



connect • monitor • manage

CONNECT - TECH NOTE

Enable RoCE Capabilities: Linux OFED with Emulex OCe14000 Network Adapters

Validate Priority Flow Control and
Network File System over RDMA



 **EMULEX**[®]

Introduction

Emulex provides RDMA over Converged Ethernet (RoCE) driver support for RedHat 6.4 and SLES 11 SP2. The Emulex OneConnect® OCe14000 family of 10Gb and 40Gb Ethernet (10GbE and 40GbE) Network Adapters and Converged Network Adapters (CNAs) support RoCE offloads to support Remote Directory Memory Access (RDMA) data transfer between hosts. RoCE is the protocol that implements RDMA operations on an Ethernet network. It is only concerned with the Ethernet level so OpenFabrics Infiniband Distribution (OFED) software is necessary for the operation to continue on, for it contains the drivers and protocols for RDMA and kernel bypass. This tech note presents the necessary hardware, software and installation of OFED and required drivers needed by the Emulex OCe14000 Network Adapters for enabling RoCE capabilities. The guide also lists the steps for testing the NFS (Network File System) application over RoCE and the switch configuration for Priority Flow Control (PFC).

Note – This tech note, however, does not give any details on performance matrix for RoCE.

The proof of concept (POC) is divided into two different working configurations.

Configuration 1 – Two hosts connected to 10Gb per second (10Gbps) networking switch. This is scalable solution for when more hosts may be desired to be added.

Configuration 2 – Two hosts connected back to back. This is a solution when a switch is not available for the infrastructure.

Figure 1 shows a diagram illustrating the network topology for two hosts connected to a 10Gbps networking switch.



Figure 1.

Figure 2 shows a diagram illustrating the network topology for two hosts connected back to back.



Figure 2.

Hardware requirements

Hardware components	Quantity	Description	POC components
Server	2	Any server which supports Linux OSes	HP DL 380p Gen8 with Intel Xeon CPU E5-2690 @ 2.90GHz
Hard drives	2	Any SAS or SATA drive	SAS 300GB Drives
RAID Controller	2	Any server which support Linux OSes	
RAM	12-96GB/server		96GB RAM/Server
OCe14000 Network Adapter	2	Emulex OCe14000 Network Adapter	Emulex OCe14102 Network Adapter
Switch	1	10Gbps switch with PFC	10Gbps Cisco Nexus 5548P
Cables	2	10Gbps optical SFP+ cables	10Gbps optical SFP+ cables

Software requirements

Supported Operating systems and OFED versions

Component	Quantity	Description		POC components
Supported operating system	2	RHEL 6.4	SLES 11 SP2	Red Hat Enterprise Linux 6.4 (RHEL 6.4)
OFED	2	OFED 3.5-1	OFED 3.5	OFED 3.5-1
		Open source software for RDMA and kernel bypass		
Firmware	2	Latest Firmware		V 10.2.354.1
NIC driver	2	NIC Ethernet driver support for Linux		Emulex provided NIC driver: RHEL 6.4
OneCommand® Manager	2	NIC managing tool for both systems		OneCommand Manager: RHEL 6.4
RoCE driver	2	Latest Emulex RoCE driver		Emulex provided RoCE driver: RHEL 6.4

Pre-installation

Before applying power and installing the software there are various logistical steps that must be taken into account.

1. Ensure that there is proper rack space for the hardware components.
2. Ensure that there is adequate power and cooling available for the systems.
3. Determine whether the systems are at the latest firmware levels.
4. Determine whether the Network Interface Card (NIC) needs driver upgrading.

For the latest firmware, drivers and instructions to install them, go to www.emulex.com/downloads.

For other components, such as LAN on Motherboard (LOM) and integrated management ports, and latest firmware and driver, please refer to the server vendor website or user's manual.

Post-installation

After completing the installation, determine which host will act as the server and which one will act as the client, before moving onto specific configuration. For this POC the NFS server will be called **roce1** and the client **roce2**. Make sure the link is “up” between the switch and systems for **Configuration 1** or between the adapters for **Configuration 2**. In addition, the RoCE profile should be enabled on the NFS server and client. For this POC, RoCE1 profile was selected.

The RoCE profile can be enabled by using PXESelect BIOS or the OneCommand Manager application.

- To configure the adapter using PXESelect BIOS, refer to the **Boot for NIC, iSCSI, Fibre Channel over Ethernet (FCoE), and RoCE Protocols User Manual** for more information on the PXESelect BIOS utility.
- To configure the adapter using the OneCommand Manager shown in Figure 3, refer to the **OneCommand Manager Application User Manual**, or the **OneCommand Manager Command Line Interface User Manual**.

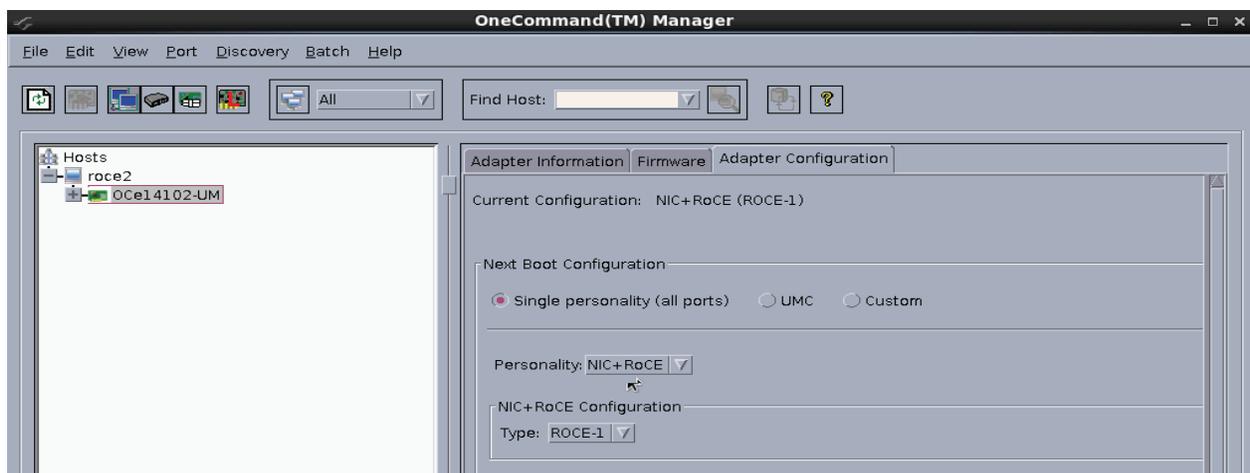


Figure 3.

Configuration 1

Implementation/Configuration using Emulex OCE14102 adapter with a switch

This POC was created using the Cisco Nexus switch to set up the basic network topology displayed in Figure 1. The key objective is to describe the configuration and validation of NFS over RoCE.

Step 1 – Create virtual LAN (VLAN) on Cisco Nexus 5548p switch

This step covers VLAN creation/configuration, as well as port configuration on the switch.

All Cisco switch configuration commands can be found in the product user manual, if you wish to look further into the switches functionality.

- a) Enter configuration mode by typing “configure terminal.”

```
# configure terminal
```

- b) Create a VLAN using the “vlan <VLAN ID>” command.

Note – The VLAN ID should be the same as the ones on the NFS server and clients. For this POC, the VLAN ID is 10.

```
# vlan 10
```

- c) Get back to base configuration mode from the VLAN mode using the “exit” command.

Note – The “exit” command gets you to the previous configuration mode/sub-mode.

```
# exit
```

- d) Enter configuration mode for a connected port using the “interface <type> <slot/port>” command.

Note – For this POC the type is ethernet, the slot/ports used are 1/15 and 1/16.

```
# interface ethernet 1/15
```

- e) Set the interface as a trunk port using the “switchport mode <type>” command.

```
# switchport mode trunk
```

- f) Allow the VLAN on the trunk port using the “switchport trunk allowed vlan <VLAN ID>” command.

```
# switchport trunk allowed vlan 10
```

- g) Select the PFC mode using the “priority-flow-control <auto/on>” command. CHOOSE “auto”. Then get back to base configuration mode using the “exit” command.

```
# priority-flow-control mode auto
```

```
# exit
```

- h) Repeat steps **d** through **f** for the other port.

- i) Enter the VLAN configuration sub mode using the “vlan <VLAN ID>” command.

```
# vlan 10
```

- j) Enable the VLAN using the “no shutdown” command. Then get back to base configuration mode using the “exit” command.

```
# no shutdown
```

```
# exit
```

Step 2 – Configure Priority Flow Control (PFC) on the switch

This step enlists the detailed steps for configuring PFC for RoCE traffic on the switch.

- a) Create a priority group for RoCE traffic with a priority of 5. There are several commands needed to accomplish this task. Type the following commands in the order shown.

```
# class-map type qos roce
# match cos 5
# exit
# class-map type queuing roce
# match qos-group 5
# exit
# class-map type network-qos roce
# match qos-group 5
# exit
```

- b) Assign the quality of service (QoS) group for the different types of traffic. Enter into QoS policy map configuration mode for **RoCE** using the “policy-map type <mode> <group>” command. Type the following commands in the order shown.

```
# policy-map type qos roce
# class roce
# set qos-group 5
# exit
# class class-fcoe
# set qos-group 1
# exit
# class class-default
# exit
# exit
```

- c) Allocate the appropriate bandwidth for the types of traffic. Enter into queuing policy map configuration mode for **RoCE** using the “policy-map type <mode> <group>” command. Type the commands in the order shown.

```
# policy-map type queuing roce
# class type queuing roce
# bandwidth percent 50
# exit
# class type queuing class-fcoe
# bandwidth percent 40
# exit
# class type queuing class-default
# bandwidth percent 10
# exit
# exit
```

- d) Set the Maximum Transition Unit (MTU) for the separate types of traffic. Enter into the network policy map configuration mode for **roce** using the “policy-map type <mode> <group>” command. Type the following commands in the order shown.

Note – Current supported Maximum MTU for OCe14102 Adapter for RoCE is 4096.

```
# policy-map type network-qos roce
# class type network-qos roce
# pause no-drop
# mtu 5000
# class type network-qos class-default
# mtu 9216
# class type network-qos class-foce
# pause no-drop
# mtu 2158
# exit
# exit
```

- e) Configure the switches service policies. Enter into the system QoS configuration mode for the switch using the “system <mode>” command. Type the commands in the following order.

```
# system qos
# service-policy type qos input roce
# service-policy type queuing input roce
# service-policy type queuing output roce
# service-policy type network-qos roce
# exit
```

- f) Save the running configuration.

```
# copy running-config startup-config
```

Figure 4 shows the entire PFC set up process in Terminal.

```

Cisco5548p-SJ(config)# class-map type qos roce
Cisco5548p-SJ(config-cmap-qos)# match cos 5
Cisco5548p-SJ(config-cmap-qos)# exit
Cisco5548p-SJ(config)# class-map type queuing roce
Cisco5548p-SJ(config-cmap-que)# match qos-group 5
Cisco5548p-SJ(config-cmap-que)# exit
Cisco5548p-SJ(config)# class-map type network-qos roce
Cisco5548p-SJ(config-cmap-nq)# match qos-group 5
Cisco5548p-SJ(config-cmap-nq)# exit
Cisco5548p-SJ(config)# policy-map type qos roce
Cisco5548p-SJ(config-pmap-qos)# class roce
Cisco5548p-SJ(config-pmap-c-qos)# set qos-group 5
Cisco5548p-SJ(config-pmap-c-qos)# exit
Cisco5548p-SJ(config-pmap-qos)# class class-fcoe
Cisco5548p-SJ(config-pmap-c-qos)# set qos-group 1
Cisco5548p-SJ(config-pmap-c-qos)# exit
Cisco5548p-SJ(config-pmap-qos)# class class-default
Cisco5548p-SJ(config-pmap-c-qos)# exit
Cisco5548p-SJ(config-pmap-qos)# exit

Cisco5548p-SJ(config)# policy-map type network-qos roce
Cisco5548p-SJ(config-pmap-nq)# class type network-qos roce
Cisco5548p-SJ(config-pmap-nq-c)# pause no-drop
Cisco5548p-SJ(config-pmap-nq-c)# mtu 4200
Cisco5548p-SJ(config-pmap-nq-c)# class type network-qos class-default
Cisco5548p-SJ(config-pmap-nq-c)# mtu 9216
Cisco5548p-SJ(config-pmap-nq-c)# class type network-qos class-fcoe
Cisco5548p-SJ(config-pmap-nq-c)# pause no-drop
Cisco5548p-SJ(config-pmap-nq-c)# mtu 2158
Cisco5548p-SJ(config-pmap-nq-c)# exit
Cisco5548p-SJ(config-pmap-nq)# exit

Cisco5548p-SJ(config)# system qos
Cisco5548p-SJ(config-sys-qos)# service-policy type qos input roce
Cisco5548p-SJ(config-sys-qos)# service-policy type queuing input roce
Cisco5548p-SJ(config-sys-qos)# service-policy type queuing output roce
Cisco5548p-SJ(config-sys-qos)# service-policy type network-qos roce

```

Figure 4.

Step 3 – Install OFED

This step goes over where to download and how to install OFED.

Note – Refer to Software Requirements to determine the proper OFED version to download.

OFED can be downloaded at <https://www.openfabrics.org/downloads/OFED>.

- a) On **roce1**, extract the tarball using the “tar -xvzf <name>” command.

```
# tar -xvzf OFED-3.5-1.tgz
```

- b) In order to enable NFS over RDMA open the **install.pl** file in the unpacked directory.

- c) Locate in the file

```
#NFSRDMA
```

```
if ($kernel =~ m/^3\.5/ or $DISTRO =~ SLES11.2|RHEL6.[23]/) {
```

- d) Change to

```
#NFSRDMA
```

```
if ($kernel =~ m/^3\.5/ or $DISTRO =~ SLES11.2|RHEL6.[234]/) {
```

- e) Install OFED by running the install script.

```
# ./install.pl
```

Select option 2 in the main menu to get to the install menu. Then select option 3 in the install menu to install everything.

Note – Several rpm packages may need to be installed first. The installer will notify you of them. If necessary, the rpm packages can be found in the OS disc and installed by clicking on the icon and selecting **install package** when the window opens.

- f) Verify the installation by rerunning the install script then selecting option 3 in the main menu. It should show all the installed packages. Then reboot the system.

- g) Repeat steps **a** through **f** on **roce2**.

Step 4 – Install RoCE driver

This step goes over where to download and how to install the RoCE driver.

Note – OFED must be installed before proceeding. The driver can be downloaded from the Emulex website.

- a) On **roce1**, extract the tarball using the “tar -xvzf elx-ocrdma-dd-<release>-<version>.tar.gz” command.

```
# tar -xvzf elx-ocrdma-dd-rhel6-10.2.351.0-1.tar.gz
```

- b) Install the driver by running the install script.

```
# ./elx_roce_install.sh
```

- c) Repeat steps **a** and **b** on **roce2**.

Step 5 – Setup VLAN

This step goes over the process of creating VLANs on the **roce1** and **roce2**.

- a) On **roce1**, load the Emulex RDMA driver library by using the “modprobe <driver name>” command.

```
# modprobe ocrdma
```

- b) List the RDMA interfaces by typing “ibv_devinfo -l”

```
# ibv_devinfo -l
```

- c) List the corresponding NIC interfaces by using the “ibdev2netdev” command.

Take note of the interface **eth<x>** that is **up**. This is where you’ll create the VLAN.

```
# ibdev2netdev
```

- d) Load the 8021q module using the “modprobe <module name>” command.

```
# modprobe 8021q
```

- e) Create a VLAN using the “vconfig add eth<x>.<VLAN ID>” command.

Note – The <x> should be the number from the up interface from part d. <VLAN ID> is a number of your choosing. For this POC the ID is 10.

```
# vconfig add eth1.10
```

- f) Configure the VLAN by typing

```
“vconfig eth<x>.<VLAN ID> x.x.x.x netmask 255.255.255.0 up”
```

Note – The ‘x.x.x.x’ is the IP address for the interface. For this POC it’s 10.8.1.16 for roce1 and 10.8.1.26 for roce2.

```
# vconfig eth1.10 10.8.1.26 netmask 255.255.255.0 up
```

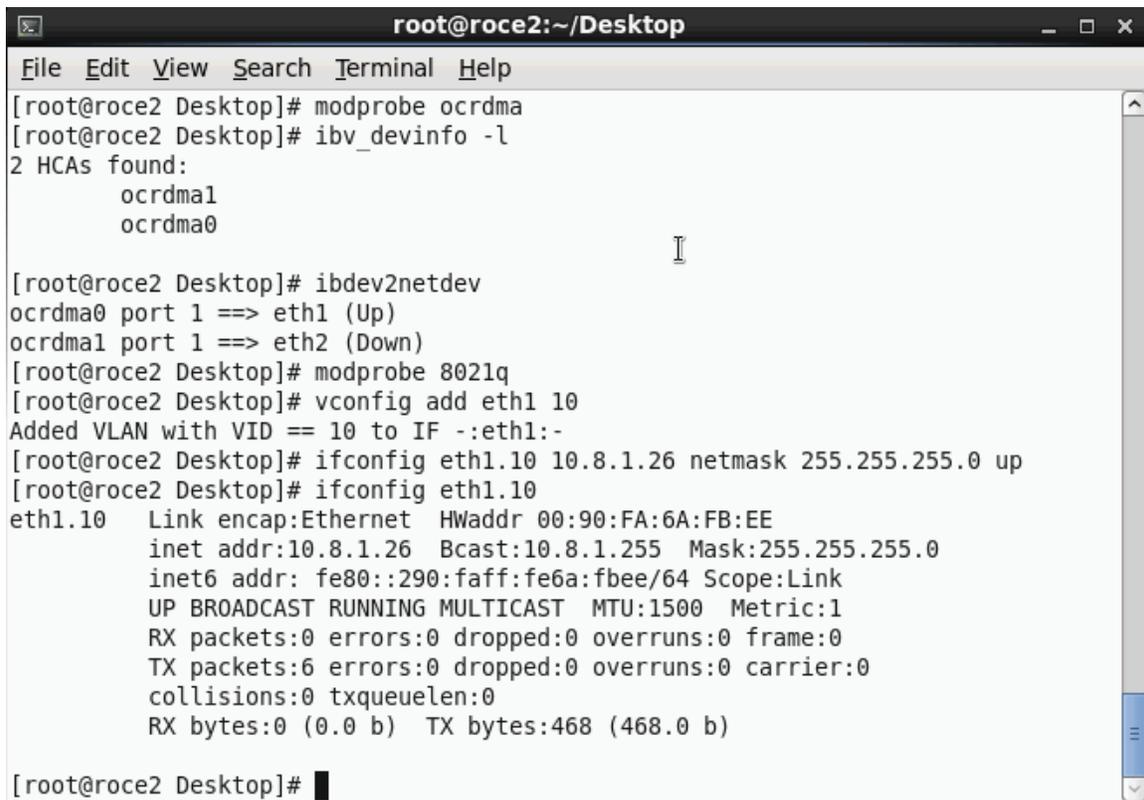
- g) Verify the configuration using the “ifconfig eth<x>.<VLAN ID>” command. This will display the interface parameters.

```
# ifconfig eth1.10
```

- h) Repeat steps **a** through **g** on **roce2**.

Note – Step 3 must be repeated after EVERY REBOOT. It is also highly recommended to “ping” one of the hosts from the other, to ensure that they are able to “talk” to one another before proceeding.

Figure 5 shows the VLAN configuration process in Terminal.



```

root@roce2:~/Desktop
File Edit View Search Terminal Help
[root@roce2 Desktop]# modprobe ocrdma
[root@roce2 Desktop]# ibv_devinfo -l
2 HCAs found:
    ocrdma1
    ocrdma0

[root@roce2 Desktop]# ibdev2netdev
ocrdma0 port 1 ==> eth1 (Up)
ocrdma1 port 1 ==> eth2 (Down)
[root@roce2 Desktop]# modprobe 8021q
[root@roce2 Desktop]# vconfig add eth1 10
Added VLAN with VID == 10 to IF -:eth1:-
[root@roce2 Desktop]# ifconfig eth1.10 10.8.1.26 netmask 255.255.255.0 up
[root@roce2 Desktop]# ifconfig eth1.10
eth1.10  Link encap:Ethernet  HWaddr 00:90:FA:6A:FB:EE
          inet addr:10.8.1.26  Bcast:10.8.1.255  Mask:255.255.255.0
          inet6 addr: fe80::290:faff:fe6a:fbee/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:6 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:0 (0.0 b)  TX bytes:468 (468.0 b)

[root@roce2 Desktop]#

```

Figure 5.

Step 6 – Setup NFS Server over RDMA

This step goes over how to enable NFS application and export a directory share on the server host.

Note – This step should only be performed on the NFS server.

- a) On roce1, load the RDMA transport module using the “modprobe <module name>” command.

```
# modprobe svcrdma
```

- b) Start NFS by running the main script. A start up checklist will be displayed.

```
# /etc/init.d/nfs start
```

- c) Set the NFS server listen port to allow RDMA using “echo “rdma <port>”>/proc/fs/nfsd/portlist” to edit the portlist directly from the command line.

Note: <port> can be any port which is available.

```
# echo “rdma 20050”>/proc/fs/nfsd/portlist
```

- d) Edit the `/etc/exports` file. Add an entry directly to it from the command line using “echo “
`<export> <client>(rw,fsid=0,insecure,no_subtree_check,async,no_root_squash)”>>/etc/exports”`

Note – In this POC the `<export>` is `/root/Desktop/nfsserver` and the `<client>` is `*` to allow universal access.

```
# echo "/root/Desktop/nfsserver *(rw,fsid=0,insecure,no_subtree_check,async,no_root_squash)">>/etc/exports
```

- e) Export the share to the client using the “exportfs” command. Add the flag “-a” to export.

```
# exportfs -a
```

- f) Verify the export using the “exportfs” command. It should show the shared directory and the recipients addressess or `<world>` if the “*” is used.

```
# exportfs
```

Note – All of Step 6 except step d must be completed after EVERY REBOOT.

Figure 6 shows the exporting process in Terminal.

```
root@roce1:~/Desktop
File Edit View Search Terminal Help
[root@roce1 Desktop]# modprobe svcrdma
[root@roce1 Desktop]# /etc/init.d/nfs start
Starting NFS services: [ OK ]
Starting NFS quotas: [ OK ]
Starting NFS mountd: [ OK ]
Stopping RPC idmapd: [ OK ]
Starting RPC idmapd: [ OK ]
Starting NFS daemon: [ OK ]
[root@roce1 Desktop]# echo "rdma 20050">/proc/fs/nfsd/portlist
[root@roce1 Desktop]# echo "/root/Desktop/nfsserver *(rw,fsid=11,insecure,no_subtree_check,async,no_root_squash)">>/etc/exports
[root@roce1 Desktop]# exportfs -a
[root@roce1 Desktop]# exportfs
/root/Desktop/nfsserver
<world>
[root@roce1 Desktop]#
```

Figure 6.

Step 7 – Mount the share from the client

This step explains how to mount an export share on a RoCE enabled client.

Note – This step should only be performed on the client host.

- a) On **roce2**, load the RDMA client module using the “`modprobe <module name>`” command.

```
# modprobe xprtrdma
```

- b) Check the status of the NFS server using the “`showmount -e <server IP>`” command. This shows the list of exports and destinations for the specified IP address.

Note – For this POC the servers IP is **10.8.1.16**.

```
# showmount -e 10.8.1.16
```

- c) Create a mount point for the share.

Note – You can use any existing directory on the client machine for the mount point, however it is recommended to create designated directories for this purpose.

For this POC the mount point will be **/root/Desktop/demo**.

- d) Mount the exported directory using the command “`mount -t nfs4 <server IP>:<file path> -o rdma,port=20050 <mount point path>`”.

The mount point will be “filled” with the contents of the exported directory.

```
# mount -t nfs4 10.8.1.16:/root/Desktop/nfsserver -o rdma,port=20050 /root/Desktop/demo
```

- e) Verify that the mount is in fact using RDMA using the “`cat /proc/mounts | grep <mount point path>`” command. Verify that the protocol field is set to rdma “**proto=rdma**”.

```
# cat /proc/mounts | grep /root/Desktop/demo
```

Note – Step 7 must be repeated after EVERY REBOOT.

Figure 7 shows the mounting process in Terminal.



```

root@roce2:~/Desktop
File Edit View Search Terminal Help
[root@roce2 Desktop]# modprobe xprtrdma
[root@roce2 Desktop]# showmount -e 10.8.1.16
Export list for 10.8.1.16:
/root/Desktop/nfsserver *
[root@roce2 Desktop]# mount -t nfs4 10.8.1.16:/root/Desktop/nfsserver -o rdma,port=20050 /root/Desktop/demo
[root@roce2 Desktop]# cat /proc/mounts | grep /root/Desktop/demo
10.8.1.16:/root/Desktop/nfsserver/ /root/Desktop/demo nfs4 rw,relatime,vers=4,rsz=262144,wsz=262144,namlen=255,hard,proto=rdma,port=0,timeo=600,retrans=2,sec=sys,clientaddr=10.8.1.26,minorversion=0,local_lock=none,addr=10.8.1.16 0 0
[root@roce2 Desktop]#

```

Figure 6.

Configuration 2

Two hosts connected back to back

The steps for this phase are exactly the same as **Configuration 1**, with the exception of the switch configuration Steps 1 and 2. If a switch is not available, this phase can be implemented instead of **Configuration 1**.

Conclusion

In summarizing the tech note, an overview of the hardware and software components needed to successfully deploy a RoCE configuration, in a switch or back to back configuration, was explained. In addition, pre and post installation steps were outlined to ensure the systems are properly updated, do not suffer from logistical issues that could compromise their functionality, and that the adapter profile is specified. Afterwards, the separate network topologies' implementations are explained and consist of installing RoCE specific software drivers, configuring the switch for Configuration 1 only, and setting up an NFS share and mount. By following the guidelines provided, configuring and enabling RoCE on Linux systems with the Emulex OCE14000 Network Adapter is fairly straightforward.

More information

OFED download

<https://www.openfabrics.org/downloads/OFED/>

Emulex downloads and resources

www.emulex.com/downloads/emulex

Cisco switch resources

www.cisco.com/c/en/us/products/switches/nexus-5548p-switch/index.html

More information on OFED

<https://www.openfabrics.org/index.php/resources/ofed-for-linux-ofed-for-windows/ofed-overview.html>

More information on RoCE

http://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

To help us improve our documents, please provide feedback at implementerslab@emulex.com.

Some of these products may not be available in the U.S. Please contact your supplier for more information.



World Headquarters 3333 Susan Street, Costa Mesa, CA 92626 +1 714 662 5600
Bangalore, India +91 80 40156789 | Beijing, China +86 10 84400221
Dublin, Ireland +35 3 (0) 1 652 1700 | Munich, Germany +49 (0) 89 97007 177
Paris, France +33 (0) 158 580 022 | Tokyo, Japan +81 3 5325 3261 | Singapore +65 6866 3768
Wokingham, United Kingdom +44 (0) 118 977 2929 | Brazil +55 11 3443 7735

www.emulex.com