Configuring NVMe with ROCE V2 in ESXi

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

General Guidelines and Limitations:

- Cisco UCS Manager release 4.2(3b) supports RoCEv2 only on ESXi 7.0 U3.
- Cisco recommends you check UCS Hardware and Software Compatibility specific to your UCS Manager release to determine support for ESXi. RoCEv2 on ESXi is supported on UCS B-Series and C-Series servers with Cisco UCS VIC 15000 Series and later adapters.
- RoCEv2 on ESXi is not supported on UCS VIC 1200, 1300 and 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.
- The COS setting is not supported on UCS Manager.

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.
- UCS Manager release 4.2.3 or later
- 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.

Procedure

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.

What to do next

Create and configure the Adapter Policy for ESXi NVMe RDMA in UCS Manager.

Configuring and Enabling RoCEv2 on UCS Manager

Configuring RoCEv2 for VMware NVMeoF on UCS Manager

UCS Manager contains a default adapter policy that is prepopulated with operational parameters, so you do not need to manually create the adapter policy. However, you do need to create the RoCEv2 interface.

Use these steps to configure the RoCEv2 interface on UCS Manager.

Procedure

Step 1 In the **Navigation** pane, click **Servers**.

Step 2 Expand **Servers** > **Service Profiles**.

Step 3 Expand the node for the organization where you want to create the policy.

If the system does not include multitenancy, expand the **root** node.

Step 4 Click on a RDMA service profile you created and expand the service profile.

Step 5 Right-click on **vNICs** and choose **Create vNIC** to create a new vNIC.

Step 6 Click on a RDMA service profile you created with the Service Policy and scroll down to **vNICs**. Right-click and choose **Create** to create a new vNIC.

The **Create VNIC** pop-up menu is displayed.

Perform the below steps to modify the vNIC policy:

a) Name the new VNIC.

b) On the **MAC address** dropdown, select the desired address or use the default in the dropdown.
c) Select which VLAN you want use use from the list.
d) In the Adapter Performance Profile, select the default adapter policy named `VMWareNVMeRoCEv2`.
e) Click **OK**. The interface is now configured for one port.

Step 7 Click **Save Changes**.

Step 8 Select **Reboot**.

What to do next

Configure the Host side for ESXi NVMe RDMA.

ESXi NVMe RDMA Host Side Configuration

NENIC RDMA Functionality

One major difference exists between the use case for RDMA on Linux and ESXi.

- In ESXi, the physical interface (vmnic) MAC is not used fo 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 that you have downloaded and installed the enic drivers.

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 trafic.

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.

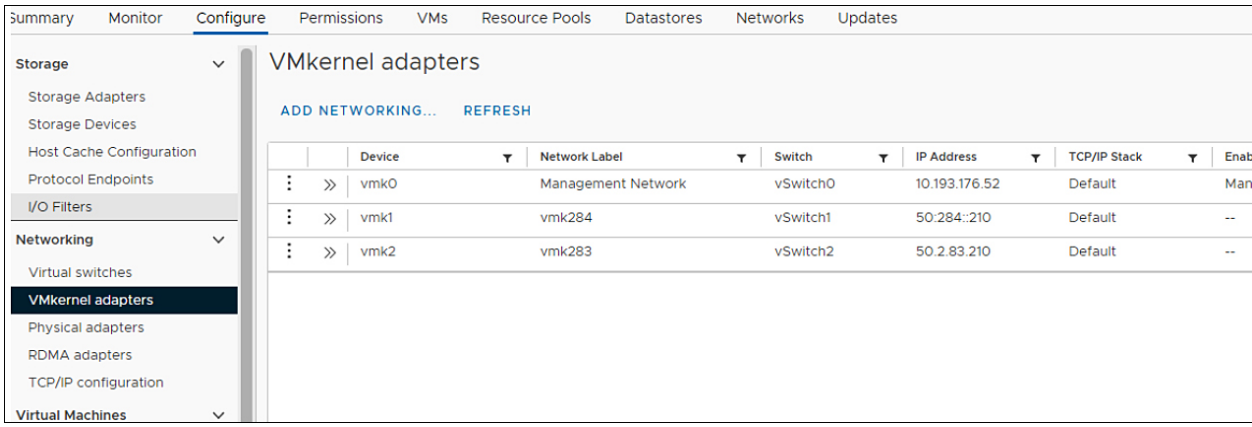
f) Configure IPV4 and/or IPV6 settings.

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

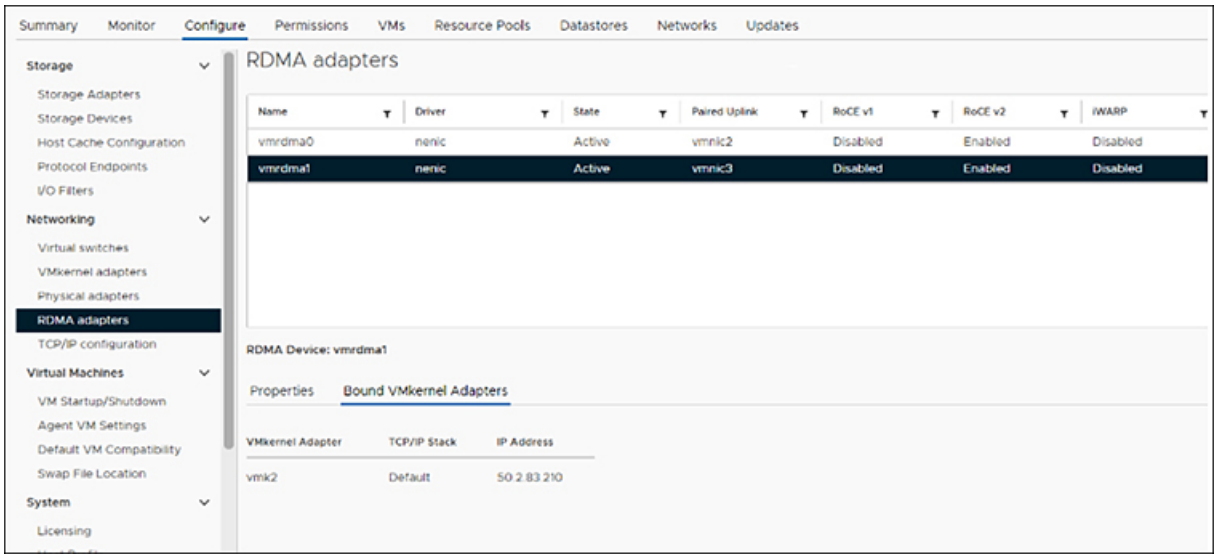
Step 9 Check the VMkernel ports for the VM Adapters or port groups with NVMe RDMA in the vSphere client, as shown in the Results below.

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

Example



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 VMHBA Ports in ESXi

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

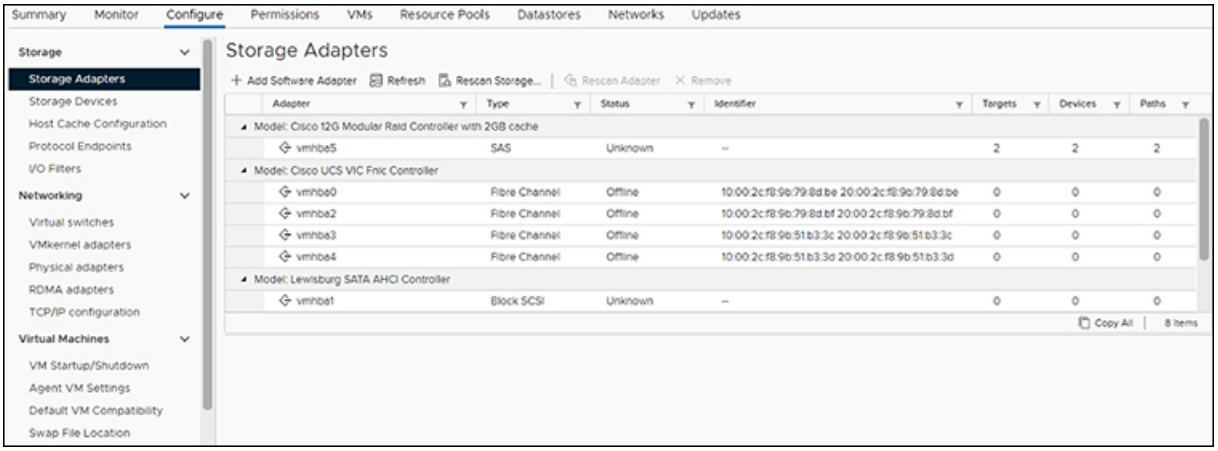
Before you begin

Create the adapter ports for storage connectivity.

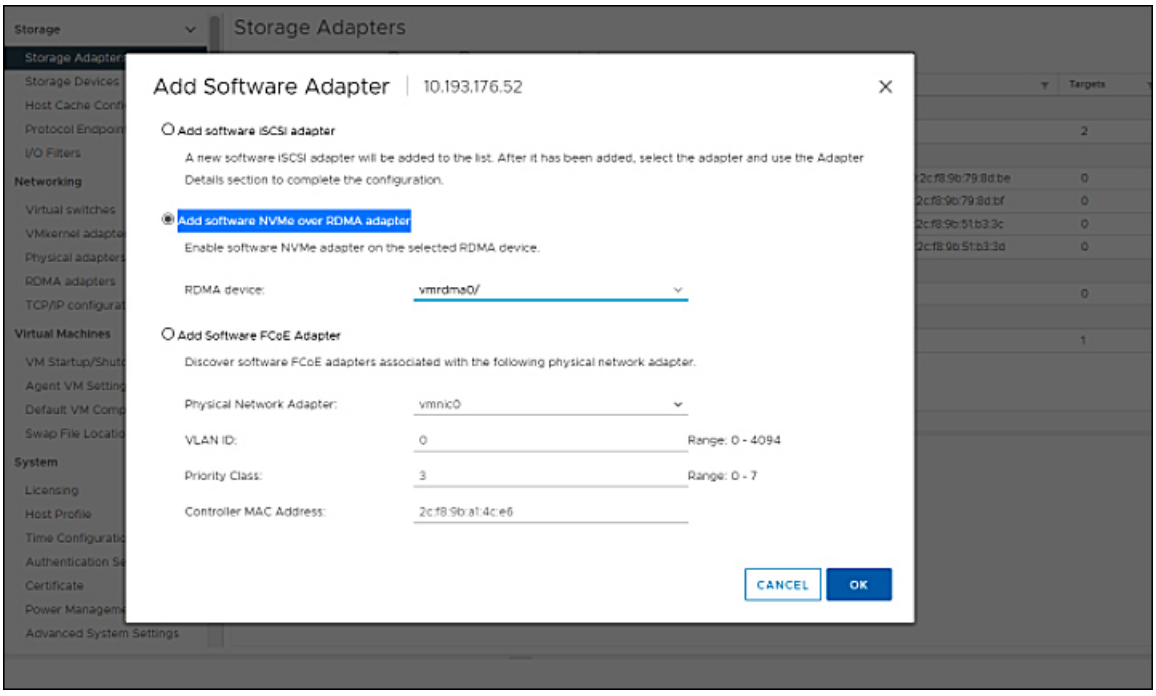
Procedure

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

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



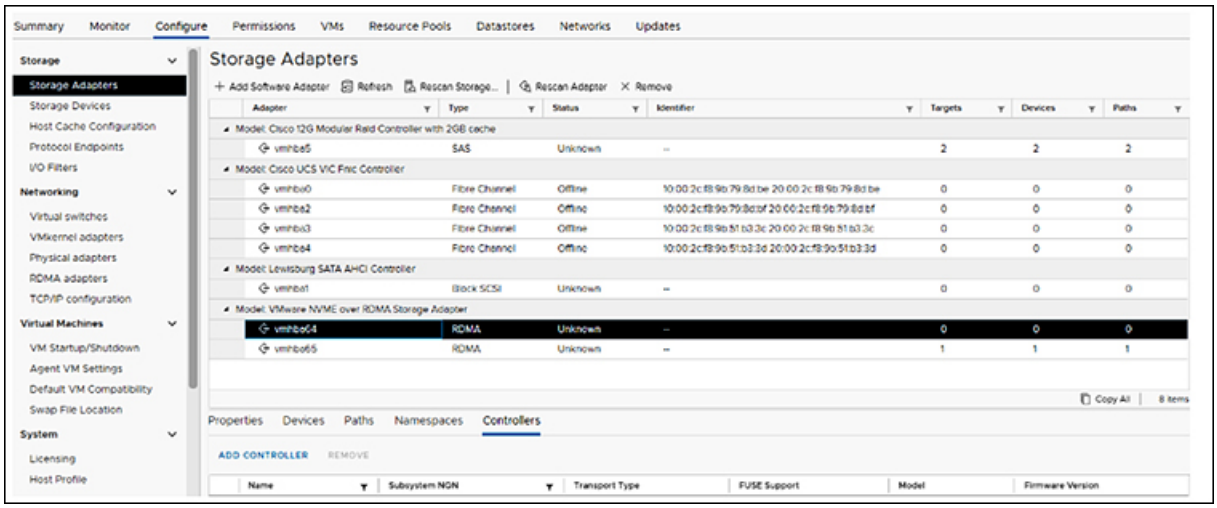
Step 3 Click **+Add Software Adapter**. The following dialog box will appear.



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



What to do next

Configure NVME.

Displaying vmnic and vmrdma Interfaces

ESXi creates a vmnic interface for each enic 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
```

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.

Step 3 Use **esxcli rdma device list** to list the vmrdma devices.

```
[root@StockholmRackServer:~] 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@StockholmRackServer:~] esxcli rdma device vmknic list
Device   Vmknic  NetStack
-------  ------  --------
vmrdma0  vmk1    defaultTcpipStack
vmrdma1  vmk2    defaultTcpipStack
```

Step 4 Use **esxcli rdma device list** to check the protocols supported by the vmrdma 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 vmrdma 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
-------  -------  -------  -----
vmrdma0    false     true  false
vmrdma1    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                                                (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@ESXi7U3Bodega:~]
```

What to do next

Configure NVME.

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.

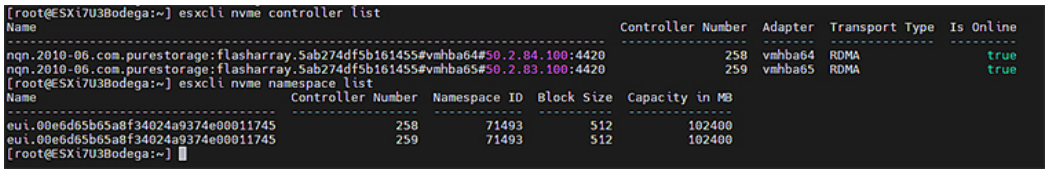
`esxcli nvme fabrics discover -a vmhba64 -l transport_address p Transport Service ID -s Subsystem NQN`

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 ESXi RoCEv2 Interface Using UCS Manager

Use these steps to remove the RoCE v2 interface for a specific port.

Procedure

Step 1 In the **Navigation** pane, click **Servers**.

Step 2 Expand **Servers** > **Service Profiles**.

Step 3 Expand the node for the profile to delete.

Step 4 Click on **vNICs** and select the desired interface. Right click and select **Delete** from the dropdown.

Step 5 Click **Save Changes**.