properties	Queue Pairs field	1024
	Memory Regions field	131072
	Resource Groups field	8
	Class of Service drop-down list	5

Step 9 Click **Save Changes**.

Step 10 Select **Reboot**.

ESXi NVMe RDMA Host Side Configuration

NENIC RDMA Functionality

One major difference between the use case for RDMA on Linux and ESXi is in ESXi terminology. The physical interface (vmnic) MAC is not used for RoCEv2 traffic. Instead, the VMkernel port (vmk) MAC is used.

The outgoing RoCe packets uses vmk MAC in ethernet source mac field and incoming RoCE packets use the vmk MAC in the ethernet destination mac field. The vmk MAC address is a VMware mac address assigned to the vmk interface when it is created.

Linux implementation used the physical interface MAC in source MAC address field in the ROCE packets. This Linux MAC is usually a Cisco MAC address configured to the VNIC.

If you ssh into the host and use the **esxcli network ip interface list** command, you can see the MAC address.

```

vmk0
  Name: vmk0
  MAC Address: 2c:f8:9b:a1:4c:e7
  Enabled: true
  Portset: vSwitch0
  Portgroup: Management Network
  Netstack Instance: defaultTcpipStack
  VDS Name: N/A
  VDS UUID: N/A
  VDS Port: N/A
  VDS Connection: -1
  Opaque Network ID: N/A
  Opaque Network Type: N/A
  External ID: N/A
  MTU: 1500
  TSO MSS: 65535
  RXDispQueue Size: 2
  Port ID: 67108881

```

You must create a vSphere Standard Switch to provide network connectivity for hosts, virtual machines, and to handle VMkernel traffic. Depending on the connection type that you want to create, you can create a new vSphere Standard Switch with a VMkernel adapter, only connect physical network adapters to the new switch, or create the switch with a virtual machine port group.

Create Network Connectivity Switches

Use these steps to create a vSphere Standard Switch to provide network connectivity for hosts, virtual machines, and to handle VMkernel traffic.

Before you begin

Download the enic and enic-rdma drivers.

-
- Step 1** In the vSphere Client, navigate to the host.
 - Step 2** On the **Configure** tab, expand **Networking** and select **Virtual Switches**.
 - Step 3** Click on **Add Networking**.

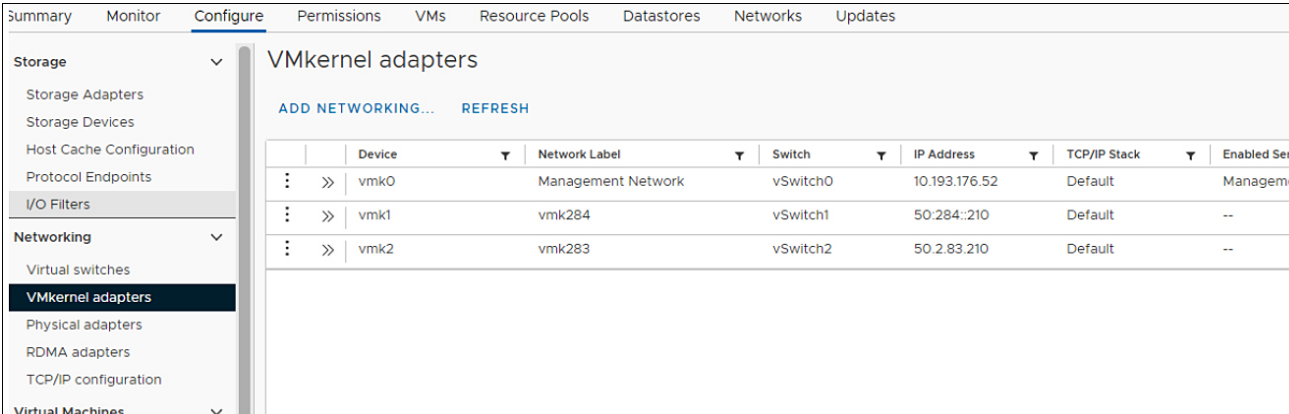
The available network adapter connection types are:

- **Vmkernel Network Adapter**
Creates a new VMkernel adapter to handle host management traffic
- **Physical Network Adapter**
Adds physical network adapters to a new or existing standard switch.
- **Virtual Machine Port Group for a Standard Switch**
Creates a new port group for virtual machine networking.

- Step 4** Select connection type **Vmkernel Network Adapter**.
- Step 5** Select **New Standard Switch** and click **Next**.
- Step 6** Add physical adapters to the new standard switch.
- Under **Assigned Adapters**, select **New Adapters**.
 - Select one or more adapters from the list and click **OK**. To promote higher throughput and create redundancy, add two or more physical network adapters to the Active list.
 - (Optional) Use the up and down arrow keys to change the position of the adapter in the Assigned Adapters list.
 - Click **Next**.
- Step 7** For the new standard switch you just created for the VMadapter or a port group, enter the connection settings for the adapter or port group.
- Enter a label that represents the traffic type for the VMkernel adapter.
 - Set a VLAN ID to identify the VLAN the VMkernel uses for routing network traffic.
 - Select IPV4 or IPV6 or both.
 - Select an MTU size from the drop-down menu. Select Custom if you wish to enter a specific MTU size. The maximum MTU size is 9000 bytes.
- Note** You can enable Jumbo Frames by setting an MTU greater than 1500.
- Select a TCP/IP stack. After setting the TCP/IP stack for the VMkernel adapter. To use the default TCP/IP stack, select it from the available services.
- Note** Be aware that the TCP/IP stack for the VMkernel adapter cannot be changed later.
- Configure IPV4 and/or IPV6 settings.
- Step 8** On the **Ready to Complete** page, click **Finish**.
- Step 9** Check the VMkernel ports for the VM Adapters or port groups with NVMe RDMA in the vSphere client, as shown in the Results below.

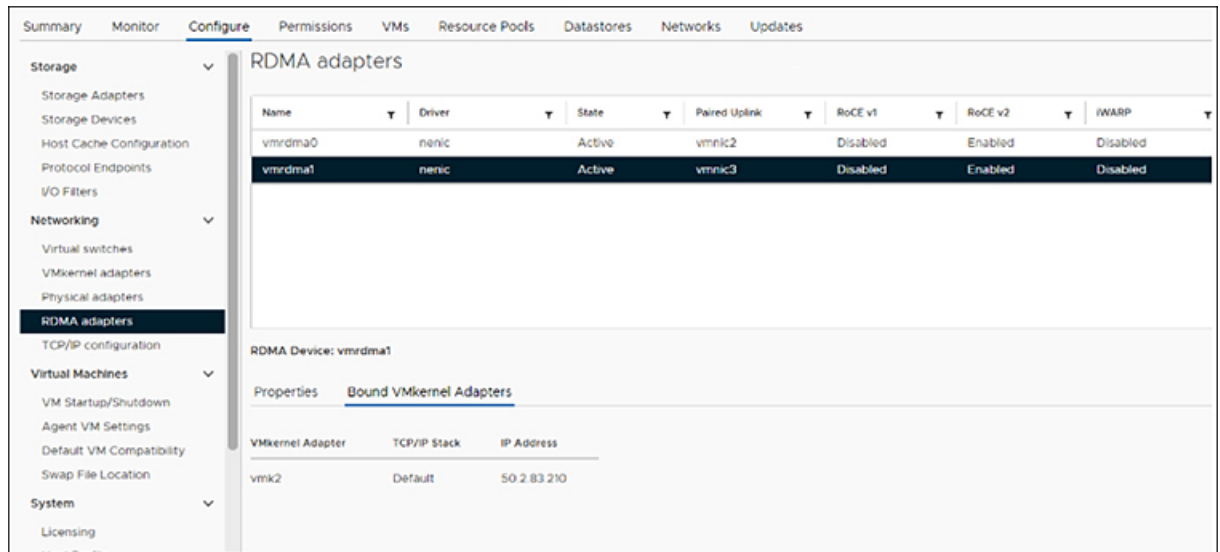
The VMkernel ports for the VM Adapters or port groups with NVMe RDMA are shown below.

Example



Device	Network Label	Switch	IP Address	TCP/IP Stack	Enabled Services
vmk0	Management Network	vSwitch0	10.193.176.52	Default	Management
vmk1	vmk284	vSwitch1	50.284::210	Default	--
vmk2	vmk283	vSwitch2	50.2.83.210	Default	--

The VRDMA Port groups created with NVMeRDMA supported vmnic appear as below.



What to do next

Create vmhba ports on top of vmrdma ports.

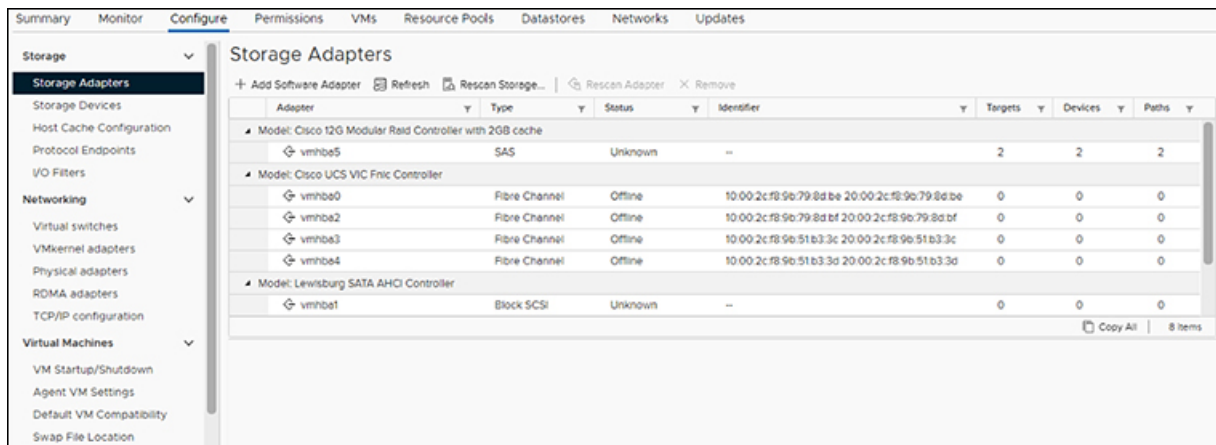
Creating VMHBA Ports in ESXi

Use the following steps for creating vmhba ports on top of the vmrdma adapter ports.

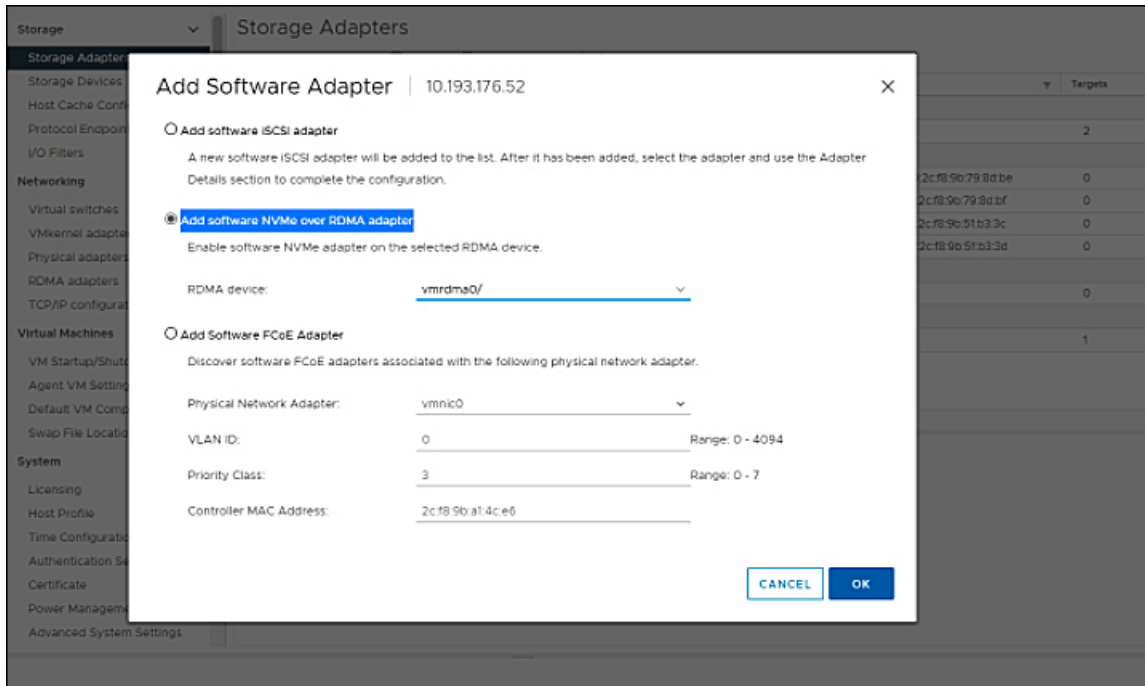
Before you begin

Create the adapter ports for network connectivity.

- Step 1** Go to vCenter where your ESXi host is connected.
- Step 2** Click on **Host>Configure>Storage adapters**.

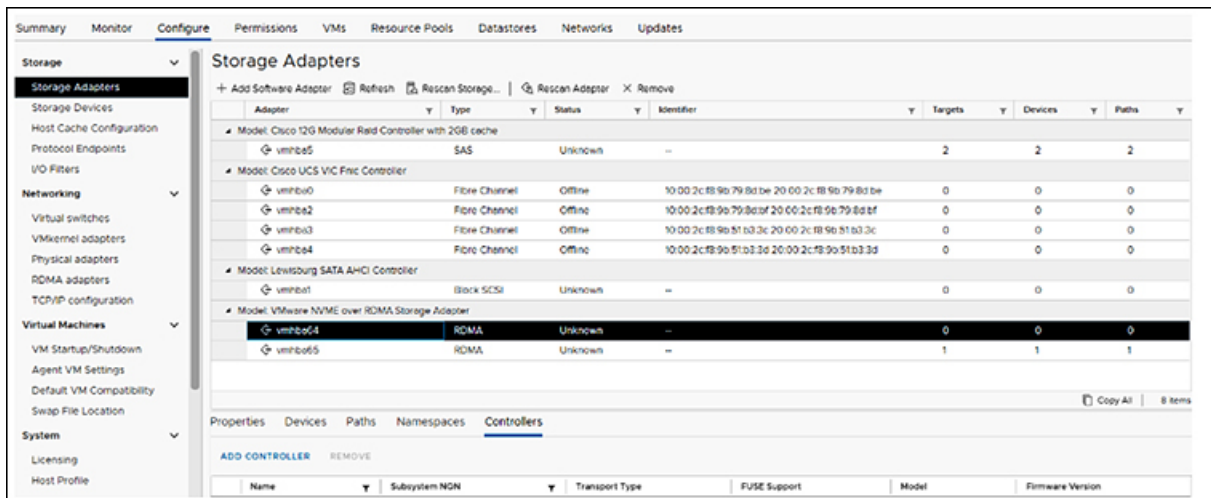


- Step 3** Click **+Add Software Adapter** . The following dialog box is displayed.



- Step 4** Select **Add software NVMe over RDMA adapter** and the vmrdma port you want to use.
- Step 5** Click **OK**.

The vmhba ports for the VMware NVMe over RDMA storage adapter is displayed as shown as in the example below.



What to do next

Configure NVME.

Displaying vmnic and vmrmda Interfaces

ESXi creates a vmnic interface for each enic VNIC configured to the host.

Before you begin

Create Network Adapters and VHBA ports.

Step 1 Ssh into the host system.

Step 2 Enter `esxcfg-nics -l` to list the vmnics on ESXi.

```
Name PCI Driver Link Speed Duplex MAC Address MTU Description
vmnic0 0000:3b:00.0 ixgben Down 0Mbps Half 2c:f8:9b:a1:4c:e6 1500 Intel(R) Ethernet Controller X550
vmnic1 0000:3b:00.1 ixgben Up 1000Mbps Full 2c:f8:9b:a1:4c:e7 1500 Intel(R) Ethernet Controller X550
vmnic2 0000:1d:00.0 nenic Up 50000Mbps Full 2c:f8:9b:79:8d:bc 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
vmnic3 0000:1d:00.1 nenic Up 50000Mbps Full 2c:f8:9b:79:8d:bd 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
vmnic4 0000:63:00.0 nenic Down 0Mbps Half 2c:f8:9b:51:b3:3a 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
vmnic5 0000:63:00.1 nenic Down 0Mbps Half 2c:f8:9b:51:b3:3b 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
```

esxcli network nic list

```
Name PCI Device Driver Admin Status Link Status Speed Duplex MAC Address MTU Description
-----
vmnic0 0000:3b:00.0 ixgben Up Down 0 Half 2c:f8:9b:a1:4c:e6 1500 Intel(R) Ethernet Controller X550
vmnic1 0000:3b:00.1 ixgben Up Up 1000 Full 2c:f8:9b:a1:4c:e7 1500 Intel(R) Ethernet Controller X550
vmnic2 0000:1d:00.0 nenic Up Up 50000 Full 2c:f8:9b:79:8d:bc 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
vmnic3 0000:1d:00.1 nenic Up Up 50000 Full 2c:f8:9b:79:8d:bd 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
vmnic4 0000:63:00.0 nenic Up Down 0 Half 2c:f8:9b:51:b3:3a 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
vmnic5 0000:63:00.1 nenic Up Down 0 Half 2c:f8:9b:51:b3:3b 1500 Cisco Systems Inc Cisco VIC Ethernet NIC
```

Step 3 When the enic driver registers with ESXi the RDMA device for a RDMA capable VNIC, ESXi creates a vmrmda device and links it to the corresponding vmnic. Enter `esxcli rdma device list` to list the vmrmda devices.

```
[root@StockholmRackServer:~] esxcli rdma device list
Name Driver State MTU Speed Paired Uplink Description
-----
vmrmda0 nenic Active 4096 50 Gbps vmnic1 Cisco UCS VIC 15XXX (A0)
vmrmda1 nenic Active 4096 50 Gbps vmnic2 Cisco UCS VIC 15XXX (A0)
[root@StockholmRackServer:~] esxcli rdma device vmknics list
Device Vmknics NetStack
-----
vmrmda0 vmk1 defaultTcpipStack
vmrmda1 vmk2 defaultTcpipStack
```

Step 4 Use `esxcli rdma device list` to check the protocols supported by the vmrmda interface.

For enic, RoCE v2 will be the only protocol supported from this list. The output of this command should match the RoCEv2 configuration on the VNIC.

Step 5 Use `esxcli rdma device protocol list` to check the protocols supported by the vmrmda interface.

For enic, RoCE v2 will be the only protocol supported from this list. The output of this command should match the RoCEv2 configuration on the VNIC.

```
[root@ESXi7U3Bodega:~] esxcli rdma device protocol list
Device RoCE v1 RoCE v2 iWARP
-----
vmrmda0 false true false
vmrmda1 false true false
[root@ESXi7U3Bodega:~]
```

Step 6 Use `esxcli nvme adapter list` to list the NVMe adapters and the vmrDMA and vmnic interfaces it is configured on.

```
[root@ESXi7U3Bodega:~] esxcli nvme adapter list
Adapter  Adapter Qualified Name      Transport Type  Driver      Associated Devices
-----  -
vmhba64  aqn:nvmerdma:2c-f8-9b-79-8d-bc  RDMA           nvmerdma   vmrDMA0, vmnic2
vmhba65  aqn:nvmerdma:2c-f8-9b-79-8d-bd  RDMA           nvmerdma   vmrDMA1, vmnic3
[root@ESXi7U3Bodega:~] █
```

Step 7 All vmhbas in the system can be listed using `esxcli storage core adapter list`.

```
[root@ESXi7U3Bodega:~] esxcli storage core adapter list
HBA Name  Driver      Link State  UID                               Capabilities      Description
-----  -
vmhba0    nfnic       link-down   fc.10002cf89b798dbe:20002cf89b798dbe  Second Level Lun ID (0000:1d:00.2) Cisco Corporation Cisco UCS VIC Fnic Controller
vmhba1    vmw_ahci    link-n/a    sata.vmhba1                        Second Level Lun ID (0000:00:11.5) Intel Corporation Lewisburg SATA AHCI Controller
vmhba2    nfnic       link-down   fc.10002cf89b798dbf:20002cf89b798dbf  Second Level Lun ID (0000:1d:00.3) Cisco Corporation Cisco UCS VIC Fnic Controller
vmhba3    nfnic       link-down   fc.10002cf89b51b33c:20002cf89b51b33c  Second Level Lun ID (0000:63:00.2) Cisco Corporation Cisco UCS VIC Fnic Controller
vmhba4    nfnic       link-down   fc.10002cf89b51b33d:20002cf89b51b33d  Second Level Lun ID (0000:63:00.3) Cisco Corporation Cisco UCS VIC Fnic Controller
vmhba5    lsi_mr3     link-n/a    sas.5cc167e9732f9b00                Second Level Lun ID (0000:3c:00.0) Broadcom Cisco 12G Modular Raid Controller with 2GB cache
vmhba64   nvmerdma    link-n/a    rdma.vmic2:2c:f8:9b:79:8d:bc          VMware NVMe over RDMA Storage Adapter on vmrDMA0
vmhba65   nvmerdma    link-n/a    rdma.vmic3:2c:f8:9b:79:8d:bd          VMware NVMe over RDMA Storage Adapter on vmrDMA1
[root@ESXi7U3Bodega:~] █
```

What to do next

Configure NVME.

NVMe Fabrics and Namespace Discovery

This process is performed through the ESXi command line interface,

Before you begin

Create and configure the adapter policy.

Step 1 Check and enable NVME on the vmrDMA device.

```
esxcli nvme fabrics enable -p RDMA -d vmrDMA0
```

The system should return a message showing if NVME is enabled.

Step 2 Discover the nvme on the array by entering the following command:

```
esxcli nvme fabrics discover -a vmhba64 -l transport_address
```

figure with `esxcli nvme fabrics discover -a vmhba64 -l 50.2.84.100`

The output will list the following information: Transport Type, Address Family, Subsystem Type, Controller ID, Admin Queue, Max Size, Transport Address, Transport Service ID, and Subsystem NQN

You will see output on the NVMe controller.

Step 3 Perform NVMe fabric interconnect.

```
esxcli nvme fabrics discover -a vmhba64 -l transport_address p Transport Service ID -s Subsystem NQN
```

Step 4 The NVMe controller should show a list of the controllers connected to NVMe The NVMe namespace list should show all the NVMe drives discovered.


```
esxcli nvme controller list RDMA -d vmrdma0
```

```
[root@ESXi7U3Bodega:~] esxcli nvme controller list
Name
-----
nqn.2010-06.com.purestorage:flasharray.Sab274df5b161455#vmhba64#50.2.84.100:4420
nqn.2010-06.com.purestorage:flasharray.Sab274df5b161455#vmhba65#50.2.83.100:4420
[root@ESXi7U3Bodega:~] esxcli nvme namespace list
Name
-----
Controller Number  Namespace ID  Block Size  Capacity in MB
-----
oui.00e6d65b65a8f34024a9374e00011745  258          71493       512          102400
oui.00e6d65b65a8f34024a9374e00011745  259          71493       512          102400
[root@ESXi7U3Bodega:~] █
```

Example

The following example shows esxcli discovery commands executed on the server.

```
[root@ESXiUCSA:~] esxcli nvme fabrics enable -p RDMA -d vmrdma0
NVMe already enabled on vmrdma0
[root@ESXiUCSA:~] esxcli nvme fabrics discover -a vmhba64 -l 50.2.84.100
Transport Type Address Family Subsystem Type Controller ID Admin Queue Max Size Transport
Address Transport Service ID Subsystem NQN
-----
RDMA          IPV4          NVM          65535        31          50.2.84.100
          4420          nq.210-06.com.purestorage:flasharray:2dp1239anjkl484

[root@ESXiUCSA:~] esxcli nvme fabrics discover -a vmhba64 -l 50.2.84.100 p 4420 -s
nq.210-06.com.purestorage:flasharray:2dp1239anjkl484
Controller already connected
```

Deleting the ESXi RoCEv2 Interface Using Cisco IMC

Use these steps to delete the ESXi RoCEv2 configuration for a specific port.

- Step 1** In the **Navigation** pane, click **Networking**.
- Step 2** Expand **Networking** and select the adapter from which you want to remove RoCEv2 configuration.
- Step 3** Select **vNICs** tab.
- Step 4** Select the vNIC from which you want to delete the ESXi RoCEv2 configuration.
- Step 5** Expand **RoCE Properties** tab and uncheck the **RoCE** check box.
- Step 6** Click **Save Changes**.
- Step 7** Reboot the server for the above changes to take effect.



CHAPTER 5

Using the Cisco IMC CLI to Configure the RoCEv2 Interface

- [Configuring RoCEv2 Interface Using Cisco IMC CLI, on page 35](#)
- [Deleting RoCEv2 Interface Using Cisco IMC CLI, on page 37](#)

Configuring RoCEv2 Interface Using Cisco IMC CLI

Use the following steps to configure RoCEv2 interface using Cisco IMC CLI interface.

Before you begin

- Ensure that you are familiar with Cisco IMC CLI interface.
- You must log in with admin privileges.

SUMMARY STEPS

1. `server # scope chassis`
2. `server/chassis # scope adapter index_number`
3. `server/chassis/adapter # create host-eth-if vNIC_name`
4. `server/chassis/adapter/host-eth-if ## set rocev2 enabled`
5. `server/chassis/adapter/host-eth-if ## set rdma-cos 5`
6. `server/chassis/adapter/host-eth-if ## set rdma_mr 131072`
7. `server/chassis/adapter/host-eth-if ## set rdma_qp 1024`
8. `server/chassis/adapter/host-eth-if ## set rdma_resgrp 8`
9. `server/chassis/adapter/host-eth-if ## scope comp-queue`
10. `server/chassis/adapter/host-eth-if/comp-queue ## set cq-count 2`
11. `server/chassis/adapter/host-eth-if/comp-queue ## exit`
12. `server/chassis/adapter/host-eth-if ## scope trans-queue`
13. `server/chassis/adapter/host-eth-if/trans-queue ## set wq-count 1`
14. `server/chassis/adapter/host-eth-if/trans-queue ## set wq-ring-size 256`
15. `server/chassis/adapter/host-eth-if/trans-queue ## exit`
16. `server/chassis/adapter/host-eth-if ## scope interrupt`
17. `server/chassis/adapter/host-eth-if/interrupt ## set interrupt-count 256`

18. server/chassis/adapter/host-eth-if/interrupt **## set interrupt-mode MSIx**
19. server/chassis/adapter/host-eth-if/interrupt **## commit**

DETAILED STEPS

	Command or Action	Purpose
Step 1	server # scope chassis	Enters chassis command mode.
Step 2	server/chassis # scope adapter <i>index_number</i>	Enters the command mode for the adapter card at the PCI slot number specified by <i>index_number</i> . Note Ensure that the server is powered on before you attempt to view or change adapter settings. To view the <i>index</i> of the adapters configured on you server, use the show adapter command.
Step 3	server/chassis/adapter # create host-eth-if <i>vNIC_name</i>	Creates a vNIC.
Step 4	server/chassis/adapter/host-eth-if ## set rocev2 enabled	Enables RoCEv2 on vNIC.
Step 5	server/chassis/adapter/host-eth-if ## set rdma-cos 5	Sets RDMA CoS 5 for RoCEv2 vNIC.
Step 6	server/chassis/adapter/host-eth-if ## set rdma_mr 131072	Sets RDMA Memory Region as 131072 for RoCEv2 vNIC.
Step 7	server/chassis/adapter/host-eth-if ## set rdma_qp 1024	Sets RDMA Queue Pairs as 1024 for RoCEv2 vNIC.
Step 8	server/chassis/adapter/host-eth-if ## set rdma_resgrp 8	Sets RDMA Resource Groups as 8 for RoCEv2 vNIC.
Step 9	server/chassis/adapter/host-eth-if ## scope comp-queue	Enters the Completion Queue command mode.
Step 10	server/chassis/adapter/host-eth-if/comp-queue ## set cq-count 2	Sets Completion Queue Count as 2 for vNIC.
Step 11	server/chassis/adapter/host-eth-if/comp-queue ## exit	Exits to host Ethernet interface command mode.
Step 12	server/chassis/adapter/host-eth-if ## scope trans-queue	Enters the Transmit Queue command mode.
Step 13	server/chassis/adapter/host-eth-if/trans-queue ## set wq-count 1	Sets Transmit Queue Count as 1 for vNIC.
Step 14	server/chassis/adapter/host-eth-if/trans-queue ## set wq-ring-size 256	Sets Transmit Queue Ring Buffer Size as 256 for vNIC.
Step 15	server/chassis/adapter/host-eth-if/trans-queue ## exit	Exits to host Ethernet interface command.
Step 16	server/chassis/adapter/host-eth-if ## scope interrupt	Enters Interrupt command mode.
Step 17	server/chassis/adapter/host-eth-if/interrupt ## set interrupt-count 256	Sets Interrupt Count as 256 for vNIC.
Step 18	server/chassis/adapter/host-eth-if/interrupt ## set interrupt-mode MSIx	Sets the Interrupt Mode as MSIx

	Command or Action	Purpose
Step 19	server/chassis/adapter/host-eth-if/interrupt *# commit	Commits the transaction to the system configuration. Note The changes take effect when the server is rebooted.

Example

```
server# scope chassis
server/chassis # scope adapter 1
server/chassis/adapter # create host-eth-if vNIC_Test
server/chassis/adapter/host-eth-if *# set rocev2 enabled
server/chassis/adapter/host-eth-if *# set rdma-cos 5
server/chassis/adapter/host-eth-if *# set rdma_mr 131072
server/chassis/adapter/host-eth-if *# set rdma_qp 1024
server/chassis/adapter/host-eth-if *# set rdma_resgrp 8
server/chassis/adapter/host-eth-if *# scope comp-queue
server/chassis/adapter/host-eth-if/comp-queue *# set cq-count 2
server/chassis/adapter/host-eth-if/comp-queue *# exit
server/chassis/adapter/host-eth-if *# scope trans-queue
server/chassis/adapter/host-eth-if/trans-queue *# set wq-count 1
server/chassis/adapter/host-eth-if/trans-queue *# set wq-ring-size 256
server/chassis/adapter/host-eth-if/trans-queue *# exit
server/chassis/adapter/host-eth-if *# scope interrupt
server/chassis/adapter/host-eth-if/interrupt *# set interrupt-count 256
server/chassis/adapter/host-eth-if/interrupt *# set interrupt-mode MSIx
server/chassis/adapter/host-eth-if/interrupt *# commit
```

Deleting RoCEv2 Interface Using Cisco IMC CLI

SUMMARY STEPS

1. server # **scope chassis**
2. server/chassis # **scope adapter** *index_number*
3. server/chassis/adapter # **scope host-eth-if** *vNIC_name*
4. server/chassis/adapter/host-eth-if # **set rocev2 disabled**
5. server/chassis/adapter/host-eth-if *# **commit**

DETAILED STEPS

	Command or Action	Purpose
Step 1	server # scope chassis	Enters the chassis command mode.
Step 2	server/chassis # scope adapter <i>index_number</i>	Enters the command mode for the adapter card at the PCI slot number specified by <i>index_number</i> .

	Command or Action	Purpose
		<p>Note Ensure that the server is powered on before you attempt to view or change adapter settings. To view the <i>index</i> of the adapters configured on your server, use the show adapter command.</p>
Step 3	server/chassis/adapter # scope host-eth-if vNIC_name	Enters the command mode for the vNIC specified by <i>vNIC_name</i> .
Step 4	server/chassis/adapter/host-eth-if # set rocev2 disabled	Disables RoCE properties on the vNIC.
Step 5	server/chassis/adapter/host-eth-if *# commit	Commits the transaction to the system configuration. <p>Note The changes take effect when the server is rebooted.</p>

Example

```
server# scope chassis
server/chassis # scope adapter 1
server/chassis/adapter # scope host-eth-if vNIC_Test
server/chassis/adapter/host-eth-if # set rocev2 disabled
server/chassis/adapter/host-eth-if *# commit
```



CHAPTER 6

Known Issues in RoCEv2

- [Known Limitations and Behavior, on page 39](#)

Known Limitations and Behavior

The following known issues are found in the RoCEv2 release.

Symptom	Conditions	Workaround
When sending high bandwidth NVMe traffic on some Cisco Nexus 9000 switches, the switch port that connected to the storage sometimes reaches the max PFC peak and does not automatically clear the buffers. In Nexus 9000 switches, the nxos command " show hardware internal buffer info pkt-stats input peak " shows that the <code>Peak_cell</code> or <code>PeakQos</code> value for the port reaches more than 1000.	The NVMe traffic will drop.	To recover the switch from this error mode. <ol style="list-style-type: none">1. Log into the switch.2. Locate the port that connected to the storage and shut down the port using "shutdown" command3. Execute the following commands one by one:<pre># clear counters # clear counter buffers module 1 # clear qos statistics</pre>4. Run no shutdown on the port that was shut down.

Symptom	Conditions	Workaround
<p>On VIC 1400 Series adapters, the neNIC driver for Windows 2019 can be installed on Windows 2016 and the Windows 2016 driver can be installed on Windows 2019. However, this is an unsupported configuration.</p>	<p>Case 1 : Installing Windows 2019 nenic driver on Windows 2016 succeeds-but on Windows 2016 RDMA is not supported.</p> <p>Case 2 : Installing Windows 2016 nenic driver on Windows 2019 succeeds-but on Windows 2019 RDMA comes with default disabled state, instead of enabled state.</p>	<p>The driver binaries for Windows 2016 and Windows 2019 are in folders that are named accordingly. Install the correct binary on the platform that is being built/upgraded.</p>