Configuring NVMeoF with RoCEv2 in ESXi

Guidelines for using NVMe over Fabrics (NVMeoF) with RoCE v2 on ESXi

General Guidelines and Limitations:

- Cisco recommends you to check the UCS Hardware and Software Compatibility to determine support for NVMeoF. NVMeoF is supported on Cisco UCS B-Series, C-Series, and X-Series servers.
- Nonvolatile Memory Express (NVMe) over RDMA with RoCE v2 is currently supported only with Cisco VIC 15000 Series adapters.
- When creating RoCE v2 interfaces, use Cisco recommended Queue Pairs, Memory Regions, Resource Groups, and Class of Service settings. NVMeoF functionality may not be guaranteed with different settings for Queue Pairs, Memory Regions, Resource Groups, and Class of Service.
- RoCE v2 supports maximum two RoCE v2 enabled interfaces per adapter.
- Booting from an NVMeoF namespace is not supported.
- Layer 3 routing is not supported.
- Saving a crashdump to an NVMeoF namespace during a system crash is not supported.
- NVMeoF cannot be used with usNIC, VxLAN, VMQ, VMMQ, NVGRE, GENEVE Offload, ENS, and DPDK features.
- Cisco Intersight does not support fabric failover for vNICs with RoCE v2 enabled.
- The Quality of Service (QoS) no drop class configuration must be properly configured on upstream switches such as Cisco Nexus 9000 series switches. QoS configurations will vary between different upstream switches.

- During the failover or failback event, the Spanning Tree Protocol (STP) can result temporary loss of network connectivity. To prevent this connectivity issue, disable STP on uplink switches.

Downgrade Guidelines: Remove the RoCEv2 configuration first and then downgrade to the release version lower than Cisco UCS Manager release 4.2(3b) version.

ESXi Requirements

Configuration and use of RoCE v2 in ESXi requires the following:

- VMWare ESXi version 7.0 Update 3.
- Cisco UCS Manager Release 4.2(3b) or later versions.
- VIC firmware 5.2(3x) or later versions.
- The driver version, *nenic-2.0.4.0-1OEM.700.1.0.15843807.x86_64.vib* that provides both standard eNIC and RDMA support with the Cisco UCS Manager 4.2(3b) release package.
- A storage array that supports NVMeoF connection.

Configuring RoCE v2 for NVMeoF on Cisco Intersight

Use these steps to configure the RoCE v2 interface on Cisco Intersight.

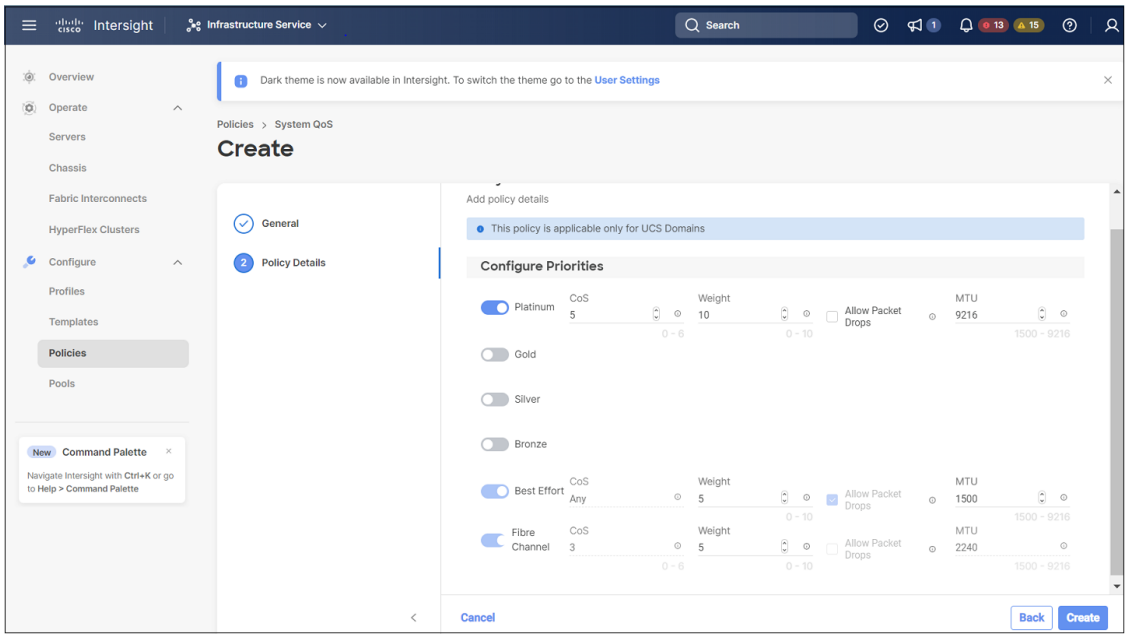
To avoid possible RDMA packet drops, ensure same no-drop COS is configured across the network. The following steps allows you to configure a no-drop class in System QoS policies and use it for RDMA supported interfaces.

Procedure

Step 1 Navigate to **CONFIGURE > Policies**. Click **Create Policy**, select **UCS Domain** platform type, search or choose **System QoS**, and click **Start**.

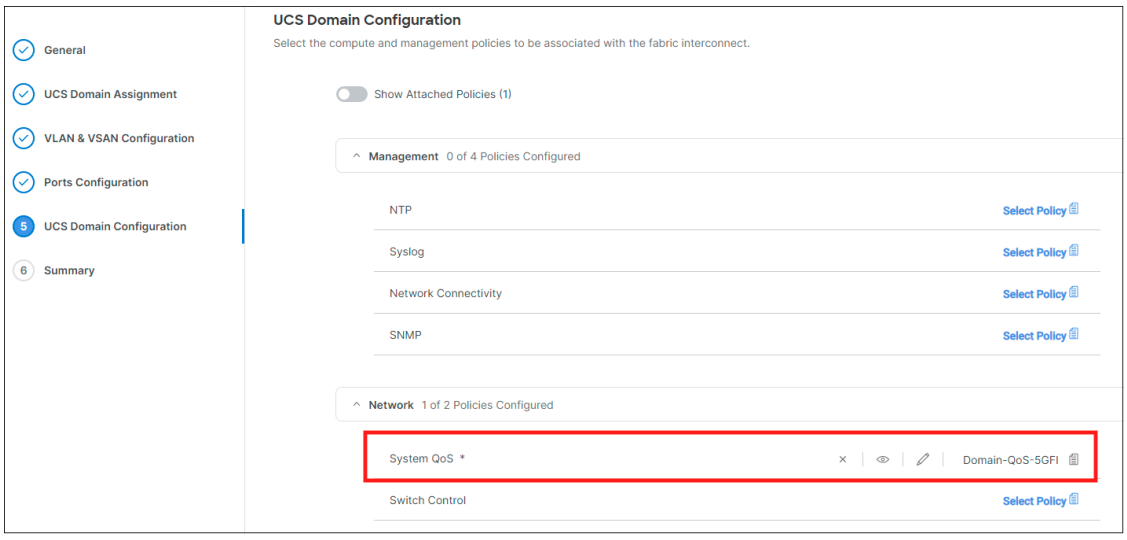
Step 2 In the **General** page, enter the policy name and click **Next**, and then in the **Policy Details** page, configure the property setting for System QoS policy as follows:

- For **Priority**, choose **Platinum**
- For **Allow Packet Drops**, uncheck the check box.
- For **MTU**, set the value as **9216**.



Step 3 Click **Create**.

Step 4 Associate the System QoS policy to the Domain Profile.



Note
For more information, see *Creating System QoS Policy* in Configuring Domain Policies and Configuring Domain Profiles.

The System QoS Policy is successfully created and deployed to the Domain Profile.

What to do next

Configure the server profile with RoCE v2 vNIC settings in LAN Connectivity policy.

Enabling RoCE Settings in LAN Connectivity Policy

Use the following steps to configure the RoCE v2 vNIC. In Cisco Intersight LAN Connectivity policy, you can enable the RoCE settings on **Ethernet Adapter policy** for Linux configuration as follows:

Procedure

Step 1 Navigate to **CONFIGURE > Policies**. Click **Create Policy**, select **UCS Server** platform type, search or choose **LAN Connectivity policy**, and click **Start**.

Step 2 In the policy **General** page, enter the policy name, select the Target Platform as **UCS Server (Standalone)** or **UCS Server (FI-Attached)**, and click **Next**.

Step 3 In the **Policy Details** page, click **Add vNIC** to create a new vNIC.

Step 4 In the **Add vNIC** page, follow the configuration parameters to enable the RoCE v2 vNIC:

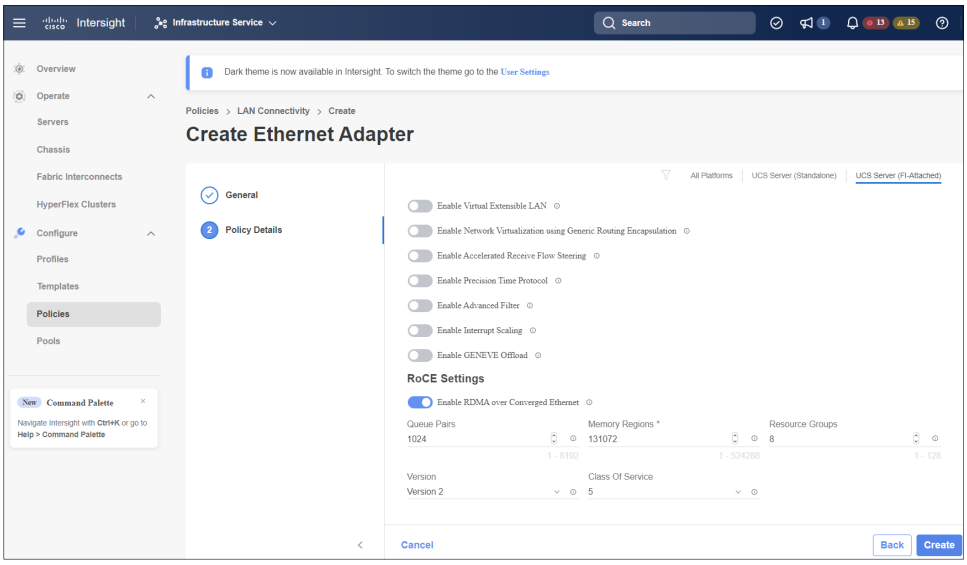
a) In the **General** section, provide a name for virtual ethernet interface.

b) Incase of a Standalone server, click the **Consistent Device Naming (CDN)** or click the **Failover** of a FI-attached server , and do the following:

- Click **Select Policy** under **Ethernet Adapter**.
- In the **Select Policy** window, click **Create New** to create an Ethernet Adapter policy.
- In the **General** page of the Ethernet Adapter Policy, enter the policy name and click **Next**.
- In the **Policy Details** page of the Ethernet Adapter Policy, modify the following property setting:
  - **RoCE Settings**
    - For **Enable RDMA over Converged Ethernet**, slide to enable and set the RoCE on this virtual interface.
    - For **Queue Pairs**, select or enter **1024**
    - For **Memory Regions**, select or enter **131072**
    - For **Resource Groups**, select or enter **8**
    - For **Version**, select **Version 2**
    - For **Class of Service**, select **5**
  - **Interrupt Settings**
    - For **Interrupts**, select or enter **256**.
    - For **Interrupt mode**, select **MSIx**.
    - For **Interrupt Timer, us,** select **125**.
    - For **Interrupt Coalescing Type**, select **Min**.
  - **Receive** Settings
    - For **Receive Queue Count**, select or enter **1**.
    - For **Receiving Ring Size**, select or enter **512**.

- **Transmit** Settings
  - For **Transmit Queue Count**, select or enter **1**.
  - For **Transmit Ring Size**, select or enter **256**.
- **Completion** Settings
  - For **Completion Queue Count**, select or enter **2**.
  - For **Completion Ring Size**, select or enter **1**.
  - For **Uplink Failback Timeout(seconds)**, select or enter **5**

- Click **Create** to create an Ethernet Adapter Policy with the above defined settings.



- Click **Add** to save the setting and add the new vNIC.

Note
All the fields with * are mandatory and ensure it is filled out or selected with appropriate policies.

Step 5 Click **Create** to complete the LAN Connectivity policy with RoCE v2 settings.

Step 6 Associate the LAN Connectivity policy to the Server Profile.

Note
For more information, see *Creating a LAN Connectivity Policy* and *Creating an Ethernet Adapter Policy* in Configuring UCS Server Policies and Configuring UCS Server Profiles.

The LAN Connectivity Policy with the Ethernet Adapter policy vNIC setting is successfully created and deployed to enable RoCE v2 configuration.

What to do next

Once the policy configuration for RoCE v2 is complete, configure RoCE v2 for NVMeoF on the Host System.

NENIC Driver Installation

Before you begin

The Ethernet Network Interface Card (eNIC) Remote Direct Memory Access (RDMA) driver requires nenic driver.

Procedure

Step 1 Copy the eNIC vSphere Installation Bundle (VIB) or offline bundle to the ESXi server.

Step 2 Use the command to install nenic driver:

```
esxcli software vib install -v {VIBFILE}
or
esxcli software vib install -d {OFFLINE_BUNDLE}
```

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 installed, you may need to put ESX into maintenance mode. This can be done through the client, or by adding the *--maintenance-mode* option to the above *esxcli*.

What to do next

Configure the Host side for ESXi NVMe RDMA.

ESXi NVMe RDMA Host Side Configuration

NENIC RDMA Functionality

One of the major difference between RDMA on Linux and ESXi is listed below:

- In ESXi, the physical interface (vmnic) MAC is not used for RoCEv2 traffic. Instead, the VMkernel port (vmk) MAC is used.

Outgoing RoCE packets use the vmk MAC in the 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.

- In Linux, the physical interface MAC is used in source MAC address field in the ROCE packets. This Linux MAC is usually a Cisco MAC address configured to the VNIC using UCS Manager.

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

Ensure you have nenic drivers. Download and install nenic drivers before proceeding with below steps:

Procedure

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.

a) Under **Assigned Adapters**, select **New Adapters**.
b) 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.
c) (Optional) Use the up and down arrow keys to change the position of the adapter in the Assigned Adapters list.
d) 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.

a) Enter a label that represents the traffic type for the VMkernel adapter.
b) Set a VLAN ID to identify the VLAN the VMkernel uses for routing network traffic.
c) Select IPV4 or IPV6 or both.
d) 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.

e) After setting the TCP/IP stack for the VMkernel adapter, select a TCP/IP stack.

   To use the default TCP/IP stack, select it from the available services.

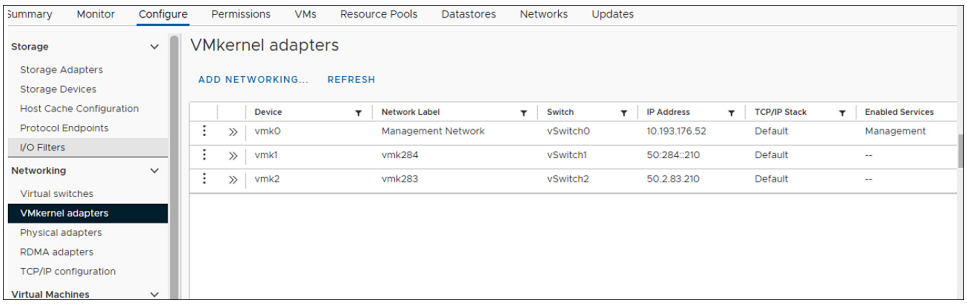
   Note
   Be aware that the TCP/IP stack for the VMkernel adapter cannot be changed later.

f) Configure IPV4 and/or IPV6 settings.

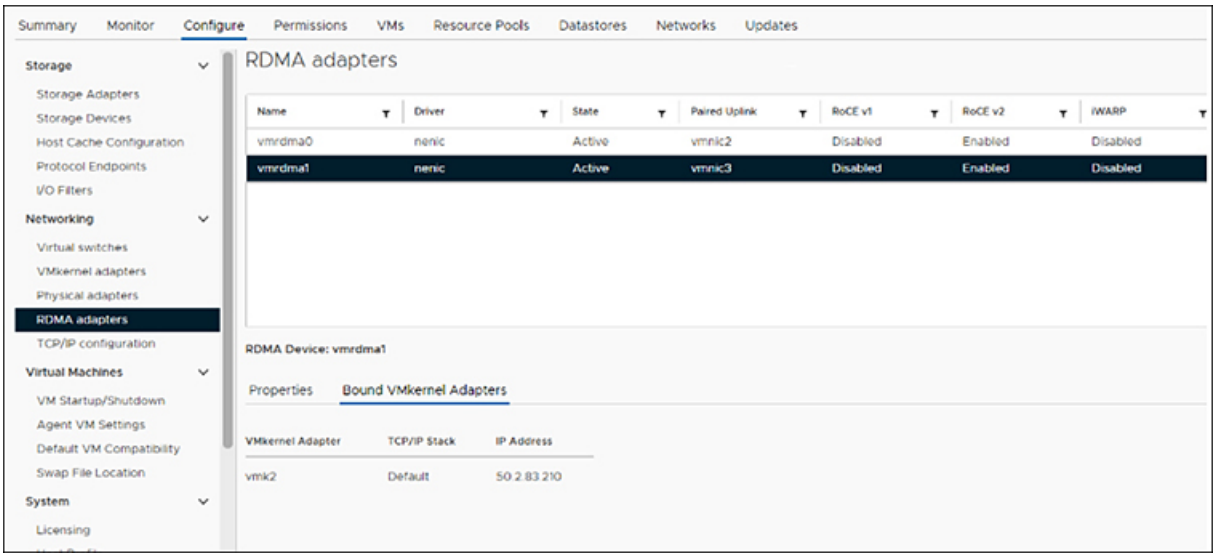
Step 8 On the **Ready to Complete** page, click **Finish**.

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.



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



What to do next

Create vmhba ports on top of vmrdma ports.

Create VMVHBA Ports in ESXi

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

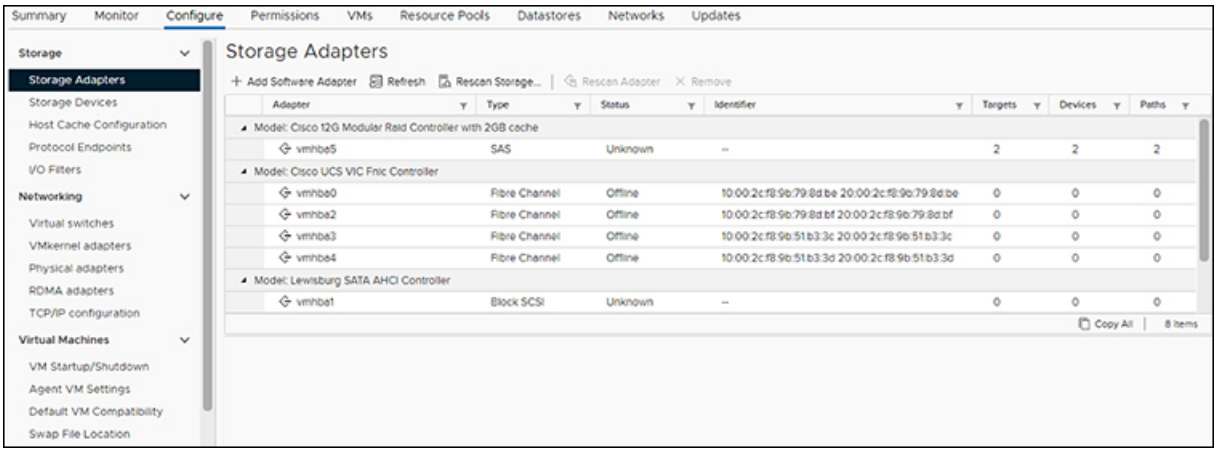
Before you begin

Create the adapter ports for storage connectivity.

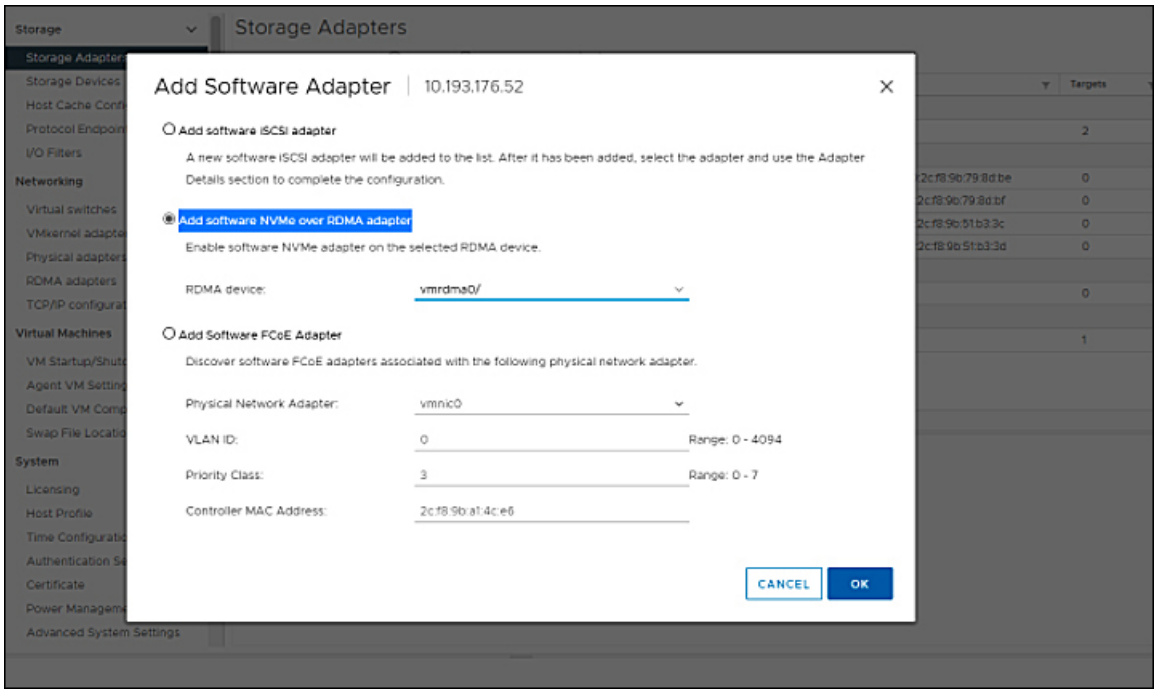
Procedure

Step 1 Go to vCenter where your ESXi host is connected.

Step 2 Click on **Host**>**Configure**>**Storage adapters**.



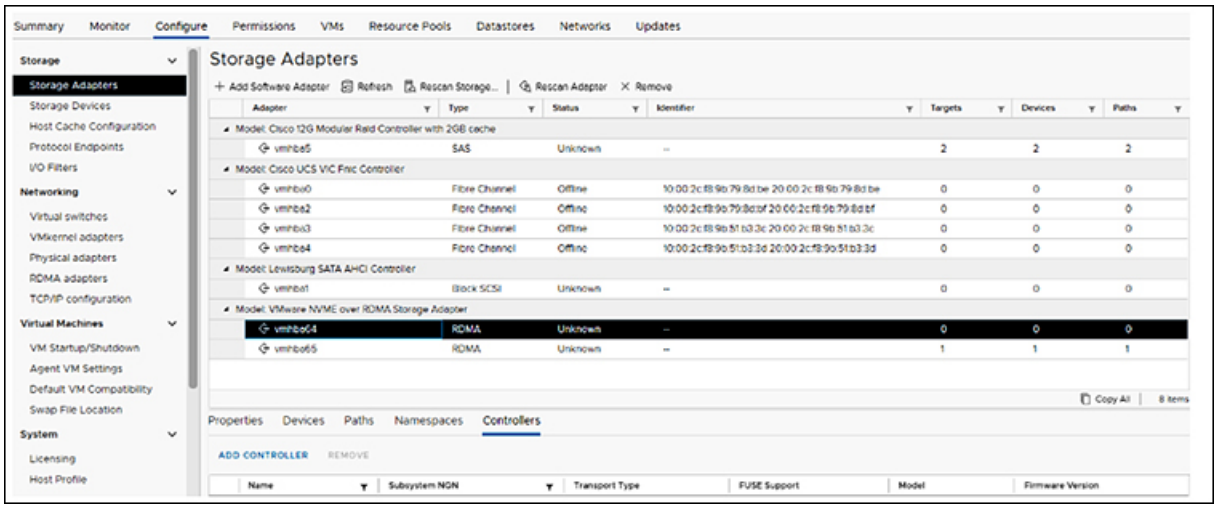
Step 3 Click **+Add Software Adapter**. The following dialog box will appear.



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 will be shown as in the example below



Displaying vmnic and vmrdma Interfaces

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

Before you begin

Create Network Adapters and VHBA ports.

Procedure

Step 1 Use **ssh** to access 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 Use **esxcli rdma device list** to list the vmrdma devices. When the enic driver registers with ESXi the RDMA device for a RDMA capable VNIC, ESXi creates a vmrdma device and links it to the corresponding vmnic.

```
[root@S        RackServer:~] esxcli rdma device list
Name     Driver  State    MTU  Speed    Paired Uplink  Description
-------  ------  ------  ----  -------  -------------  -----------
vmrdma0  nenic   Active  4096  50 Gbps  vmnic1         Cisco UCS VIC 15XXX (A0)
vmrdma1  nenic   Active  4096  50 Gbps  vmnic2         Cisco UCS VIC 15XXX (A0)
[root@S        RackServer:~] esxcli rdma device vmknic list
Device   Vmknic  NetStack
-------  ------  --------
vmrdma0  vmk1    defaultTcpipStack
vmrdma1  vmk2    defaultTcpipStack
```

Step 4 Use **esxcli rdma device protocol list** to check the protocols supported by the vmrdma interface.

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

```
[root@ESXi7U3      :~] esxcli rdma device protocol list
Device    RoCE v1  RoCE v2  iWARP
--------  -------  -------  -----
vmrdma0     false     true  false
vmrdma1     false     true  false
[root@ESXi7U3      :~]
```

Step 5 Use **esxcli nvme adapter list** to list the NVMe adapters and the vmrdma and vmnic interfaces it is configured on.

```
[root@ESXi7U3      :~] 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@ESXi7U3      :~]
```

Step 6 All vmhbas in the system can be listed using **esxcli storage core adapter list**. The vmhba configured over RDMA.

```
[root@ESXi7U3      :~] 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                                       (0000:3c:00.0) Broadcom Cisco 12G Modular Raid Controller with 2GB cache
vmhba64   nvmerdma  link-n/a    rdma.vmnic2:2c:f8:9b:79:8d:bc                              VMware NVMe over RDMA Storage Adapter on vmrdma0
vmhba65   nvmerdma  link-n/a    rdma.vmnic3:2c:f8:9b:79:8d:bd                              VMware NVMe over RDMA Storage Adapter on vmrdma1
[root@ESXi7U3      :~]
```

Note
For vmhba64 and vmhba65, you may observe that the driver's Link State displays *link-n/a* instead of *Online*. This is a known issue in ESXi 7.0 Update 3. For more information, see Known Issues - ESXi.

NVMe Fabrics and Namespace Discovery

This procedure is performed through the ESXi command line interface.

Before you begin

Create and configure NVMe on the adapter's VMHBAs. The maximum number of adapters is two, and it is a best practice to configure both for fault tolerance.

Procedure

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 fabric 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.

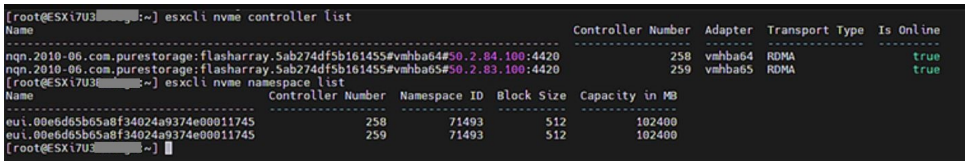
`esxcli nvme fabrics discover -a vmhba64 -l transport_address p Transport Service ID -s Subsystem NQN`

Step 4 Repeat steps 1 through 4 to configure the second adapter.

Step 5 Verify the configuration.

a) Display the controller list to verify the NVMe controller is present and operating.

`esxcli nvme controller list RDMA -d vmrdma0`



b) Verify that the fabric is enabled on the controller through the adapter, and verify the controller is accessible through the port on the adapter.

```
[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 RoCE v2 Interface Using Cisco Intersight

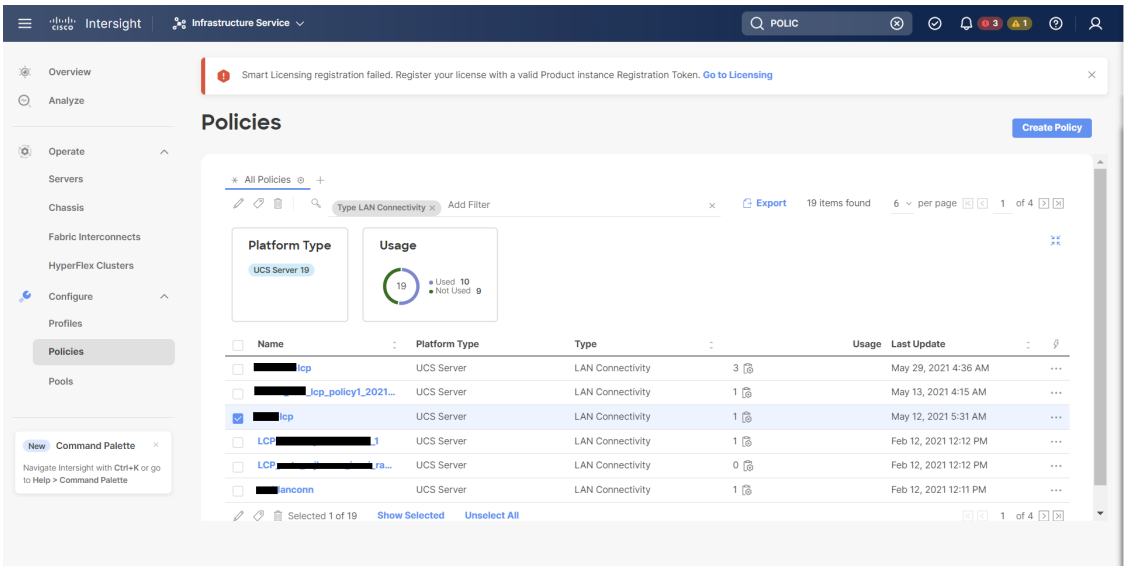
Use these steps to remove the RoCE v2 interface.

Procedure

Step 1 Navigate to **CONFIGURE > Policies**. In the **Add Filter** field, select **Type: LAN Connectivity**.

Step 2 Select the appropriate LAN Connectivity policy created for RoCE V2 configuration and use the delete icon on the top or bottom of the policy list.

Step 3 Click **Delete** to delete the policy.



Step 4 Upon deleting the RoCE v2 configuration, re-deploy the server profile and reboot the server.