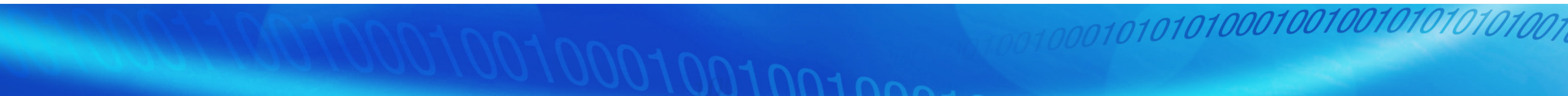
User’s Guide

Converged Network Adapters

41*xxx* Series

Third party information brought to you courtesy of Dell EMC.

AH0054602-00 H



Document Revision History	
Revision A, April 28, 2017	
Revision B, August 24, 2017	
Revision C, October 1, 2017	
Revision D, January 24, 2018	
Revision E, March 15, 2018	
Revision F, April 19, 2018	
Revision G, May 22, 2018	
Revision H, August 23, 2018	
Changes	**Sections Affected**
Updated some of the chapter content descriptions.	“What Is in This Guide” on page xix
In Table 2-2, added support for RHEL 6.10.	“System Requirements” on page 4
For Linux driver installation, added a new procedure for importing and enrolling a QLogic public key for Secure Boot.	“Importing a Public Key for Secure Boot” on page 16
Converted the following one-row table into text: **QLogic 41*xxx* Series Adapter VMware FCoE Driver**	“FCoE Support” on page 33
Converted the following one-row table into text: **QLogic 41*xxx* Series Adapter iSCSI Driver**	“iSCSI Support” on page 33
For VMware driver installation: ■ Updated the steps in the first procedure, **To install the VMware driver**. ■ Added a new procedure, **To upgrade the existing driver bundle**. ■ Changed the last procedure title from **To upgrade an existing driver** to **To upgrade an individual driver**.	“Installing VMware Drivers” on page 29
Added “NIC” to the section title and Table 3-6 (was “VMware Driver Optional Parameters”).	“VMware NIC Driver Optional Parameters” on page 30
Added a note regarding the user space library libqedr and RDMA-Core on SLES 12 SP3 and RHEL 7.4.	“Planning for RoCE” on page 64
In Step 2, added: “These packages are required for earlier versions, including RHEL 7.3 and 7.2.”	“RoCE Configuration for RHEL” on page 77

In Step 2, added: “These packages are required for earlier versions, including SLES 12 SP1/SP2.”	“RoCE Configuration for SLES” on page 77
Added a new section for configuring DCQCN for RoCE.	“Configuring DCQCN” on page 93
In the command input/output following Step 3, made all of the **`esxcli`** command input bold and changed this command: **`esxcli iscsi networkportal add -A vmhba67 -n vmk1`** (was `esxcli iscsi networkportal add -n vmk1 -A vmhba65`).	“Configuring iSER for ESXi 6.7” on page 119
Removed the note: “For more information on how to install FastLinQ drivers, see Chapter 3 Driver Installation.”	“Configuring qedi.ko” on page 154
Changed the section title to Configuring iSCSI Boot from SAN (previously titled “Open-iSCSI and Boot from SAN Considerations”).	“Configuring iSCSI Boot from SAN” on page 156
■ Moved the section to later in the chapter. ■ In Step 4, changed the command to `inst.dd modprobe.blacklist=qed,qede,qedr,qedi,qedf` (was `linux dd modprobe.blacklist=qed modprobe.blacklist=qede modprobe.blacklist=qedr modprobe.blacklist=qedi modprobe.blacklist=qedf`). ■ Updated Step 17 and Step 19. ■ Following Step 19, added a note providing a workaround for an iSCSI target on a different subnet than the initiator.	“Configuring iSCSI Boot from SAN for RHEL 7.4” on page 156
In Step 4, changed the command to `inst.dd modprobe.blacklist=qed,qede,qedr,qedi,qedf` (was `linux dd modprobe.blacklist=qed modprobe.blacklist=qede modprobe.blacklist=qedr modprobe.blacklist=qedi modprobe.blacklist=qedf`) and updated Step 13.	“Configuring iSCSI Boot from SAN for RHEL 7.5” on page 160
Added a new section for iSCSI BFS for SLES 12.	“Configuring iSCSI Boot from SAN for SLES 12 SP3 and Later” on page 161
Added the RHEL 6.10 distribution to the first sentence.	“Configuring iSCSI Boot from SAN for Other Linux Distributions” on page 162

Changed the procedure names in this section and updated the procedure steps. Following the first step of each procedure, added a note: "To retain the original contents of the inbox RPM during installation, you must use the `-ivh` option (instead of the `-Uvh` option), followed by the `--force` option."	"Migrating from Software iSCSI Installation to Offload iSCSI" on page 163
Added support for RHEL 6.10 to the RHEL 6.9 procedure. Updated and removed some of the procedure steps.	"Migrating to Offload iSCSI for RHEL 6.9/6.10" on page 163
Updated and removed some of the procedure steps.	"Migrating to Offload iSCSI for RHEL 7.2/7.3" on page 165
Updated and removed some of the procedure steps.	"Migrating to Offload iSCSI for SLES 11 SP4" on page 166
Updated and removed some of the procedure steps.	"Migrating to Offload iSCSI for SLES 12 SP1/SP2" on page 168
Added support for RHEL 6.10 to the procedure.	"Migrating and Configuring MPIO to Offloaded Interface for RHEL 6.9/6.10" on page 170
Changed the following major sections into subsections under Configuring Linux FCoE Offload: ■ "Differences Between qedf and bnx2fc" on page 186 ■ "Configuring qedf.ko" on page 186 ■ "Verifying FCoE Devices in Linux" on page 187 ■ "Prerequisites for FCoE Boot from SAN" on page 188 ■ "Configuring FCoE Boot from SAN" on page 189	"Configuring Linux FCoE Offload" on page 185
In Step 3, added a second command, `lsscsi`. Following Step 3, changed "In the preceding example, `X` could 4 or 5" to "In the preceding example, `X` is 4."	"Verifying FCoE Devices in Linux" on page 187
Added a new section for RHEL 6.10 BFS.	"Configuring FCoE Boot from SAN for RHEL 6.10" on page 190
Added a new section for FCoE kdump target.	"Using an FCoE Boot Device as a kdump Target" on page 194
Added a new section for VMware FCoE BFS.	"VMware FCoE Boot from SAN" on page 194

Added a new section with details for enabling IOMMU for Linux SR-IOV configuration.	“Enabling IOMMU for SR-IOV in UEFI-based Linux OS Installations” on page 210
In the **To configure the network or memory settings** procedure, deleted the `# echo noop > /sys/block/nvme0n1/queue/scheduler` command.	“CPU Frequency (cpufreq.sh)” on page 226
Replaced Figure 13-9 and Figure 13-10, and updated Step 7 through Step 11 to change the priority from 4 to 5.	“Configuring QoS by Disabling DCBX on the Adapter” on page 235
Added a note following Step 2. Updated Step 7 to change the priority from 4 to 5.	“Configuring QoS by Enabling DCBX on the Adapter” on page 239
Removed the obsolete section, “Setting the VMMQ Max QPs Default and Non-Default VPort”.	“Configuring VMMQ” on page 243

Table of Contents

Preface

1 Product Overview

2 Hardware Installation

3 Driver Installation

4 Upgrading the Firmware

5 Adapter Preboot Configuration

6 RoCE Configuration

11 SR-IOV Configuration

12 NVMe-oF Configuration with RDMA

13 Windows Server 2016

14 Troubleshooting

List of Figures

List of Tables

Preface

This preface lists the supported products, specifies the intended audience, explains the typographic conventions used in this guide, and describes legal notices.

Supported Products

This user's guide describes the following Cavium™ products:

- QL41112HFCU-DE 10Gb Converged Network Adapter, full-height bracket
- QL41112HLCU-DE 10Gb Converged Network Adapter, low-profile bracket
- QL41162HFRJ-DE 10Gb Converged Network Adapter, full-height bracket
- QL41162HLRJ-DE 10Gb Converged Network Adapter, low-profile bracket
- QL41162HMRJ-DE 10Gb Converged Network Adapter
- QL41164HMCU-DE 10Gb Converged Network Adapter
- QL41164HMRJ-DE 10Gb Converged Network Adapter
- QL41262HFCU-DE 10/25Gb Converged Network Adapter, full-height bracket
- QL41262HLCU-DE 10/25Gb Converged Network Adapter, low-profile bracket
- QL41262HMCU-DE 10/25Gb Converged Network
- QL41264HMCU-DE 10/25Gb Converged Network Adapter

Intended Audience

This guide is intended for system administrators and other technical staff members responsible for configuring and managing adapters installed on Dell® PowerEdge® servers in Windows®, Linux®, or VMware® environments.

What Is in This Guide

Following this preface, the remainder of this guide is organized into the following chapters and appendices:

- Chapter 1 Product Overview provides a product functional description, a list of features, and the adapter specifications.
- Chapter 2 Hardware Installation describes how to install the adapter, including the list of system requirements and a preinstallation checklist.
- Chapter 3 Driver Installation describes the installation of the adapter drivers on Windows, Linux, and VMware.
- Chapter 4 Upgrading the Firmware describes how to use the Dell Update Package (DUP) to upgrade the adapter firmware,
- Chapter 5 Adapter Preboot Configuration describes the preboot adapter configuration tasks using the Human Infrastructure Interface (HII) application.
- Chapter 6 RoCE Configuration describes how to configure the adapter, the Ethernet switch, and the host to use RDMA over converged Ethernet (RoCE).
- Chapter 7 iWARP Configuration provides procedures for configuring Internet wide area RDMA protocol (iWARP) on Windows, Linux, and VMware ESXi 6.7 systems.
- Chapter 8 iSER Configuration describes how to configure iSCSI Extensions for RDMA (iSER) for Linux RHEL, SLES, Ubuntu, and ESXi 6.7.
- Chapter 9 iSCSI Configuration describes iSCSI boot, iSCSI crash dump, and iSCSI offload for Windows and Linux.
- Chapter 10 FCoE Configuration describes Fibre Channel over Ethernet (FCoE) boot from SAN and booting from SAN after installation on Windows, Linux, and VMware systems.
- Chapter 11 SR-IOV Configuration provides procedures for configuring single root input/output virtualization (SR-IOV) on Windows, Linux, and VMware systems.
- Chapter 12 NVMe-oF Configuration with RDMA demonstrates how to configure NVMe-oF on a simple network for 41*xxx* Series Adapters.
- Chapter 13 Windows Server 2016 describes the Windows Server 2016 features.
- Chapter 14 Troubleshooting describes a variety of troubleshooting methods and resources.
- Appendix A Adapter LEDS lists the adapter LEDs and their significance.

- Appendix B Cables and Optical Modules lists the cables, optical modules, and switches that the 41*xxx* Series Adapters support.
- Appendix C Dell Z9100 Switch Configuration describes how to configure the Dell Z9100 switch port for 25Gbps.
- Appendix D Feature Constraints provides information about feature constraints implemented in the current release.

At the end of this guide is a glossary of terms.

Documentation Conventions

This guide uses the following documentation conventions:

- **NOTE** provides additional information.
- **CAUTION** without an alert symbol indicates the presence of a hazard that could cause damage to equipment or loss of data.
- **⚠ CAUTION** with an alert symbol indicates the presence of a hazard that could cause minor or moderate injury.
- **⚠ WARNING** indicates the presence of a hazard that could cause serious injury or death.
- Text in blue font indicates a hyperlink (jump) to a figure, table, or section in this guide, and links to Web sites are shown in underlined blue. For example:
  - ❑ Table 9-2 lists problems related to the user interface and remote agent.
  - ❑ See “Installation Checklist” on page 6.
  - ❑ For more information, visit www.cavium.com.
- Text in **bold** font indicates user interface elements such as a menu items, buttons, check boxes, or column headings. For example:
  - ❑ Click the **Start** button, point to **Programs**, point to **Accessories**, and then click **Command Prompt**.
  - ❑ Under **Notification Options**, select the **Warning Alarms** check box.
- Text in `Courier` font indicates a file name, directory path, or command line text. For example:
  - ❑ To return to the root directory from anywhere in the file structure: Type **`cd/ root`** and press ENTER.
  - ❑ Issue the following command: **`sh ./install.bin`**.

- Key names and key strokes are indicated with UPPERCASE:
  - Press CTRL+P.
  - Press the UP ARROW key.
- Text in *italics* indicates terms, emphasis, variables, or document titles. For example:
  - What are *shortcut keys*?
  - To enter the date type *mm/dd/yyyy* (where *mm* is the month, *dd* is the day, and *yyyy* is the year).
- Topic titles between quotation marks identify related topics either within this manual or in the online help, which is also referred to as *the help system* throughout this document.
- Command line interface (CLI) command syntax conventions include the following:
  - Plain text indicates items that you must type as shown. For example:
    - **`qaucli -pr nic -ei`**
  - `< >` (angle brackets) indicate a variable whose value you must specify. For example:
    - `<serial_number>`

    NOTE

    For CLI commands only, variable names are always indicated using angle brackets instead of *italics*.
  - `[ ]` (square brackets) indicate an optional parameter. For example:
    - `[<file_name>]` means specify a file name, or omit it to select the default file name.
  - `|` (vertical bar) indicates mutually exclusive options; select one option only. For example:
    - `on|off`
    - `1|2|3|4`
  - `...` (ellipsis) indicates that the preceding item may be repeated. For example:
    - `x...` means *one* or more instances of `x`.
    - `[x...]` means *zero* or more instances of `x`.

- ⋮ (vertical ellipsis) within command example output indicate where portions of repetitious output data have been intentionally omitted.
- `( )` (parentheses) and `{ }` (braces) are used to avoid logical ambiguity. For example:
  - `a|b c` is ambiguous
    `{(a|b) c}` means `a` or `b`, followed by `c`
    `{a|(b c)}` means either `a`, or `b c`

Legal Notices

Legal notices covered in this section include laser safety (FDA notice), agency certification, and product safety compliance.

Laser Safety—FDA Notice

This product complies with DHHS Rules 21CFR Chapter I, Subchapter J. This product has been designed and manufactured according to IEC60825-1 on the safety label of laser product.

CLASS I LASER

Class 1 Laser Product	**Caution**—Class 1 laser radiation when open Do not view directly with optical instruments
Appareil laser de classe 1	**Attention**—Radiation laser de classe 1 Ne pas regarder directement avec des instruments optiques
Produkt der Laser Klasse 1	**Vorsicht**—Laserstrahlung der Klasse 1 bei geöffneter Abdeckung Direktes Ansehen mit optischen Instrumenten vermeiden
Luokan 1 Laserlaite	**Varoitus**—Luokan 1 lasersäteilyä, kun laite on auki Älä katso suoraan laitteeseen käyttämällä optisia instrumenttej

Agency Certification

The following sections summarize the EMC and EMI test specifications performed on the 41*xxx* Series Adapters to comply with emission, immunity, and product safety standards.

EMI and EMC Requirements

FCC Part 15 compliance: Class A

FCC compliance information statement: This device complies with Part 15 of the FCC Rules. Operation is subject to the following two conditions: (1) this device may not cause harmful interference, and (2) this device must accept any interference received, including interference that may cause undesired operation.

ICES-003 Compliance: Class A

This Class A digital apparatus complies with Canadian ICES-003. Cet appareil numériqué de la classe A est conformé à la norme NMB-003 du Canada.

CE Mark 2014/30/EU, 2014/35/EU EMC Directive Compliance:

EN55032:2012/ CISPR 32:2015 Class A

EN55024: 2010
EN61000-3-2: Harmonic Current Emission
EN61000-3-3: Voltage Fluctuation and Flicker

Immunity Standards
EN61000-4-2: ESD
EN61000-4-3: RF Electro Magnetic Field
EN61000-4-4: Fast Transient/Burst
EN61000-4-5: Fast Surge Common/ Differential
EN61000-4-6: RF Conducted Susceptibility
EN61000-4-8: Power Frequency Magnetic Field
EN61000-4-11: Voltage Dips and Interrupt

VCCI: 2015-04; Class A

AS/NZS; CISPR 32: 2015 Class A

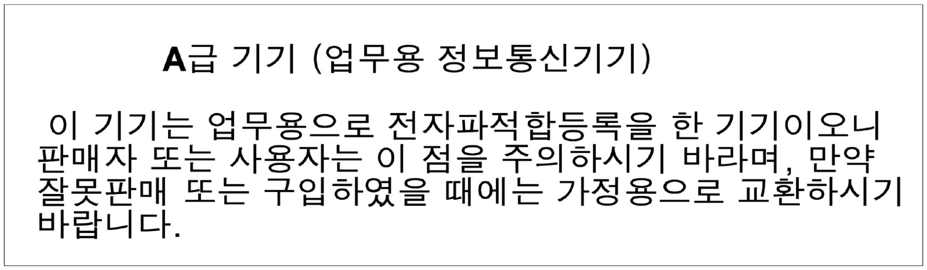
CNS 13438: 2006 Class A

KCC: Class A

Korea RRA Class A Certified

	Product Name/Model: Converged Network Adapters and Intelligent Ethernet Adapters Certification holder: QLogic Corporation Manufactured date: Refer to date code listed on product Manufacturer/Country of origin: QLogic Corporation/USA
A class equipment (Business purpose info/telecommunications equipment)	As this equipment has undergone EMC registration for business purpose, the seller and/or the buyer is asked to beware of this point and in case a wrongful sale or purchase has been made, it is asked that a change to household use be made.

Korean Language Format—Class A

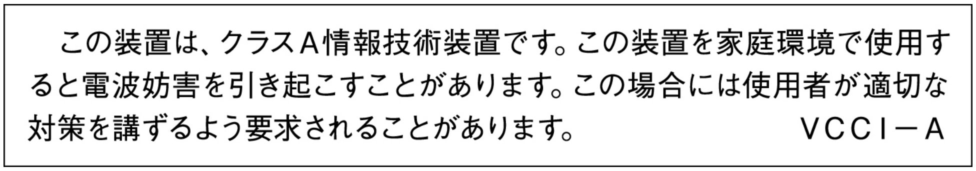


A급 기기 (업무용 정보통신기기)

이 기기는 업무용으로 전자파적합등록을 한 기기이오니 판매자 또는 사용자는 이 점을 주의하시기 바라며, 만약 잘못판매 또는 구입하였을 때에는 가정용으로 교환하시기 바랍니다.

VCCI: Class A

This is a Class A product based on the standard of the Voluntary Control Council for Interference (VCCI). If this equipment is used in a domestic environment, radio interference may occur, in which case the user may be required to take corrective actions.



この装置は、クラスA情報技術装置です。この装置を家庭環境で使用すると電波妨害を引き起こすことがあります。この場合には使用者が適切な対策を講ずるよう要求されることがあります。　VCCI－A

Product Safety Compliance

UL, cUL product safety:

UL 60950-1 (2nd Edition) A1 + A2 2014-10-14
CSA C22.2 No.60950-1-07 (2nd Edition) A1 +A2 2014-10

Use only with listed ITE or equivalent.

Complies with 21 CFR 1040.10 and 1040.11, 2014/30/EU, 2014/35/EU.

2006/95/EC low voltage directive:

TUV EN60950-1:2006+A11+A1+A12+A2 2nd Edition
TUV IEC 60950-1: 2005 2nd Edition Am1: 2009 + Am2: 2013 CB

CB Certified to IEC 60950-1 2nd Edition

1 Product Overview

This chapter provides the following information for the 41*xxx* Series Adapters:

- Functional Description
- Features
- “Adapter Specifications” on page 3

Functional Description

The Cavium FastLinQ 41000 Series Adapters include 10 and 25Gb Converged Network Adapters and Intelligent Ethernet Adapters that are designed to perform accelerated data networking for server systems. The 41000 Series Adapter includes a 10/25Gb Ethernet MAC with full-duplex capability.

Using the operating system’s teaming feature, you can split your network into virtual LANs (VLANs), as well as group multiple network adapters together into teams to provide network load balancing and fault tolerance. For more information about teaming, see your operating system documentation.

Features

The 41*xxx* Series Adapters provide the following features. Some features may not be available on all adapters:

- NIC partitioning (NPAR)
- Single-chip solution:
  - ❑ 10/25Gb MAC
  - ❑ SerDes interface for direct attach copper (DAC) transceiver connection
  - ❑ PCI Express® (PCIe®) 3.0 x8
  - ❑ Zero copy capable hardware

- Performance features:
  - TCP, IP, UDP checksum offloads
  - TCP segmentation offload (TSO)
    - Large segment offload (LSO)
    - Generic segment offload (GSO)
    - Large receive offload (LRO)
    - Receive segment coalescing (RSC)
    - Microsoft® dynamic virtual machine queue (VMQ), and Linux Multiqueue
- Adaptive interrupts:
  - Transmit/receive side scaling (TSS/RSS)
  - Stateless offloads for Network Virtualization using Generic Routing Encapsulation (NVGRE) and virtual LAN (VXLAN) L2/L3 GRE tunneled traffic[1]
- Manageability:
  - System management bus (SMB) controller
  - *Advanced Configuration and Power Interface* (ACPI) 1.1a compliant (multiple power modes)
  - Network controller-sideband interface (NC-SI) support
- Advanced network features:
  - Jumbo frames (up to 9,600 bytes). The OS and the link partner must support jumbo frames.
  - Virtual LANs (VLAN)
  - Flow control (IEEE Std 802.3x)
- Logical link control (IEEE Std 802.2)
- High-speed on-chip reduced instruction set computer (RISC) processor
- Integrated 96KB frame buffer memory (not applicable to all models)
- 1,024 classification filters (not applicable to all models)
- Support for multicast addresses through 128-bit hashing hardware function
- Serial flash NVRAM memory
- *PCI Power Management Interface* (v1.1)
- 64-bit base address register (BAR) support

[1] This feature requires OS or Hypervisor support to use the offloads.

- EM64T processor support
- iSCSI and FCoE boot support[2]

Adapter Specifications

The 41*xxx* Series Adapter specifications include the adapter's physical characteristics and standards-compliance references.

Physical Characteristics

The 41*xxx* Series Adapters are standard PCIe cards and ship with either a full-height or a low-profile bracket for use in a standard PCIe slot.

Standards Specifications

Supported standards specifications include:

- *PCI Express Base Specification*, rev. 3.1
- *PCI Express Card Electromechanical Specification*, rev. 3.0
- *PCI Bus Power Management Interface Specification*, rev. 1.2
- IEEE Specifications:
  - ❑ *802.1ad* (QinQ)
  - ❑ *802.1AX* (Link Aggregation)
  - ❑ *802.1p* (Priority Encoding)
  - ❑ *802.1q* (VLAN)
  - ❑ *802.3-2015 IEEE Standard for Ethernet* (flow control)
  - ❑ *802.3-2015 Clause 78 Energy Efficient Ethernet (EEE)*
  - ❑ *1588-2002 PTPv1* (Precision Time Protocol)
  - ❑ *1588-2008 PTPv2*
- IPv4 (RFQ 791)
- IPv6 (RFC 2460)

[2] Hardware support limit of SR-IOV VFs varies. The limit may be lower in some OS environments; refer to the appropriate section for your OS.

2 Hardware Installation

This chapter provides the following hardware installation information:

- System Requirements
- “Safety Precautions” on page 5
- “Preinstallation Checklist” on page 6
- “Installing the Adapter” on page 6

System Requirements

Before you install a Cavium 41*xxx* Series Adapter, verify that your system meets the hardware and operating system requirements shown in Table 2-1 and Table 2-2. For a complete list of supported operating systems, visit the Cavium Web site.

Table 2-1. Host Hardware Requirements

Hardware	Requirement
Architecture	IA-32 or EMT64 that meets operating system requirements
PCIe	PCIe Gen 2 x8 (2x10G NIC) PCIe Gen 3 x8 (2x25G NIC) Full dual-port 25Gb bandwidth is supported on PCIe Gen 3 x8 or faster slots.
Memory	8GB RAM (minimum)
Cables and Optical Modules	The 41*xxx* Series Adapters have been tested for interoperability with a variety of 1G, 10G, and 25G cables and optical modules. See “Tested Cables and Optical Modules” on page 274.

Table 2-2. Minimum Host Operating System Requirements

Operating System	Requirement
Windows Server	2012, 2012 R2, 2016 (including Nano)
Linux	RHEL® 6.8, 6.9, 6.10, 7.2, 7.3, 7.4 SLES® 11 SP4, SLES 12 SP2, SLES 12 SP3
VMware	ESXi 6.0 U3 and later for 25G adapters

NOTE

Table 2-2 denotes minimum host OS requirements. For a complete list of supported operating systems, visit the Cavium Web site.

Safety Precautions

⚠ WARNING

The adapter is being installed in a system that operates with voltages that can be lethal. Before you open the case of your system, observe the following precautions to protect yourself and to prevent damage to the system components.

- Remove any metallic objects or jewelry from your hands and wrists.
- Make sure to use only insulated or nonconducting tools.
- Verify that the system is powered OFF and is unplugged before you touch internal components.
- Install or remove adapters in a static-free environment. The use of a properly grounded wrist strap or other personal antistatic devices and an antistatic mat is strongly recommended.

Preinstallation Checklist

Before installing the adapter, complete the following:

1. Verify that the system meets the hardware and software requirements listed under “System Requirements” on page 4.
2. Verify that the system is using the latest BIOS.

NOTE

If you acquired the adapter software from the Cavium Web site, verify the path to the adapter driver files.

3. If the system is active, shut it down.
4. When system shutdown is complete, turn off the power and unplug the power cord.
5. Remove the adapter from its shipping package and place it on an anti-static surface.
6. Check the adapter for visible signs of damage, particularly on the edge connector. Never attempt to install a damaged adapter.

Installing the Adapter

The following instructions apply to installing the Cavium 41*xxx* Series Adapters in most systems. For details about performing these tasks, refer to the manuals that were supplied with the system.

To install the adapter:

1. Review “Safety Precautions” on page 5 and “Preinstallation Checklist” on page 6. Before you install the adapter, ensure that the system power is OFF, the power cord is unplugged from the power outlet, and that you are following proper electrical grounding procedures.
2. Open the system case, and select the slot that matches the adapter size, which can be PCIe Gen 2 x8 or PCIe Gen 3 x8. A lesser-width adapter can be seated into a greater-width slot (x8 in an x16), but a greater-width adapter cannot be seated into a lesser-width slot (x8 in an x4). If you do not know how to identify a PCIe slot, refer to your system documentation.
3. Remove the blank cover-plate from the slot that you selected.
4. Align the adapter connector edge with the PCIe connector slot in the system.

5. Applying even pressure at both corners of the card, push the adapter card into the slot until it is firmly seated. When the adapter is properly seated, the adapter port connectors are aligned with the slot opening, and the adapter faceplate is flush against the system chassis.

   CAUTION

   Do not use excessive force when seating the card, because this may damage the system or the adapter. If you have difficulty seating the adapter, remove it, realign it, and try again.

6. Secure the adapter with the adapter clip or screw.
7. Close the system case and disconnect any personal anti-static devices.

3 Driver Installation

This chapter provides the following information about driver installation:

- Installing Linux Driver Software
- "Installing Windows Driver Software" on page 17
- "Installing VMware Driver Software" on page 27

Installing Linux Driver Software

This section describes how to install Linux drivers with or without remote direct memory access (RDMA). It also describes the Linux driver optional parameters, default values, messages, statistics, and public key for Secure Boot.

- Installing the Linux Drivers Without RDMA
- Installing the Linux Drivers with RDMA
- Linux Driver Optional Parameters
- Linux Driver Operation Defaults
- Linux Driver Messages
- Statistics
- Importing a Public Key for Secure Boot

The 41*xxx* Series Adapter Linux drivers and supporting documentation are available on the Dell Support page:

dell.support.com

Table 3-1 describes the 41*xxx* Series Adapter Linux drivers.

Table 3-1. Cavium QLogic 41xxx Series Adapters Linux Drivers

Linux Driver	Description
qed	The qed core driver module directly controls the firmware, handles interrupts, and provides the low-level API for the protocol specific driver set. The qed interfaces with the qede, qedr, qedi, and qedf drivers. The Linux core module manages all PCI device resources (registers, host interface queues, and so on). The qed core module requires Linux kernel version 2.6.32 or later. Testing was concentrated on the x86_64 architecture.
qede	Linux Ethernet driver for the 41*xxx* Series Adapter. This driver directly controls the hardware and is responsible for sending and receiving Ethernet packets on behalf of the Linux host networking stack. This driver also receives and processes device interrupts on behalf of itself (for L2 networking). The qede driver requires Linux kernel version 2.6.32 or later. Testing was concentrated on the x86_64 architecture.
qedr	Linux RoCE driver that works in the OpenFabrics Enterprise Distribution (OFED™) environment in conjunction with the qed core module and the qede Ethernet driver. RDMA user space applications also require that the libqedr user library is installed on the server.
qedi	Linux iSCSI-Offload driver for the 41*xxx* Series Adapters. This driver works with the Open iSCSI library.
qedf	Linux FCoE-Offload driver for the 41*xxx* Series Adapters. This driver works with Open FCoE library.

Install the Linux drivers using either a source Red Hat® Package Manager (RPM) package or a kmod RPM package. The RHEL RPM packages are as follows:

- `qlgc-fastlinq-<version>.<OS>.src.rpm`
- `qlgc-fastlinq-kmp-default-<version>.<arch>.rpm`

The SLES source and kmp RPM packages are as follows:

- `qlgc-fastlinq-<version>.<OS>.src.rpm`
- `qlgc-fastlinq-kmp-default-<version>.<OS>.<arch>.rpm`

The following kernel module (kmod) RPM installs Linux drivers on SLES hosts running the Xen Hypervisor:

- `qlgc-fastlinq-kmp-xen-<version>.<OS>.<arch>.rpm`

The following source RPM installs the RDMA library code on RHEL and SLES hosts:

- `qlgc-libqedr-<version>.<OS>.<arch>.src.rpm`

The following source code TAR BZip2 (BZ2) compressed file installs Linux drivers on RHEL and SLES hosts:

- `fastlinq-<version>.tar.bz2`

NOTE

For network installations through NFS, FTP, or HTTP (using a network boot disk), you may require a driver disk that contains the qede driver. Compile the Linux boot drivers by modifying the makefile and the make environment.

Installing the Linux Drivers Without RDMA

To install the Linux drivers without RDMA:

1. Download the 41*xxx* Series Adapter Linux drivers from Dell:

   dell.support.com

2. Remove the existing Linux drivers, as described in "Removing the Linux Drivers" on page 10.

3. Install the new Linux drivers using one of the following methods:

   - ❑ Installing Linux Drivers Using the src RPM Package
   - ❑ Installing Linux Drivers Using the kmp/kmod RPM Package
   - ❑ Installing Linux Drivers Using the TAR File

Removing the Linux Drivers

There are two procedures for removing Linux drivers: one for a non-RDMA environment and another for an RDMA environment. Choose the procedure that matches your environment.

To remove Linux drivers in a non-RDMA environment, unload and remove the drivers:

Follow the procedure that relates to the original installation method and the OS.

- If the Linux drivers were installed using an RPM package, issue the following commands:

```
rmmod qede
rmmod qed
depmod -a
rpm -e qlgc-fastlinq-kmp-default-<version>.<arch>
```

- If the Linux drivers were installed using a TAR file, issue the following commands:

```
rmmod qede
rmmod qed
depmod -a
```

- For RHEL:

  ```
  cd /lib/modules/<version>/extra/qlgc-fastlinq
  rm -rf qed.ko qede.ko qedr.ko
  ```

- For SLES:

  ```
  cd /lib/modules/<version>/updates/qlgc-fastlinq
  rm -rf qed.ko qede.ko qedr.ko
  ```

To remove Linux drivers in a non-RDMA environment:

1. To get the path to the currently installed drivers, issue the following command:

   ```
   modinfo <driver name>
   ```

2. Unload and remove the Linux drivers.

   - If the Linux drivers were installed using an RPM package, issue the following commands:

     ```
     modprobe -r qede
     depmod -a
     rpm -e qlgc-fastlinq-kmp-default-<version>.<arch>
     ```

   - If the Linux drivers were installed using a TAR file, issue the following commands:

     ```
     modprobe -r qede
     depmod -a
     ```

     NOTE

     If the qedr is present, issue the `modprobe -r qedr` command instead.

3. Delete the `qed.ko`, `qede.ko`, and `qedr.ko` files from the directory in which they reside. For example, in SLES, issue the following commands:

   ```
   cd /lib/modules/<version>/updates/qlgc-fastlinq
   rm -rf qed.ko
   rm -rf qede.ko
   rm -rf qedr.ko
   depmod -a
   ```

To remove Linux drivers in an RDMA environment:

1. To get the path to the installed drivers, issue the following command:

   ```
   modinfo <driver name>
   ```

2. Unload and remove the Linux drivers.

   ```
   modprobe -r qedr
   modprobe -r qede
   modprobe -r qed
   depmod -a
   ```

3. Remove the driver module files:

   ❑ If the drivers were installed using an RPM package, issue the following command:

   ```
   rpm -e qlgc-fastlinq-kmp-default-<version>.<arch>
   ```

   ❑ If the drivers were installed using a TAR file, issue the following commands for your operating system:

   For RHEL:

   ```
   cd /lib/modules/<version>/extra/qlgc-fastlinq
   rm -rf qed.ko qede.ko qedr.ko
   ```

   For SLES:

   ```
   cd /lib/modules/<version>/updates/qlgc-fastlinq
   rm -rf qed.ko qede.ko qedr.ko
   ```

Installing Linux Drivers Using the src RPM Package

To install Linux drivers using the src RPM package:

1. Issue the following at a command prompt:

   ```
   rpm -ivh RPMS/<arch>/qlgc-fastlinq-<version>.src.rpm
   ```

2. Change the directory to the RPM path and build the binary RPM for the kernel.

   For RHEL:

   ```
   cd /root/rpmbuild
   rpmbuild -bb SPECS/fastlinq-<version>.spec
   ```

   For SLES:

   ```
   cd /usr/src/packages
   ```

```
rpmbuild -bb SPECS/fastlinq-<version>.spec
```

3. Install the newly compiled RPM:

```
rpm -ivh RPMS/<arch>/qlgc-fastlinq-<version>.<arch>.rpm
```

NOTE

The `--force` option may be needed on some Linux distributions if conflicts are reported.

The drivers will be installed in the following paths.

For SLES:

```
/lib/modules/<version>/updates/qlgc-fastlinq
```

For RHEL:

```
/lib/modules/<version>/extra/qlgc-fastlinq
```

4. Turn on all ethX interfaces as follows:

```
ifconfig <ethX> up
```

5. For SLES, use YaST to configure the Ethernet interfaces to automatically start at boot by setting a static IP address or enabling DHCP on the interface.

Installing Linux Drivers Using the kmp/kmod RPM Package

To install kmod RPM package:

1. Issue the following command at a command prompt:

```
rpm -ivh qlgc-fastlinq-<version>.<arch>.rpm
```

2. Reload the driver:

```
modprobe -r qede
modprobe qede
```

Installing Linux Drivers Using the TAR File

To install Linux drivers using the TAR file:

1. Create a directory and extract the TAR files to the directory:

```
tar xjvf fastlinq-<version>.tar.bz2
```

2. Change to the recently created directory, and then install the drivers:

```
cd fastlinq-<version>
make clean; make install
```

The qed and qede drivers will be installed in the following paths.

For SLES:

```
/lib/modules/<version>/updates/qlgc-fastlinq
```

For RHEL:

```
/lib/modules/<version>/extra/qlgc-fastlinq
```

3. Test the drivers by loading them (unload the existing drivers first, if necessary):

```
rmmod qede
rmmod qed
modprobe qed
modprobe qede
```

Installing the Linux Drivers with RDMA

For information on iWARP, see Chapter 7 iWARP Configuration.

To install Linux drivers in an inbox OFED environment:

1. Download the 41*xxx* Series Adapter Linux drivers from the Dell:

   dell.support.com

2. Configure RoCE on the adapter, as described in "Configuring RoCE on the Adapter for Linux" on page 76.

3. Remove existing Linux drivers, as described in "Removing the Linux Drivers" on page 10.

4. Install the new Linux drivers using one of the following methods:

   ❑ Installing Linux Drivers Using the kmp/kmod RPM Package

   ❑ Installing Linux Drivers Using the TAR File

5. Install libqedr libraries to work with RDMA user space applications. The libqedr RPM is available only for inbox OFED. You must select which RDMA (RoCE, RoCEv2, or iWARP) is used in UEFI until concurrent RoCE+iWARP capability is supported in the firmware). None is enabled by default. Issue the following command:

```
rpm -ivh qlgc-libqedr-<version>.<arch>.rpm
```

6. To build and install the libqedr user space library, issue the following command:

```
'make libqedr_install'
```

7. Test the drivers by loading them as follows:

```
modprobe qedr
make install_libeqdr
```

Linux Driver Optional Parameters

Table 3-2 describes the optional parameters for the qede driver.

Table 3-2. qede Driver Optional Parameters

Parameter	Description
`debug`	Controls driver verbosity level similar to `ethtool -s <dev> msglvl`.
`int_mode`	Controls interrupt mode other than MSI-X.
`gro_enable`	Enables or disables the hardware generic receive offload (GRO) feature. This feature is similar to the kernel's software GRO, but is only performed by the device hardware.
`err_flags_override`	A bitmap for disabling or forcing the actions taken in case of a hardware error: ■ bit #31 – An enable bit for this bitmask ■ bit #0 – Prevent hardware attentions from being reasserted ■ bit #1 – Collect debug data ■ bit #2 – Trigger a recovery process ■ bit #3 – Call WARN to get a call trace of the flow that led to the error

Linux Driver Operation Defaults

Table 3-3 lists the qed and qede Linux driver operation defaults.

Table 3-3. Linux Driver Operation Defaults

Operation	qed Driver Default	qede Driver Default
`Speed`	Auto-negotiation with speed advertised	Auto-negotiation with speed advertised
`MSI/MSI-X`	Enabled	Enabled
`Flow Control`	—	Auto-negotiation with RX and TX advertised
`MTU`	—	1500 (range is 46–9600)
`Rx Ring Size`	—	1000

Table 3-3. Linux Driver Operation Defaults (Continued)

Operation	qed Driver Default	qede Driver Default
`Tx Ring Size`	—	4078 (range is 128–8191)
`Coalesce Rx Microseconds`	—	24 (range is 0–255)
`Coalesce Tx Microseconds`	—	48
`TSO`	—	Enabled

Linux Driver Messages

To set the Linux driver message detail level, issue one of the following commands:

- **`ethtool -s <interface> msglvl <value>`**
- **`modprobe qede debug=<value>`**

  Where `<value>` represents bits 0–15, which are standard Linux networking values, and bits 16 and greater are driver-specific.

Statistics

To view detailed statistics and configuration information, use the ethtool utility. See the ethtool man page for more information.

Importing a Public Key for Secure Boot

Linux drivers require that you import and enroll the QLogic public key to load the drivers in a Secure Boot environment. Before you begin, ensure that your server supports Secure Boot. This section provides two methods for importing and enrolling the public key.

To import and enroll the QLogic public key:

1. Download the public key from the following Web page:

   http://ldriver.qlogic.com/Module-public-key/

2. To install the public key, issue the following command:

   ```
   # mokutil --root-pw --import cert.der
   ```

   Where the `--root-pw` option enables direct use of the root user.

3. Reboot the system.
4. Review the list of certificates that are prepared to be enrolled:

   ```
   # mokutil --list-new
   ```

5. Reboot the system again.
6. When the shim launches MokManager, enter the root password to confirm the certificate importation to the Machine Owner Key (MOK) list.
7. To determine if the newly imported key was enrolled:

```
# mokutil --list-enrolled
```

To launch MOK manually and enroll the QLogic public key:

1. Issue the following command:

```
# reboot
```

2. In the **GRUB 2** menu, press the C key.
3. Issue the following commands:

```
chainloader $efibootdir/MokManager.efi
- boot
```

4. Select **Enroll key from disk**.
5. Navigate to the `cert.der` file and then press ENTER.
6. Follow the instructions to enroll the key. Generally this includes pressing the 0 (zero) key and then pressing the Y key to confirm.

NOTE

The firmware menu may provide more methods to add a new key to the Signature Database.

For additional information about Secure Boot, refer to the following Web page:

https://www.suse.com/documentation/sled-12/book_sle_admin/data/sec_uefi_secboot.html

Installing Windows Driver Software

For information on iWARP, see Chapter 7 iWARP Configuration.

- Installing the Windows Drivers
- Removing the Windows Drivers
- Managing Adapter Properties
- Setting Power Management Options

Installing the Windows Drivers

Install Windows driver software using the Dell Update Package (DUP):

- Running the DUP in the GUI
- DUP Installation Options
- DUP Installation Examples

Running the DUP in the GUI

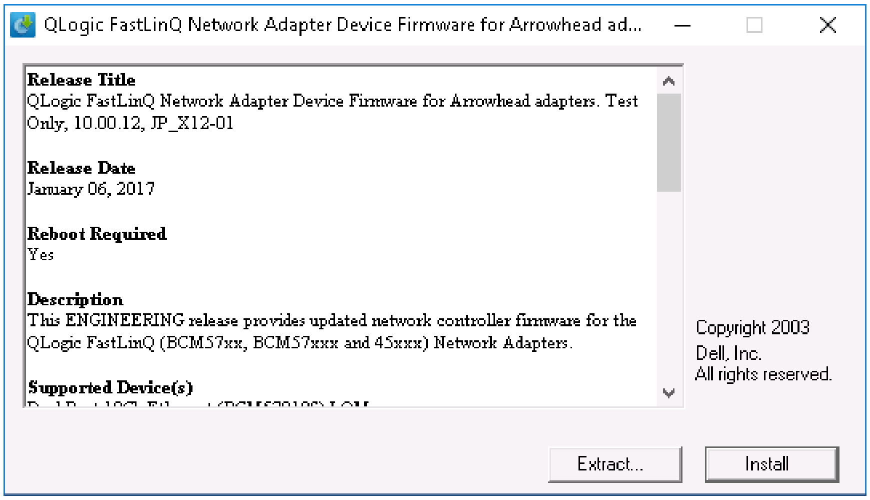
To run the DUP in the GUI:

1. Double-click the icon representing the Dell Update Package file.

   NOTE

   The actual file name of the Dell Update Package varies.

2. In the Dell Update Package window (Figure 3-1), click **Install**.



QLogic FastLinQ Network Adapter Device Firmware for Arrowhead ad...
Extract...
Install

Release Title
QLogic FastLinQ Network Adapter Device Firmware for Arrowhead adapters. Test Only, 10.00.12, JP_X12-01

Release Date
January 06, 2017

Reboot Required
Yes

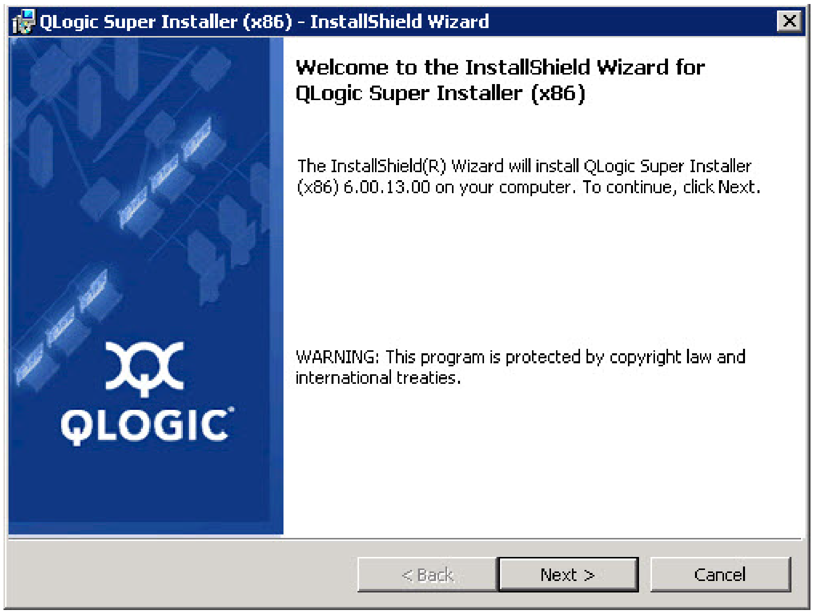
Description
This ENGINEERING release provides updated network controller firmware for the QLogic FastLinQ (BCM57xx, BCM57xxx and 45xxx) Network Adapters.

Supported Device(s)

Copyright 2003
Dell, Inc.
All rights reserved.

Figure 3-1. Dell Update Package Window

3. In the QLogic Super Installer—InstallShield® Wizard’s Welcome window (Figure 3-2), click **Next**.



Welcome to the InstallShield Wizard for QLogic Super Installer (x86)

The InstallShield(R) Wizard will install QLogic Super Installer (x86) 6.00.13.00 on your computer. To continue, click Next.

WARNING: This program is protected by copyright law and international treaties.

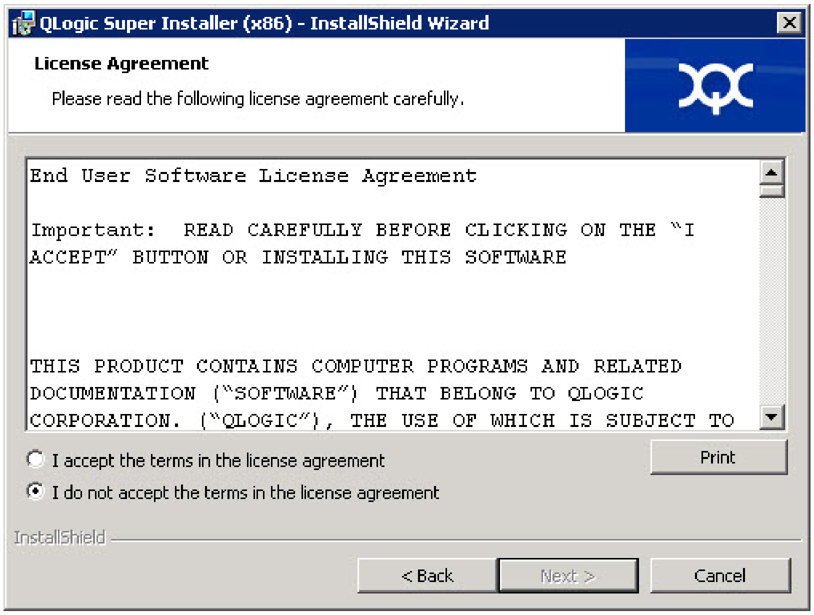
Figure 3-2. QLogic InstallShield Wizard: Welcome Window

4. Complete the following in the wizard's License Agreement window (Figure 3-3):

   a. Read the QLogic End User Software License Agreement.

   b. To continue, select **I accept the terms in the license agreement**.

   c. Click **Next**.



QLogic Super Installer (x86) - InstallShield Wizard

License Agreement

Please read the following license agreement carefully.

```
End User Software License Agreement

Important:  READ CAREFULLY BEFORE CLICKING ON THE "I
ACCEPT" BUTTON OR INSTALLING THIS SOFTWARE


THIS PRODUCT CONTAINS COMPUTER PROGRAMS AND RELATED
DOCUMENTATION ("SOFTWARE") THAT BELONG TO QLOGIC
CORPORATION. ("QLOGIC"), THE USE OF WHICH IS SUBJECT TO
```

○ I accept the terms in the license agreement

● I do not accept the terms in the license agreement

Figure 3-3. QLogic InstallShield Wizard: License Agreement Window

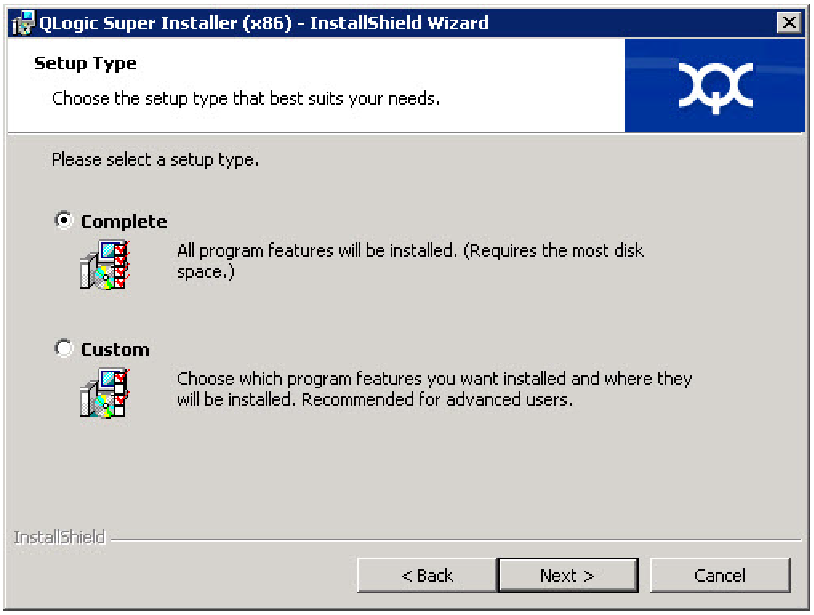
5. Complete the wizard's Setup Type window (Figure 3-4) as follows:

   a. Select one of the following setup types:

      - Click **Complete** to install all program features.
      - Click **Custom** to manually select the features to be installed.

   b. To continue, click **Next**.

   If you clicked **Complete**, proceed directly to Step 6b.



QLogic Super Installer (x86) - InstallShield Wizard

Setup Type

Choose the setup type that best suits your needs.

Please select a setup type.

● **Complete**

All program features will be installed. (Requires the most disk space.)

○ **Custom**

Choose which program features you want installed and where they will be installed. Recommended for advanced users.

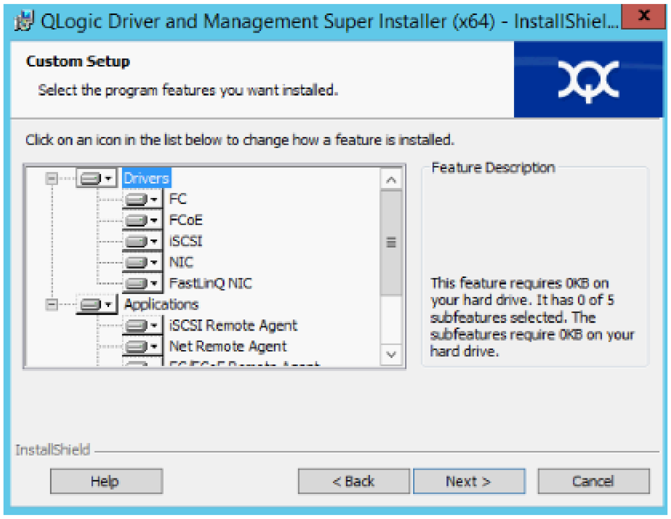
Figure 3-4. InstallShield Wizard: Setup Type Window

6. If you selected **Custom** in Step 5, complete the Custom Setup window (Figure 3-5) as follows:

   a. Select the features to install. By default, all features are selected. To change a feature's install setting, click the icon next to it, and then select one of the following options:

      - **This feature will be installed on the local hard drive**—Marks the feature for installation without affecting any of its subfeatures.
      - **This feature, and all subfeatures, will be installed on the local hard drive**—Marks the feature and all of its subfeatures for installation.
      - **This feature will not be available**—Prevents the feature from being installed.

   b. Click **Next** to continue.



QLogic Driver and Management Super Installer (x64) - InstallShiel...

Custom Setup

Select the program features you want installed.

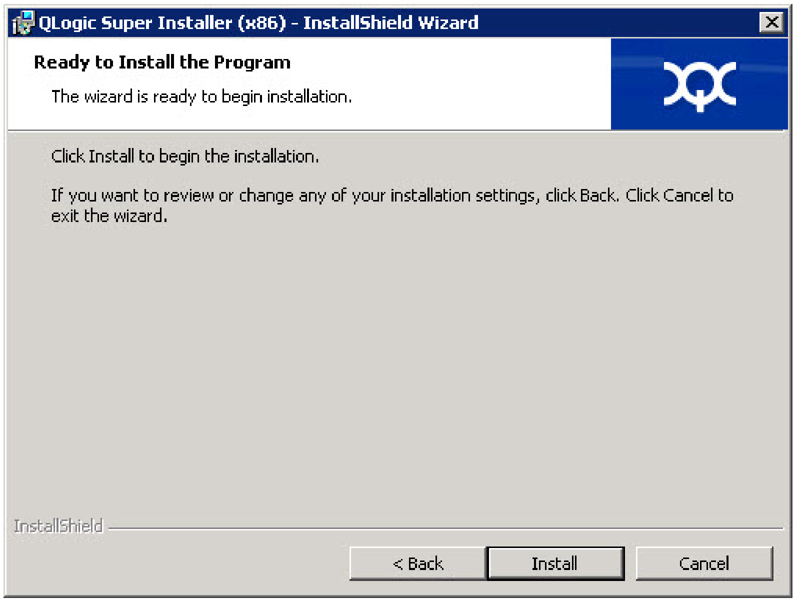
Click on an icon in the list below to change how a feature is installed.

Feature Description

This feature requires 0KB on your hard drive. It has 0 of 5 subfeatures selected. The subfeatures require 0KB on your hard drive.

Figure 3-5. InstallShield Wizard: Custom Setup Window

7. In the InstallShield Wizard's Ready To Install window (Figure 3-6), click **Install**. The InstallShield Wizard installs the QLogic Adapter drivers and Management Software Installer.



QLogic Super Installer (x86) - InstallShield Wizard

Ready to Install the Program

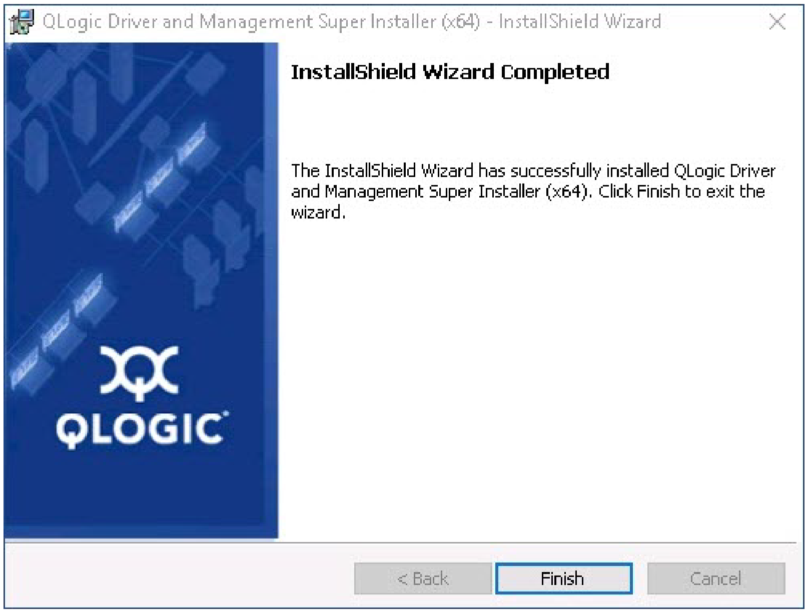
The wizard is ready to begin installation.

Click Install to begin the installation.

If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.

Figure 3-6. InstallShield Wizard: Ready to Install the Program Window

8. When the installation is complete, the InstallShield Wizard Completed window appears (Figure 3-7). Click **Finish** to dismiss the installer.



QLogic Driver and Management Super Installer (x64) - InstallShield Wizard

InstallShield Wizard Completed

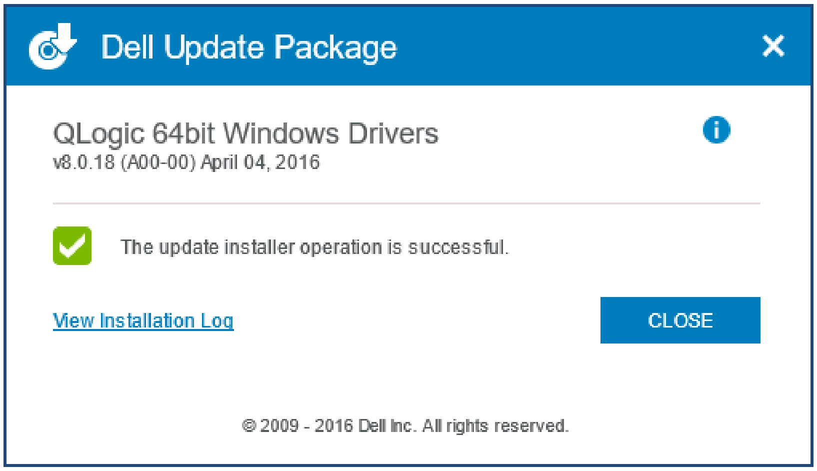
The InstallShield Wizard has successfully installed QLogic Driver and Management Super Installer (x64). Click Finish to exit the wizard.

Figure 3-7. InstallShield Wizard: Completed Window

9. In the Dell Update Package window (Figure 3-8), “Update installer operation was successful” indicates completion.

   ❑ (Optional) To open the log file, click **View Installation Log**. The log file shows the progress of the DUP installation, any previous installed versions, any error messages, and other information about the installation.

   ❑ To close the Update Package window, click **CLOSE**.



Dell Update Package

QLogic 64bit Windows Drivers
v8.0.18 (A00-00) April 04, 2016

The update installer operation is successful.

© 2009 - 2016 Dell Inc. All rights reserved.

Figure 3-8. Dell Update Package Window

DUP Installation Options

To customize the DUP installation behavior, use the following command line options.

- To extract only the driver components to a directory:

  `/drivers=<path>`

  NOTE

  This command requires the **`/s`** option.

- To install or update only the driver components:

  `/driveronly`

  NOTE

  This command requires the **`/s`** option.

- (Advanced) Use the `/passthrough` option to send all text following `/passthrough` directly to the QLogic installation software of the DUP. This mode suppresses any provided GUIs, but not necessarily those of the QLogic software.

  `/passthrough`

- (Advanced) To return a coded description of this DUP’s supported features:

  `/capabilities`

NOTE

This command requires the `/s` option.

DUP Installation Examples

The following examples show how to use the installation options.

To update the system silently:

```
<DUP_file_name>.exe /s
```

To extract the update contents to the `C:\mydir\` directory:

```
<DUP_file_name>.exe /s /e=C:\mydir
```

To extract the driver components to the `C:\mydir\` directory:

```
<DUP_file_name>.exe /s /drivers=C:\mydir
```

To install only the driver components:

```
<DUP_file_name>.exe /s /driveronly
```

To change from the default log location to `C:\my path with spaces\log.txt`:

```
<DUP_file_name>.exe /l="C:\my path with spaces\log.txt"
```

Removing the Windows Drivers

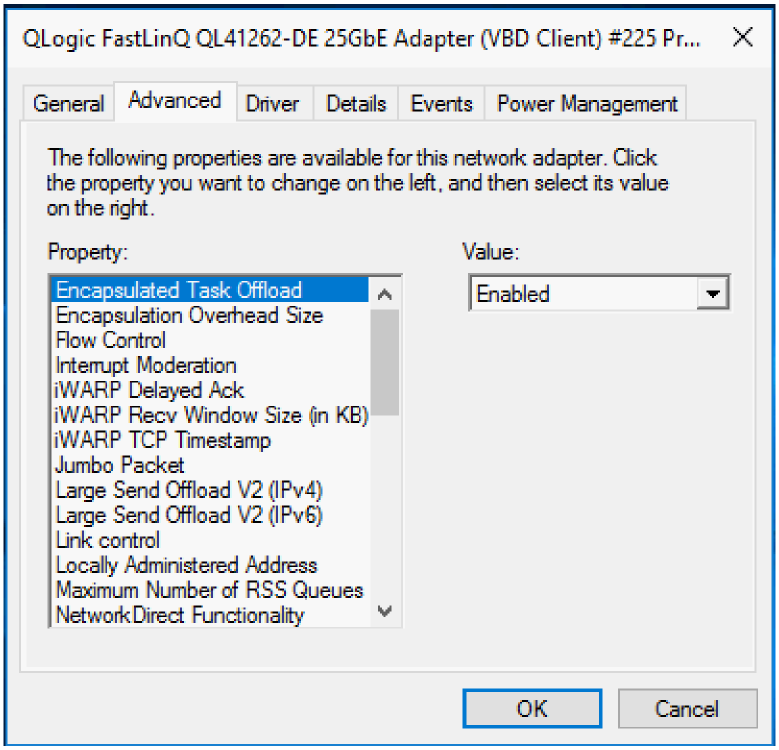
To remove the Windows drivers:

1. In the Control Panel, click **Programs**, and then click **Programs and Features**.
2. In the list of programs, select **QLogic FastLinQ Driver Installer**, and then click **Uninstall**.
3. Follow the instructions to remove the drivers.

Managing Adapter Properties

To view or change the 41*xxx* Series Adapter properties:

1. In the Control Panel, click **Device Manager**.
2. On the properties of the selected adapter, click the **Advanced** tab.
3. On the Advanced page (Figure 3-9), select an item under **Property** and then change the **Value** for that item as needed.


QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...
General
Advanced
Driver
Details
Events
Power Management
OK
Cancel

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

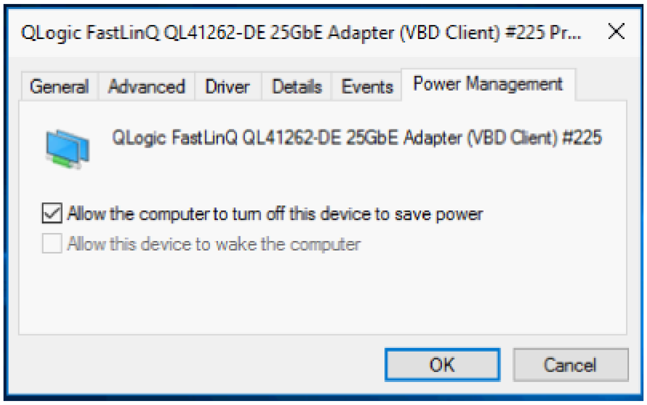
- Encapsulated Task Offload
- Encapsulation Overhead Size
- Flow Control
- Interrupt Moderation
- iWARP Delayed Ack
- iWARP Recv Window Size (in KB)
- iWARP TCP Timestamp
- Jumbo Packet
- Large Send Offload V2 (IPv4)
- Large Send Offload V2 (IPv6)
- Link control
- Locally Administered Address
- Maximum Number of RSS Queues
- NetworkDirect Functionality

Value: Enabled

Figure 3-9. Setting Advanced Adapter Properties

Setting Power Management Options

You can set power management options to allow the operating system to turn off the controller to save power or to allow the controller to wake up the computer. If the device is busy (servicing a call, for example), the operating system will not shut down the device. The operating system attempts to shut down every possible device only when the computer attempts to go into hibernation. To have the controller remain on at all times, do not select the **Allow the computer to turn off the device to save power** check box (Figure 3-10).


QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...

QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225

☑ Allow the computer to turn off this device to save power
□ Allow this device to wake the computer

Figure 3-10. Power Management Options

NOTE

- The Power Management page is available only for servers that support power management.
- Do not select the **Allow the computer to turn off the device to save power** check box for any adapter that is a member of a team.

Installing VMware Driver Software

This section describes the qedentv VMware ESXi driver for the 41*xxx* Series Adapters:

- VMware Drivers and Driver Packages
- Installing VMware Drivers
- VMware NIC Driver Optional Parameters
- VMware Driver Parameter Defaults
- Removing the VMware Driver

- FCoE Support
- iSCSI Support

VMware Drivers and Driver Packages

Table 3-4 lists the VMware ESXi drivers for the protocols.

Table 3-4. VMware Drivers

VMware Driver	Description
qedentv	Native networking driver
qedrntv	Native RDMA-Offload (RoCE and RoCEv2) driver [a]
qedf	Native FCoE-Offload driver
qedil	Legacy iSCSI-Offload driver

[a] The certified RoCE driver is not included in this release. The uncertified driver may be available as an early preview.

The ESXi drivers are included as individual driver packages and are not bundled together, except as noted. Table 3-5 lists the ESXi versions and applicable driver versions.

Table 3-5. ESXi Driver Packages by Release

ESXi Release [a]	Protocol	Driver Name	Driver Version
ESXi 6.5 [b]	NIC	qedentv	3.0.7.5
	FCoE	qedf	1.2.24.0
	iSCSI	qedil	1.0.19.0
	RoCE	qedrntv	3.0.7.5.1
ESXi 6.0u3	NIC	qedentv	2.0.7.5
	FCoE	qedf	1.2.24.0
	iSCSI	qedil	1.0.19.0

[a] Additional ESXi drivers may be available after this user’s guide has been published. For more information, refer to the release notes.

[b] For ESXi 6.5, the NIC and RoCE drivers have been packaged together and can be installed as a single offline bundle using the standard ESXi installation commands. The package name is *qedentv_3.0.7.5_qedrntv_3.0.7.5.1_signed_drivers.zip*. The recommended installation sequence is NIC and RoCE drivers, followed by FCoE and iSCSI drivers.

Install individual drivers using either:

- Standard ESXi package installation commands (see Installing VMware Drivers)
- Procedures in the individual driver Read Me files
- Procedures in the following VMware KB article:

  https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2137853

You should install the NIC driver first, followed by the storage drivers.

Installing VMware Drivers

You can use the driver ZIP file to install a new driver or update an existing driver. Be sure to install the entire driver set from the same driver ZIP file. Mixing drivers from different ZIP files will cause problems.

To install the VMware driver:

1. Download the VMware driver for the 41*xxx* Series Adapter from the VMware support page:

   www.vmware.com/support.html

2. Power up the ESX host, and then log into an account with administrator authority.

3. Use the Linux scp utility to copy the driver bundle from a local system into the `/tmp` directory on an ESX server with IP address 10.10.10.10. For example, issue the following command:

   ```
   # scp qedentv-bundle-2.0.3.zip root@10.10.10.10:/tmp
   ```

   You can place the file anywhere that is accessible to the ESX console shell.

4. Place the host in maintenance mode by issuing the following command:

   ```
   # esxcli --maintenance-mode
   ```

5. Select one of the following installation options:

   - ❑ **Option 1:** Install the driver bundle (which will install all of the driver VIBs at one time) by issuing the following command:

     ```
     # esxcli software vib install -d /tmp/qedentv-2.0.3.zip
     ```

   - ❑ **Option 2:** Install the `.vib` directly on an ESX server using either the CLI or the VMware Update Manager (VUM). To do this, unzip the driver ZIP file, and then extract the `.vib` file.

- To install the .vib file using the CLI, issue the following command. Be sure to specify the full .vib file path:

```
# esxcli software vib install -v /tmp/qedentv-1.0.3.11-1OEM.550.0.0.1331820.x86_64.vib
```

- To install the .vib file using the VUM, see the knowledge base article here:

  Updating an ESXi/ESX host using VMware vCenter Update Manager 4.x and 5.x (1019545)

To upgrade the existing driver bundle:

- Issue the following command:

  ```
  # esxcli software vib update -d /tmp/qedentv-bundle-2.0.3.zip
  ```

To upgrade an individual driver:

Follow the steps for a new installation (see **To install the VMware driver)**, except replace the command in Option 1 with the following:

```
# esxcli software vib update -v /tmp/qedentv-1.0.3.11-1OEM.550.0.0.1331820.x86_64.vib
```

VMware NIC Driver Optional Parameters

Table 3-6 describes the optional parameters that can be supplied as command line arguments to the esxcfg-module command.

Table 3-6. VMware NIC Driver Optional Parameters

Parameter	Description
hw_vlan	Globally enables (1) or disables (0) hardware VLAN insertion and removal. Disable this parameter when the upper layer needs to send or receive fully formed packets. hw_vlan=1 is the default.
num_queues	Specifies the number of TX/RX queue pairs. num_queues can be 1-11 or one of the following: ■ -1 allows the driver to determine the optimal number of queue pairs (default). ■ 0 uses the default queue. You can specify multiple values delimited by commas for multiport or multi-function configurations.
multi_rx_filters	Specifies the number of RX filters per RX queue, excluding the default queue. multi_rx_filters can be 1-4 or one of the following values: ■ -1 uses the default number of RX filters per queue. ■ 0 disables RX filters.

Table 3-6. VMware NIC Driver Optional Parameters (Continued)

Parameter	Description
`disable_tpa`	Enables (`0`) or disables (`1`) the TPA (LRO) feature. `disable_tpa=0` is the default.
`max_vfs`	Specifies the number of virtual functions (VFs) per physical function (PF). `max_vfs` can be `0` (disabled) or `64` VFs on a single port (enabled). The 64 VF maximum support for ESXi is an OS resource allocation constraint.
`RSS`	Specifies the number of receive side scaling queues used by the host or virtual extensible LAN (VXLAN) tunneled traffic for a PF. `RSS` can be `2`, `3`, `4`, or one of the following values: ■ `-1` uses the default number of queues. ■ `0` or `1` disables RSS queues. You can specify multiple values delimited by commas for multiport or multifunction configurations.
`debug`	Specifies the level of data that the driver records in the `vmkernel` log file. `debug` can have the following values, shown in increasing amounts of data: ■ `0x80000000` indicates Notice level. ■ `0x40000000` indicates Information level (includes the Notice level). ■ `0x3FFFFFFF` indicates Verbose level for all driver submodules (includes the Information and Notice levels).
`auto_fw_reset`	Enables (`1`) or disables (`0`) the driver automatic firmware recovery capability. When this parameter is enabled, the driver attempts to recover from events such as transmit timeouts, firmware asserts, and adapter parity errors. The default is `auto_fw_reset=1`.
`vxlan_filter_en`	Enables (`1`) or disables (`0`) the VXLAN filtering based on the outer MAC, the inner MAC, and the VXLAN network (VNI), directly matching traffic to a specific queue. The default is `vxlan_filter_en=1`. You can specify multiple values delimited by commas for multiport or multifunction configurations.
`enable_vxlan_offld`	Enables (`1`) or disables (`0`) the VXLAN tunneled traffic checksum offload and TCP segmentation offload (TSO) capability. The default is `enable_vxlan_offld=1`. You can specify multiple values delimited by commas for multiport or multifunction configurations.

VMware Driver Parameter Defaults

Table 3-7 lists the VMware driver parameter default values.

Table 3-7. VMware Driver Parameter Defaults

Parameter	Default
Speed	Autonegotiation with all speeds advertised. The speed parameter must be the same on all ports. If autonegotiation is enabled on the device, all of the device ports will use autonegotiation.
Flow Control	Autonegotiation with RX and TX advertised
MTU	1,500 (range 46–9,600)
Rx Ring Size	8,192 (range 128–8,192)
Tx Ring Size	8,192 (range 128–8,192)
MSI-X	Enabled
Transmit Send Offload (TSO)	Enabled
Large Receive Offload (LRO)	Enabled
RSS	Enabled (four RX queues)
HW VLAN	Enabled
Number of Queues	Enabled (eight RX/TX queue pairs)
Wake on LAN (WoL)	Disabled

Removing the VMware Driver

To remove the .vib file (qedentv), issue the following command:

```
# esxcli software vib remove --vibname qedentv
```

To remove the driver, issue the following command:

```
# vmkload_mod -u qedentv
```

FCoE Support

The QLogic VMware FCoE qedf driver included in the VMware software package supports QLogic FCoE converged network interface controllers (C-NICs). The driver is a kernel-mode driver that provides a translation layer between the VMware SCSI stack and the QLogic FCoE firmware and hardware. The FCoE and DCB feature set is supported on VMware ESXi 5.0 and later.

iSCSI Support

The QLogic VMware iSCSI qedil HBA driver, similar to qedf, is a kernel mode driver that provides a translation layer between the VMware SCSI stack and the QLogic iSCSI firmware and hardware. The qedil driver leverages the services provided by the VMware iscsid infrastructure for session management and IP services.

4 Upgrading the Firmware

This chapter provides information about upgrading the firmware using the Dell Update Package (DUP).

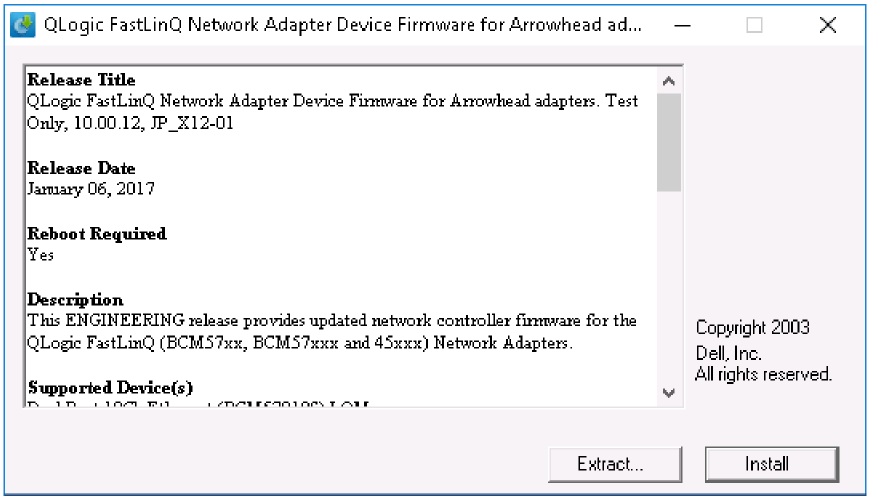
The firmware DUP is a Flash update utility only; it is not used for adapter configuration. You can run the firmware DUP by double-clicking the executable file. Alternatively, you can run the firmware DUP from the command line with several supported command line options.

- Running the DUP by Double-Clicking
- "Running the DUP from a Command Line" on page 36
- "Running the DUP Using the .bin File" on page 37 (Linux only)

Running the DUP by Double-Clicking

To run the firmware DUP by double-clicking the executable file:

1. Double-click the icon representing the firmware Dell Update Package file.
2. The Dell Update Package splash screen appears, as shown in Figure 4-1. Click **Install** to continue.



QLogic FastLinQ Network Adapter Device Firmware for Arrowhead ad...

Release Title
QLogic FastLinQ Network Adapter Device Firmware for Arrowhead adapters. Test Only, 10.00.12, JP_X12-01

Release Date
January 06, 2017

Reboot Required
Yes

Description
This ENGINEERING release provides updated network controller firmware for the QLogic FastLinQ (BCM57xx, BCM57xxx and 45xxx) Network Adapters.

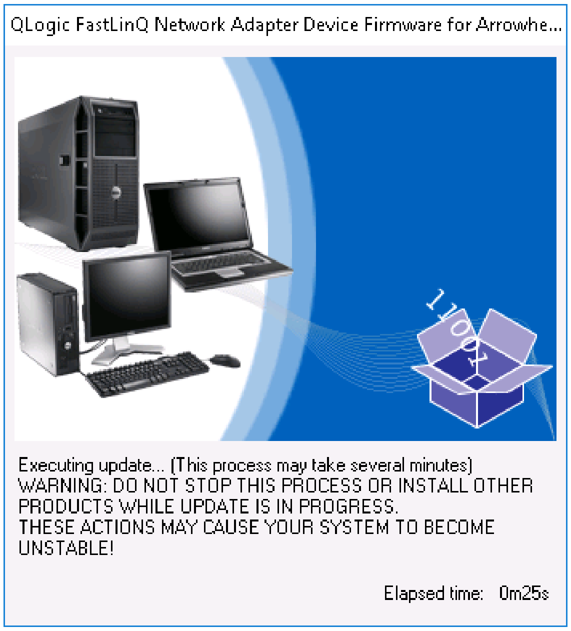
Supported Device(s)

Copyright 2003
Dell, Inc.
All rights reserved.

Figure 4-1. Dell Update Package: Splash Screen

3. Follow the on-screen instructions. In the Warning dialog box, click **Yes** to continue the installation.

   The installer indicates that it is loading the new firmware, as shown in Figure 4-2.



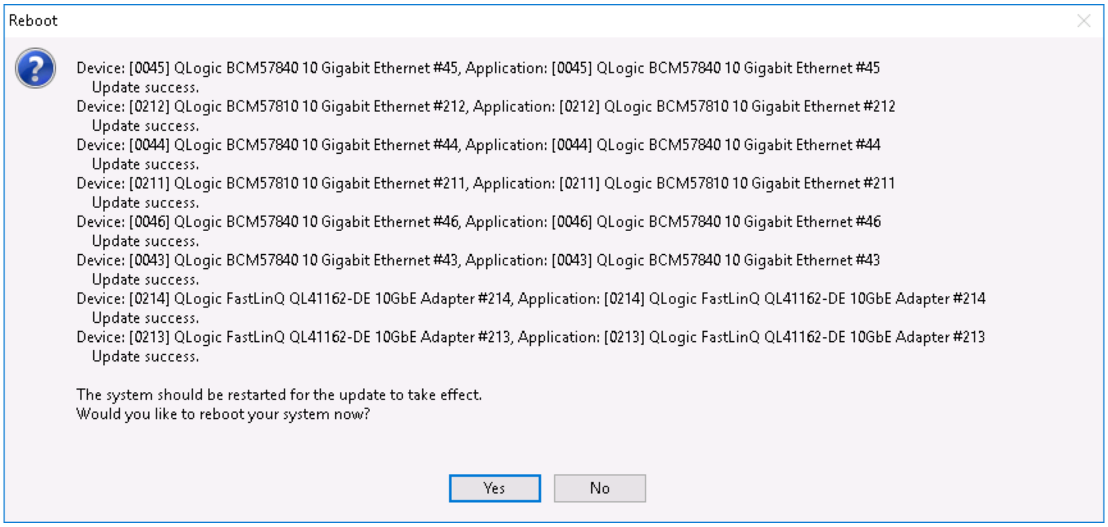
QLogic FastLinQ Network Adapter Device Firmware for Arrowhe...

Executing update... (This process may take several minutes)
WARNING: DO NOT STOP THIS PROCESS OR INSTALL OTHER PRODUCTS WHILE UPDATE IS IN PROGRESS.
THESE ACTIONS MAY CAUSE YOUR SYSTEM TO BECOME UNSTABLE!

Elapsed time: 0m25s

Figure 4-2. Dell Update Package: Loading New Firmware

When complete, the installer indicates the result of the installation, as shown in Figure 4-3.

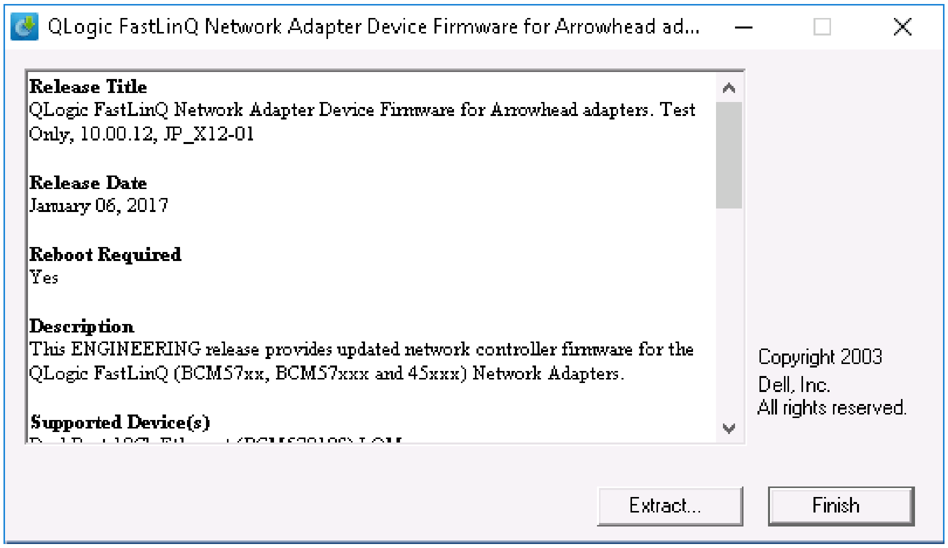


Reboot

```
Device: [0045] QLogic BCM57840 10 Gigabit Ethernet #45, Application: [0045] QLogic BCM57840 10 Gigabit Ethernet #45
  Update success.
Device: [0212] QLogic BCM57810 10 Gigabit Ethernet #212, Application: [0212] QLogic BCM57810 10 Gigabit Ethernet #212
  Update success.
Device: [0044] QLogic BCM57840 10 Gigabit Ethernet #44, Application: [0044] QLogic BCM57840 10 Gigabit Ethernet #44
  Update success.
Device: [0211] QLogic BCM57810 10 Gigabit Ethernet #211, Application: [0211] QLogic BCM57810 10 Gigabit Ethernet #211
  Update success.
Device: [0046] QLogic BCM57840 10 Gigabit Ethernet #46, Application: [0046] QLogic BCM57840 10 Gigabit Ethernet #46
  Update success.
Device: [0043] QLogic BCM57840 10 Gigabit Ethernet #43, Application: [0043] QLogic BCM57840 10 Gigabit Ethernet #43
  Update success.
Device: [0214] QLogic FastLinQ QL41162-DE 10GbE Adapter #214, Application: [0214] QLogic FastLinQ QL41162-DE 10GbE Adapter #214
  Update success.
Device: [0213] QLogic FastLinQ QL41162-DE 10GbE Adapter #213, Application: [0213] QLogic FastLinQ QL41162-DE 10GbE Adapter #213
  Update success.

The system should be restarted for the update to take effect.
Would you like to reboot your system now?
```

Figure 4-3. Dell Update Package: Installation Results

4. Click **Yes** to reboot the system.
5. Click **Finish** to complete the installation, as shown in Figure 4-4.


QLogic FastLinQ Network Adapter Device Firmware for Arrowhead ad...
Extract...
Finish

Release Title
QLogic FastLinQ Network Adapter Device Firmware for Arrowhead adapters. Test Only, 10.00.12, JP_X12-01

Release Date
January 06, 2017

Reboot Required
Yes

Description
This ENGINEERING release provides updated network controller firmware for the QLogic FastLinQ (BCM57xx, BCM57xxx and 45xxx) Network Adapters.

Supported Device(s)

Copyright 2003
Dell, Inc.
All rights reserved.

Figure 4-4. Dell Update Package: Finish Installation

Running the DUP from a Command Line

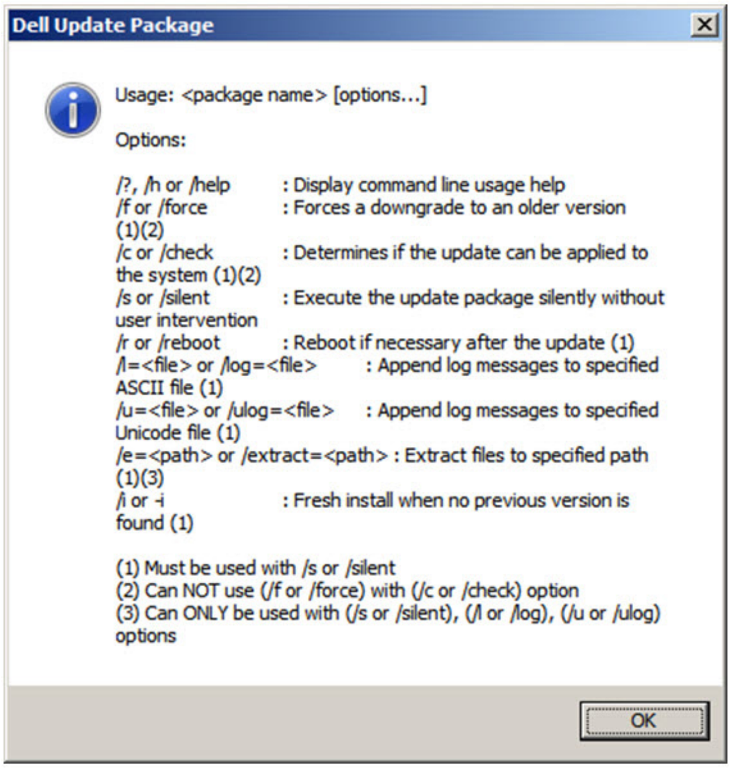
Running the firmware DUP from the command line, with no options specified, results in the same behavior as double-clicking the DUP icon. Note that the actual file name of the DUP will vary.

To run the firmware DUP from a command line:

- Issue the following command:

```
C:\> Network_Firmware_2T12N_WN32_<version>_X16.EXE
```

Figure 4-5 shows the options that you can use to customize the Dell Update Package installation.



```
Usage: <package name> [options...]

Options:

/?, /h or /help        : Display command line usage help
/f or /force           : Forces a downgrade to an older version
(1)(2)
/c or /check           : Determines if the update can be applied to
the system (1)(2)
/s or /silent          : Execute the update package silently without
user intervention
/r or /reboot          : Reboot if necessary after the update (1)
/l=<file> or /log=<file>      : Append log messages to specified
ASCII file (1)
/u=<file> or /ulog=<file>     : Append log messages to specified
Unicode file (1)
/e=<path> or /extract=<path> : Extract files to specified path
(1)(3)
/i or -i               : Fresh install when no previous version is
found (1)

(1) Must be used with /s or /silent
(2) Can NOT use (/f or /force) with (/c or /check) option
(3) Can ONLY be used with (/s or /silent), (/l or /log), (/u or /ulog)
options
```

Figure 4-5. DUP Command Line Options

Running the DUP Using the .bin File

The following procedure is supported only on Linux OS.

To update the DUP using the .bin file:

1. Copy the `Network_Firmware_NJCX1_LN_X.Y.Z.BIN` file to the system under test (SUT).

2. Change the file type into an executable file as follows:

   ```
   chmod 777 Network_Firmware_NJCX1_LN_X.Y.Z.BIN
   ```

3. To start the update process, issue the following command:

   ```
   ./Network_Firmware_NJCX1_LN_X.Y.Z.BIN
   ```

4. After the firmware is updated, reboot the system.

Example output from the SUT during the DUP update:

```
./Network_Firmware_NJCX1_LN_08.07.26.BIN
Collecting inventory...
Running validation...
BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
The version of this Update Package is the same as the currently installed
version.
Software application name: BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
Package version: 08.07.26
Installed version: 08.07.26
BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
The version of this Update Package is the same as the currently installed
version.
Software application name: BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
Package version: 08.07.26
Installed version: 08.07.26
Continue? Y/N:Y
Y entered; update was forced by user
Executing update...
WARNING: DO NOT STOP THIS PROCESS OR INSTALL OTHER DELL PRODUCTS WHILE UPDATE
IS IN PROGRESS.
THESE ACTIONS MAY CAUSE YOUR SYSTEM TO BECOME UNSTABLE!
...........................................................................
Device: BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
  Application: BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
  Update success.
Device: BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
  Application: BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
  Update success.
Would you like to reboot your system now?
Continue? Y/N:Y
```

5 Adapter Preboot Configuration

During the host boot process, you have the opportunity to pause and perform adapter management tasks using the Human Infrastructure Interface (HII) application. These tasks include the following:

- “Getting Started” on page 40
- “Displaying Firmware Image Properties” on page 43
- “Configuring Device-level Parameters” on page 44
- “Configuring NIC Parameters” on page 45
- “Configuring Data Center Bridging” on page 49
- “Configuring FCoE Boot” on page 50
- “Configuring iSCSI Boot” on page 51
- “Configuring Partitions” on page 56

NOTE

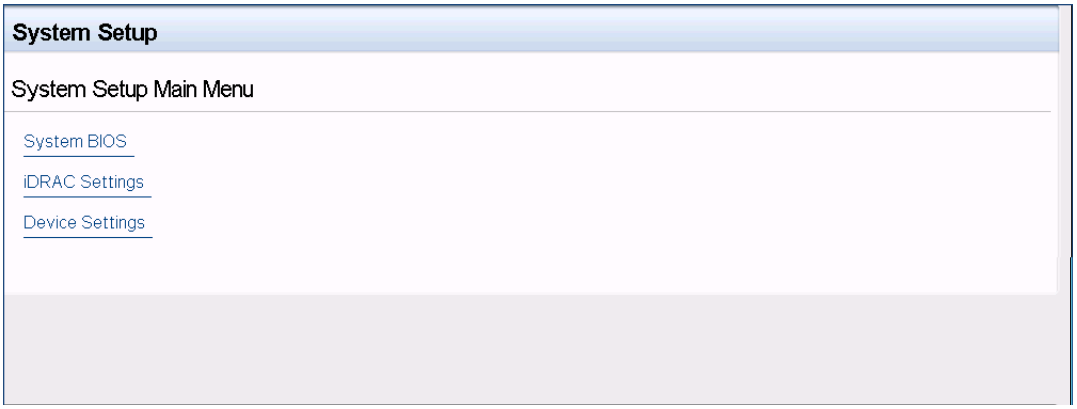
The HII screen shots in this chapter are representative and may not match the screens that you see on your system.

Getting Started

To start the HII application:

1. Open the System Setup window for your platform. For information about launching the System Setup, consult the user guide for your system.

2. In the System Setup window (Figure 5-1), select **Device Settings**, and then press ENTER.



System Setup

System Setup Main Menu

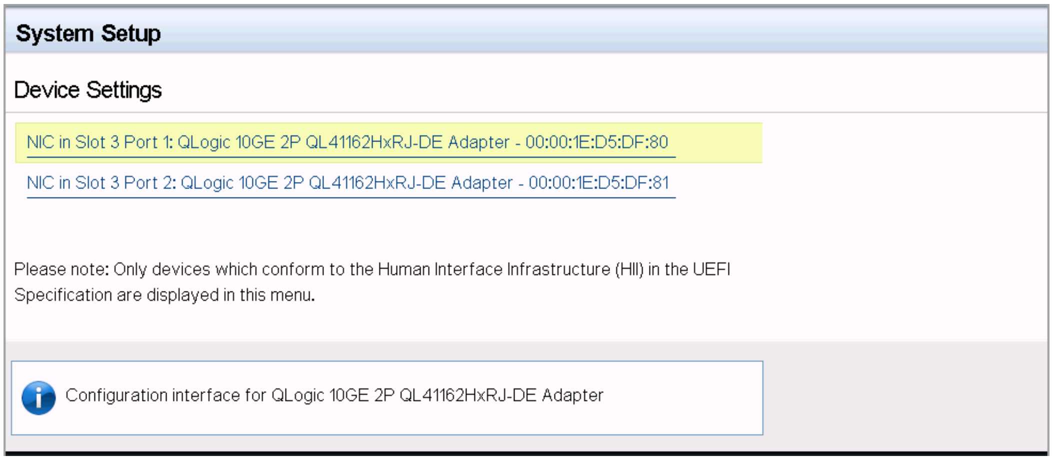
System BIOS

iDRAC Settings

Device Settings

Figure 5-1. System Setup

3. In the Device Settings window (Figure 5-2), select the 41*xxx* Series Adapter port that you want to configure, and then press ENTER.



System Setup

Device Settings

NIC in Slot 3 Port 1: QLogic 10GE 2P QL41162HxRJ-DE Adapter - 00:00:1E:D5:DF:80

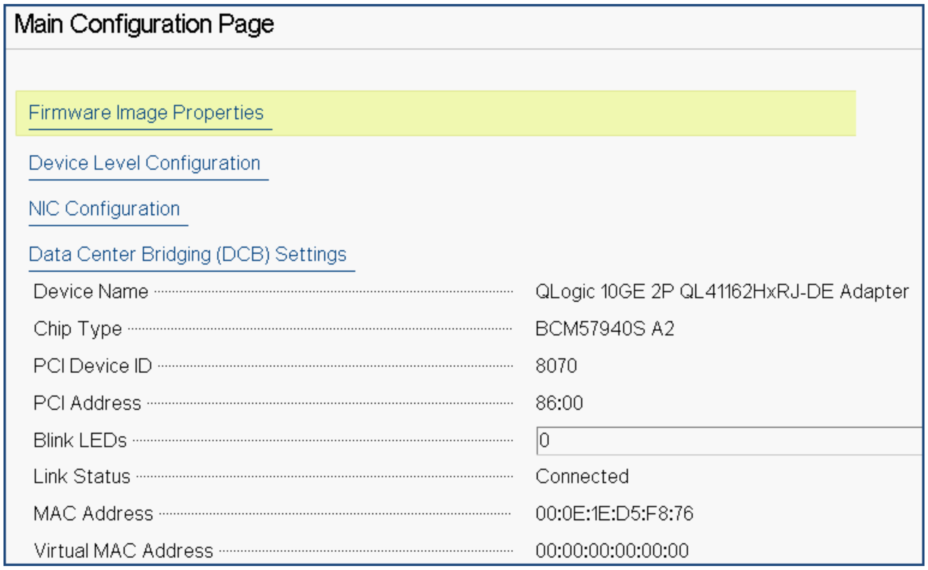
NIC in Slot 3 Port 2: QLogic 10GE 2P QL41162HxRJ-DE Adapter - 00:00:1E:D5:DF:81

Please note: Only devices which conform to the Human Interface Infrastructure (HII) in the UEFI Specification are displayed in this menu.

Configuration interface for QLogic 10GE 2P QL41162HxRJ-DE Adapter

Figure 5-2. System Setup: Device Settings

The Main Configuration Page (Figure 5-3) presents the adapter management options where you can set the partitioning mode.



Main Configuration Page

Firmware Image Properties

Device Level Configuration

NIC Configuration

Data Center Bridging (DCB) Settings

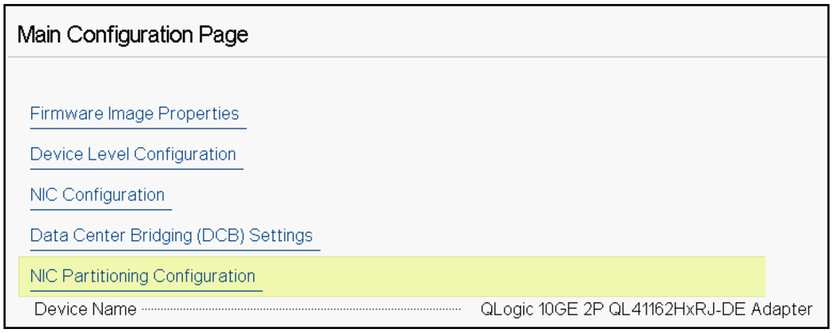
Device Name .... QLogic 10GE 2P QL41162HxRJ-DE Adapter
Chip Type BCM57940S A2
PCI Device ID 8070
PCI Address 86:00
Blink LEDs 0
Link Status .... Connected
MAC Address 00:0E:1E:D5:F8:76
Virtual MAC Address 00:00:00:00:00:00

Figure 5-3. Main Configuration Page

4. Under **Device Level Configuration**, set the **Partitioning Mode** to **NPAR** to add the **NIC Partitioning Configuration** option to the Main Configuration Page, as shown in Figure 5-4.

NOTE

NPAR is not available on ports with a maximum speed of 1G.



Main Configuration Page

Firmware Image Properties

Device Level Configuration

NIC Configuration

Data Center Bridging (DCB) Settings

NIC Partitioning Configuration

Device Name .... QLogic 10GE 2P QL41162HxRJ-DE Adapter

Figure 5-4. Main Configuration Page, Setting Partitioning Mode to NPAR

In Figure 5-3 and Figure 5-4, the Main Configuration Page shows the following:

- **Firmware Image Properties** (see "Displaying Firmware Image Properties" on page 43)

- **Device Level Configuration** (see "Configuring Device-level Parameters" on page 44)
- **NIC Configuration** (see "Configuring NIC Parameters" on page 45)
- **iSCSI Configuration** (if iSCSI remote boot is allowed by enabling iSCSI offload in NPAR mode on the port's third partition) (see "Configuring iSCSI Boot" on page 51)
- **FCoE Configuration** (if FCoE boot from SAN is allowed by enabling FCoE offload in NPAR mode on the port's second partition) (see "Configuring FCoE Boot" on page 50)
- **Data Center Bridging (DCB) Settings** (see "Configuring Data Center Bridging" on page 49)
- **NIC Partitioning Configuration** (if **NPAR** is selected on the Device Level Configuration page) (see "Configuring Partitions" on page 56)

In addition, the Main Configuration Page presents the adapter properties listed in Table 5-1.

Table 5-1. Adapter Properties

Adapter Property	Description
Device Name	Factory-assigned device name
Chip Type	ASIC version
PCI Device ID	Unique vendor-specific PCI device ID
PCI Address	PCI device address in bus-device function format
Blink LEDs	User-defined blink count for the port LED
Link Status	External link status
MAC Address	Manufacturer-assigned permanent device MAC address
Virtual MAC Address	User-defined device MAC address
iSCSI MAC Address [a]	Manufacturer-assigned permanent device iSCSI Offload MAC address
iSCSI Virtual MAC Address [a]	User-defined device iSCSI Offload MAC address
FCoE MAC Address [b]	Manufacturer-assigned permanent device FCoE Offload MAC address
FCoE Virtual MAC Address [b]	User-defined device FCoE Offload MAC address

Table 5-1. Adapter Properties (Continued)

Adapter Property	Description
FCoE WWPN [b]	Manufacturer-assigned permanent device FCoE Offload WWPN (world wide port name)
FCoE Virtual WWPN [b]	User-defined device FCoE Offload WWPN
FCoE WWNN [b]	Manufacturer-assigned permanent device FCoE Offload WWNN (world wide node name)
FCoE Virtual WWNN [b]	User-defined device FCoE Offload WWNN

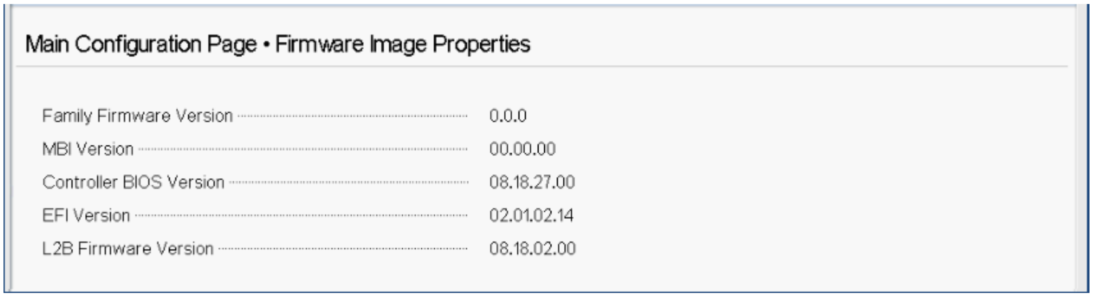
[a] This property is visible only if **iSCSI Offload** is enabled on the NIC Partitioning Configuration page.

[b] This property is visible only if **FCoE Offload** is enabled on the NIC Partitioning Configuration page.

Displaying Firmware Image Properties

To view the properties for the firmware image, select **Firmware Image Properties** on the Main Configuration Page, and then press ENTER. The Firmware Image Properties page (Figure 5-5) specifies the following view-only data:

- **Family Firmware Version** is the multiboot image version, which comprises several firmware component images.
- **MBI Version** is the Cavium QLogic bundle image version that is active on the device.
- **Controller BIOS Version** is the management firmware version.
- **EFI Driver Version** is the extensible firmware interface (EFI) driver version.
- **L2B Firmware Version** is the NIC offload firmware version for boot.



Main Configuration Page • Firmware Image Properties

```
Family Firmware Version ........ 0.0.0
MBI Version .................... 00.00.00
Controller BIOS Version ........ 08.18.27.00
EFI Version .................... 02.01.02.14
L2B Firmware Version ........... 08.18.02.00
```

Figure 5-5. Firmware Image Properties

Configuring Device-level Parameters

NOTE

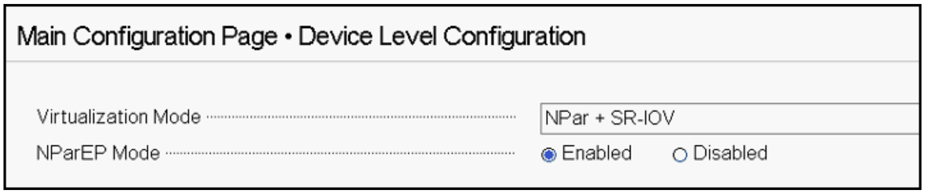
The iSCSI physical functions (PFs) are listed when the iSCSI Offload feature is enabled in NPAR mode only. The FCoE PFs are listed when the FCoE Offload feature is enabled in NPAR mode only. Not all adapter models support iSCSI Offload and FCoE Offload. Only one offload can be enabled per port, and only in NPAR mode.

Device-level configuration includes the following parameters:

- **Virtualization Mode**
- **NPAREP Mode**

To configure device-level parameters:

1. On the Main Configuration Page, select **Device Level Configuration** (see Figure 5-3 on page 41), and then press ENTER.
2. On the **Device Level Configuration** page, select values for the device-level parameters, as shown in Figure 5-6.



Main Configuration Page • Device Level Configuration

Virtualization Mode: NPar + SR-IOV

NParEP Mode: ● Enabled ○ Disabled

Figure 5-6. Device Level Configuration

NOTE

QL41264HMCU-DE (part number 5V6Y4) and QL41264HMRJ-DE (part number 0D1WT) adapters show support for NPAR, SR-IOV and NPAR-EP in the Device Level Configuration, though these features are not supported on 1Gbps ports 3 and 4.

3. For **Virtualization Mode**, select one of the following modes to apply to all adapter ports:
   - **None** (default) specifies that no virtualization mode is enabled.
   - **NPAR** sets the adapter to switch-independent NIC partitioning mode.
   - **SR-IOV** sets the adapter to SR-IOV mode.
   - **NPar + SR-IOV** sets the adapter to SR-IOV over NPAR mode.

4. **NParEP Mode** configures the maximum quantity of partitions per adapter. This parameter is visible when you select either **NPAR** or **NPar + SR-IOV** as the **Virtualization Mode** in Step 2.

   - **Enabled** allows you to configure up to 16 partitions per adapter.
   - **Disabled** allows you to configures up to 8 partitions per adapter.

5. Click **Back**.

6. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

Configuring NIC Parameters

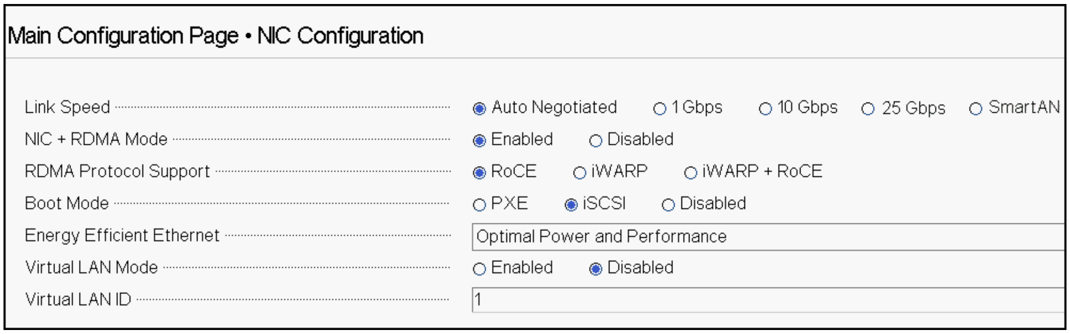
NIC configuration includes setting the following parameters:

- **Link Speed**
- **NIC + RDMA Mode**
- **RDMA Protocol Support**
- **Boot Mode**
- **FEC Mode**
- **Energy Efficient Ethernet**
- **Virtual LAN Mode**
- **Virtual LAN ID**

To configure NIC parameters:

1. On the Main Configuration Page, select **NIC Configuration** (Figure 5-3 on page 41), and then click **Finish**.

   Figure 5-7 shows the NIC Configuration page.



Main Configuration Page • NIC Configuration

Link Speed ● Auto Negotiated ○ 1 Gbps ○ 10 Gbps ○ 25 Gbps ○ SmartAN
NIC + RDMA Mode ● Enabled ○ Disabled
RDMA Protocol Support ● RoCE ○ iWARP ○ iWARP + RoCE
Boot Mode ○ PXE ● iSCSI ○ Disabled
Energy Efficient Ethernet Optimal Power and Performance
Virtual LAN Mode ○ Enabled ● Disabled
Virtual LAN ID 1

Figure 5-7. NIC Configuration

2. Select one of the following **Link Speed** options for the selected port. Not all speed selections are available on all adapters.
    - **Auto Negotiated** enables Auto Negotiation mode on the port. FEC mode selection is not available for this speed mode.
    - **1 Gbps** enables 1GbE fixed speed mode on the port. This mode is intended only for 1GbE interfaces and should not be configured for adapter interfaces that operate at other speeds. FEC mode selection is not available for this speed mode. This mode is not available on all adapters.
    - **10 Gbps** enables 10GbE fixed speed mode on the port. This mode is not available on all adapters.
    - **25 Gbps** enables 25GbE fixed speed mode on the port. This mode is not available on all adapters.
    - **SmartAN** (Default) enables FastLinQ SmartAN™ link speed mode on the port. No FEC mode selection is available for this speed mode. The **SmartAN** setting cycles through all possible link speeds and FEC modes until a link is established. This mode is intended for use only with 25G interfaces. If you configure SmartAN for a 10Gb interface, the system will apply settings for a 10G interface.This mode is not available on all adapters.
3. For **NIC + RDMA Mode**, select either **Enabled** or **Disabled** for RDMA on the port. This setting applies to all partitions of the port, if in NPAR mode.
4. **FEC Mode** is visible when **25 Gbps** fixed speed mode is selected as the **Link Speed** in Step 2. For **FEC Mode**, select one of the following options. Not all FEC modes are available on all adapters.
    - **None** disables all FEC modes.
    - **Fire Code** enables Fire Code (BASE-R) FEC mode.
    - **Reed Solomon** enables Reed Solomon FEC mode.
    - **Auto** enables the port to cycle through **None**, **Fire Code**, and **Reed Solomon** FEC modes (at that link speed) in a round-robin fashion, until a link is established.

5. The **RDMA Protocol Support** setting applies to all partitions of the port, if in NPAR mode. This setting appears if the **NIC + RDMA Mode** in Step 3 is set to **Enabled**. **RDMA Protocol Support** options include the following:

   ❑ **RoCE** enables RoCE mode on this port.

   ❑ **iWARP** enables iWARP mode on this port.

   ❑ **iWARP + RoCE** enables iWARP and RoCE modes on this port. This is the default. Additional configuration for Linux is required for this option as described in "Configuring iWARP and RoCE" on page 104.

6. For **Boot Mode**, select one of the following values:

   ❑ **PXE** enables PXE boot.

   ❑ **FCoE** enables FCoE boot from SAN over the hardware offload pathway. The **FCoE** mode is available only if **FCoE Offload** is enabled on the second partition in NPAR mode (see "Configuring Partitions" on page 56).

   ❑ **iSCSI** enables iSCSI remote boot over the hardware offload pathway. The **iSCSI** mode is available only if **iSCSI Offload** is enabled on the third partition in NPAR mode (see "Configuring Partitions" on page 56).

   ❑ **Disabled** prevents this port from being used as a remote boot source.

7. The **Energy Efficient Ethernet** (EEE) parameter is visible only on 100BASE-T or 10GBASE-T RJ45 interfaced adapters. Select from the following EEE options:

   ❑ **Disabled** disables EEE on this port.

   ❑ **Optimal Power and Performance** enables EEE in optimal power and performance mode on this port.

   ❑ **Maximum Power Savings** enables EEE in maximum power savings mode on this port.

   ❑ **Maximum Performance** enables EEE in maximum performance mode on this port.

8. The **Virtual LAN Mode** parameter applies to the entire port when in PXE remote install mode. It is not persistent after a PXE remote install finishes. Select from the following VLAN options:

   ❑ **Enabled** enables VLAN mode on this port for PXE remote install mode.

   ❑ **Disabled** disables VLAN mode on this port.

9. The **Virtual LAN ID** parameter specifies the VLAN tag ID to be used on this port for PXE remote install mode. This setting applies only when **Virtual LAN Mode** is enabled in the previous step.
10. Click **Back**.
11. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

To configure the port to use RDMA:

NOTE

Follow these steps to enable RDMA on all partitions of an NPAR mode port.

1. Set **NIC + RDMA Mode** to **Enabled**.
2. Click **Back**.
3. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

To configure the port's boot mode:

1. For a UEFI PXE remote installation, select **PXE** as the **Boot Mode**.
2. Click **Back**.
3. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

To configure the port’s PXE remote install to use a VLAN:

NOTE

This VLAN is not persistent after the PXE remote install is finished.

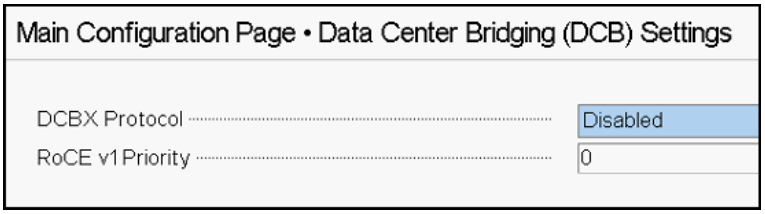
1. Set the **Virtual LAN Mode** to **Enabled**.
2. In the **Virtual LAN ID** box, enter the number to be used.
3. Click **Back**.
4. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

Configuring Data Center Bridging

The data center bridging (DCB) settings comprise the DCBX protocol and the RoCE priority.

To configure the DCB settings:

1. On the Main Configuration Page (Figure 5-3 on page 41), select **Data Center Bridging (DCB) Settings**, and then click **Finish**.
2. On the Data Center Bridging (DCB) Settings page (Figure 5-8), select the appropriate **DCBX Protocol** option:
   - ❑ **Disabled** disables DCBX on this port.
   - ❑ **CEE** enables the legacy Converged Enhanced Ethernet (CEE) protocol DCBX mode on this port.
   - ❑ **IEEE** enables the IEEE DCBX protocol on this port.
   - ❑ **Dynamic** enables dynamic application of either the CEE or IEEE protocol to match the attached link partner.
3. On the Data Center Bridging (DCB) Settings page, enter the **RoCE v1 Priority** as a value from **0–7**. This setting indicates the DCB traffic class priority number used for RoCE traffic and should match the number used by the DCB-enabled switching network for RoCE traffic.
   - ❑ **0** specifies the usual priority number used by the lossy default or common traffic class.
   - ❑ **3** specifies the priority number used by lossless FCoE traffic.
   - ❑ **4** specifies the priority number used by lossless iSCSI-TLV over DCB traffic.
   - ❑ **1**, **2**, **5**, **6**, and **7** specify DCB traffic class priority numbers available for RoCE use. Follow the respective OS RoCE setup instructions for using this RoCE control.



Main Configuration Page • Data Center Bridging (DCB) Settings

DCBX Protocol: Disabled
RoCE v1 Priority: 0

Figure 5-8. System Setup: Data Center Bridging (DCB) Settings

4. Click **Back**.

5. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

NOTE

When DCBX is enabled, the adapter periodically sends link layer discovery protocol (LLDP) packets with a dedicated unicast address that serves as the source MAC address. This LLDP MAC address is different from the factory-assigned adapter Ethernet MAC address. If you examine the MAC address table for the switch port that is connected to the adapter, you will see two MAC addresses: one for LLDP packets and one for the adapter Ethernet interface.

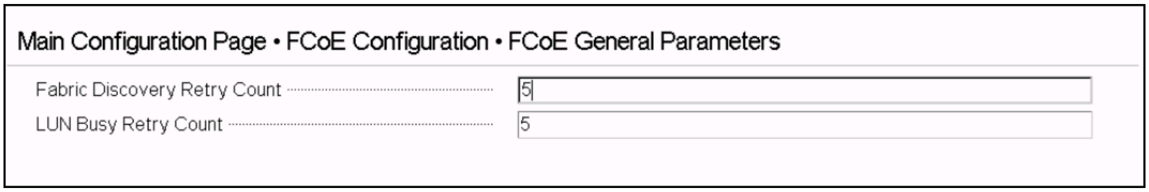
Configuring FCoE Boot

NOTE

The FCoE Boot Configuration Menu is only visible if **FCoE Offload Mode** is enabled on the second partition in NPAR mode (see Figure 5-18 on page 59). It is not visible in non-NPAR mode.

To configure the FCoE boot configuration parameters:

1. On the Main Configuration Page, select **FCoE Configuration**, and then select the following as needed:
   - ❑ **FCoE General Parameters** (Figure 5-9)
   - ❑ **FCoE Target Configuration** (Figure 5-10)
2. Press ENTER.
3. Choose values for the FCoE General or FCoE Target Configuration parameters.

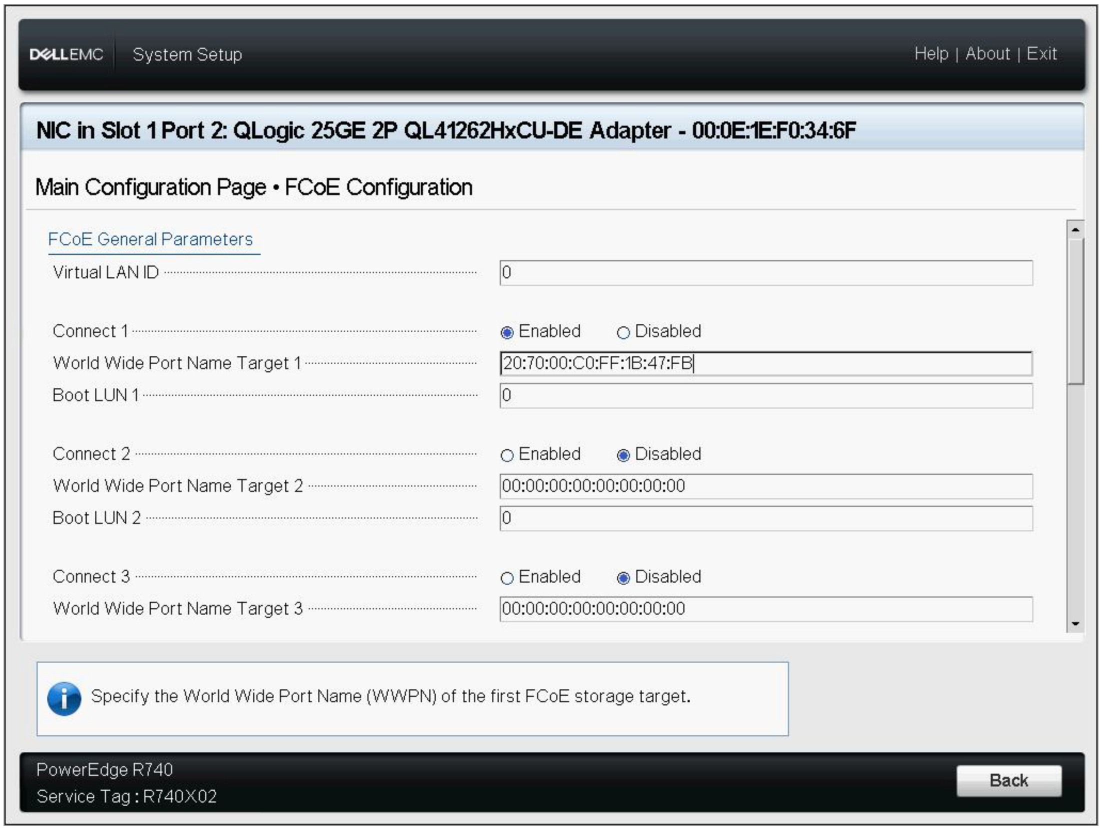


Main Configuration Page • FCoE Configuration • FCoE General Parameters

Fabric Discovery Retry Count 5

LUN Busy Retry Count 5

Figure 5-9. FCoE General Parameters



NIC in Slot 1 Port 2: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6F

Main Configuration Page • FCoE Configuration

FCoE General Parameters

Virtual LAN ID	0
Connect 1	● Enabled ○ Disabled
World Wide Port Name Target 1	20:70:00:C0:FF:1B:47:FB
Boot LUN 1	0
Connect 2	○ Enabled ● Disabled
World Wide Port Name Target 2	00:00:00:00:00:00:00:00
Boot LUN 2	0
Connect 3	○ Enabled ● Disabled
World Wide Port Name Target 3	00:00:00:00:00:00:00:00

Specify the World Wide Port Name (WWPN) of the first FCoE storage target.

PowerEdge R740
Service Tag : R740X02

Figure 5-10. FCoE Target Configuration

4. Click **Back**.
5. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

Configuring iSCSI Boot

NOTE

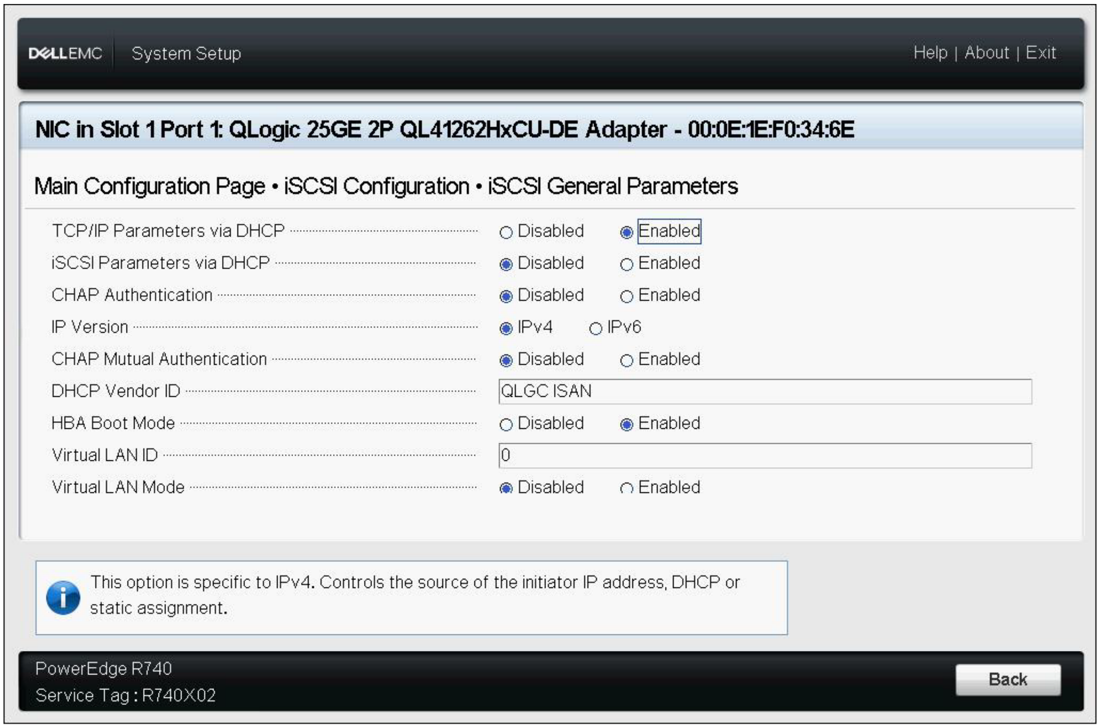
The iSCSI Boot Configuration Menu is only visible if **iSCSI Offload Mode** is enabled on the third partition in NPAR mode (see Figure 5-19 on page 60). It is not visible in non-NPAR mode.

To configure the iSCSI boot configuration parameters:

1. On the Main Configuration Page, select **iSCSI Boot Configuration Menu**, and then select one of the following options:
    - **iSCSI General Configuration**
    - **iSCSI Initiator Configuration**
    - **iSCSI First Target Configuration**
    - **iSCSI Second Target Configuration**
2. Press ENTER.
3. Choose values for the appropriate iSCSI configuration parameters:
    - **iSCSI General Parameters** (Figure 5-11 on page 53)
        - TCP/IP Parameters Via DHCP
        - iSCSI Parameters Via DHCP
        - CHAP Authentication
        - CHAP Mutual Authentication
        - IP Version
        - ARP Redirect
        - DHCP Request Timeout
        - Target Login Timeout
        - DHCP Vendor ID
    - **iSCSI Initiator Parameters** (Figure 5-12 on page 54)
        - IPv4 Address
        - IPv4 Subnet Mask
        - IPv4 Default Gateway
        - IPv4 Primary DNS
        - IPv4 Secondary DNS
        - VLAN ID
        - iSCSI Name
        - CHAP ID
        - CHAP Secret
    - **iSCSI First Target Parameters** (Figure 5-13 on page 54)
        - Connect
        - IPv4 Address
        - TCP Port
        - Boot LUN
        - iSCSI Name
        - CHAP ID
        - CHAP Secret

- ❑ **iSCSI Second Target Parameters** (Figure 5-14 on page 55)
  - Connect
  - IPv4 Address
  - TCP Port
  - Boot LUN
  - iSCSI Name
  - CHAP ID
  - CHAP Secret

4. Click **Back**.
5. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

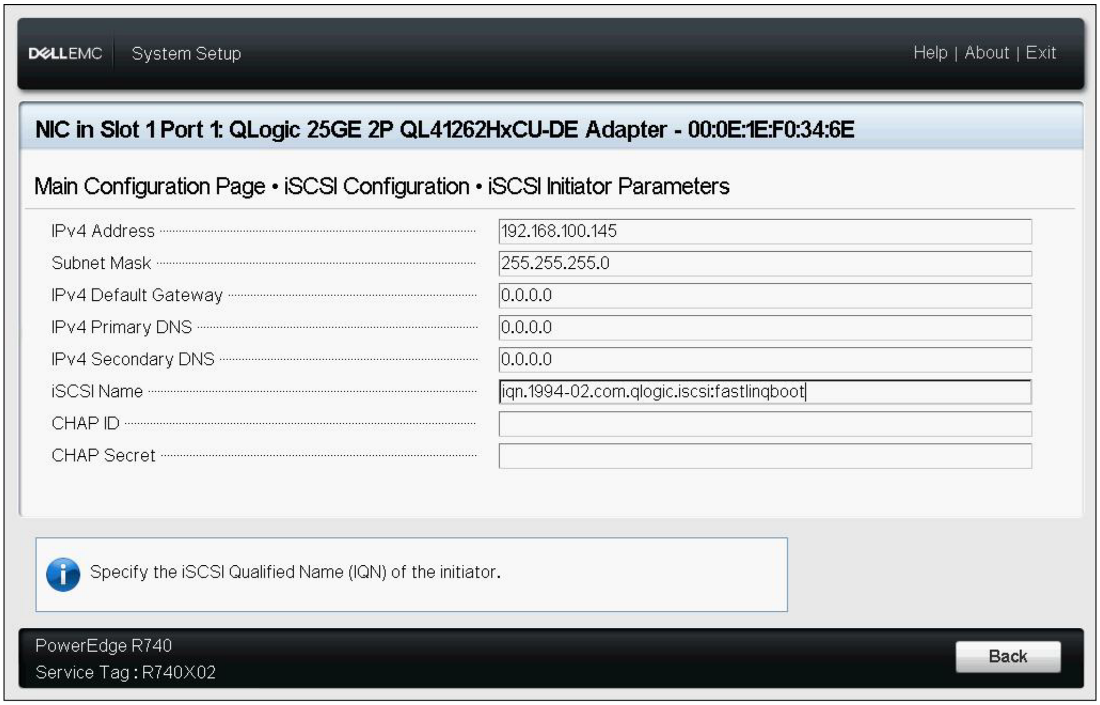
Main Configuration Page • iSCSI Configuration • iSCSI General Parameters

TCP/IP Parameters via DHCP	○ Disabled	● Enabled
iSCSI Parameters via DHCP	● Disabled	○ Enabled
CHAP Authentication	● Disabled	○ Enabled
IP Version	● IPv4	○ IPv6
CHAP Mutual Authentication	● Disabled	○ Enabled
DHCP Vendor ID	QLGC ISAN	
HBA Boot Mode	○ Disabled	● Enabled
Virtual LAN ID	0	
Virtual LAN Mode	● Disabled	○ Enabled

This option is specific to IPv4. Controls the source of the initiator IP address, DHCP or static assignment.

PowerEdge R740
Service Tag : R740X02

Figure 5-11. iSCSI General Parameters



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

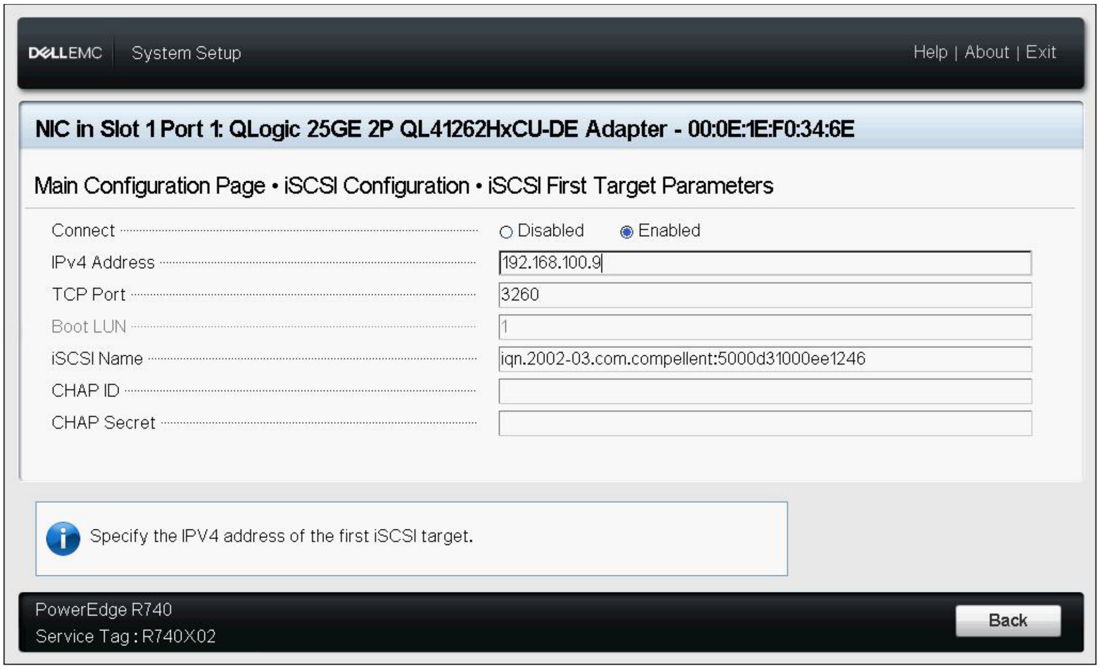
Main Configuration Page • iSCSI Configuration • iSCSI Initiator Parameters

IPv4 Address	192.168.100.145
Subnet Mask	255.255.255.0
IPv4 Default Gateway	0.0.0.0
IPv4 Primary DNS	0.0.0.0
IPv4 Secondary DNS	0.0.0.0
iSCSI Name	iqn.1994-02.com.qlogic.iscsi:fastlinqboot
CHAP ID	
CHAP Secret	

Specify the iSCSI Qualified Name (IQN) of the initiator.

PowerEdge R740
Service Tag : R740X02

Figure 5-12. iSCSI Initiator Configuration Parameters



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

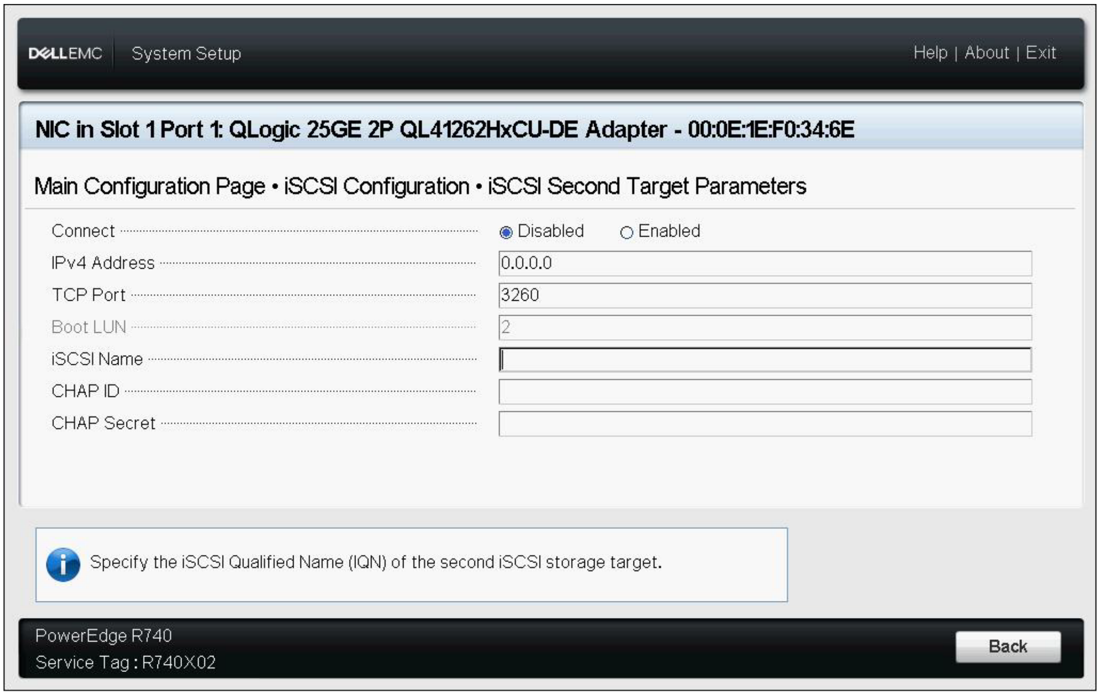
Main Configuration Page • iSCSI Configuration • iSCSI First Target Parameters

Connect	○ Disabled ● Enabled
IPv4 Address	192.168.100.9
TCP Port	3260
Boot LUN	1
iSCSI Name	iqn.2002-03.com.compellent:5000d31000ee1246
CHAP ID	
CHAP Secret	

Specify the IPV4 address of the first iSCSI target.

PowerEdge R740
Service Tag : R740X02

Figure 5-13. iSCSI First Target Parameters



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration • iSCSI Second Target Parameters

Connect	● Disabled ○ Enabled
IPv4 Address	0.0.0.0
TCP Port	3260
Boot LUN	2
iSCSI Name	
CHAP ID	
CHAP Secret	

Specify the iSCSI Qualified Name (IQN) of the second iSCSI storage target.

PowerEdge R740
Service Tag : R740X02

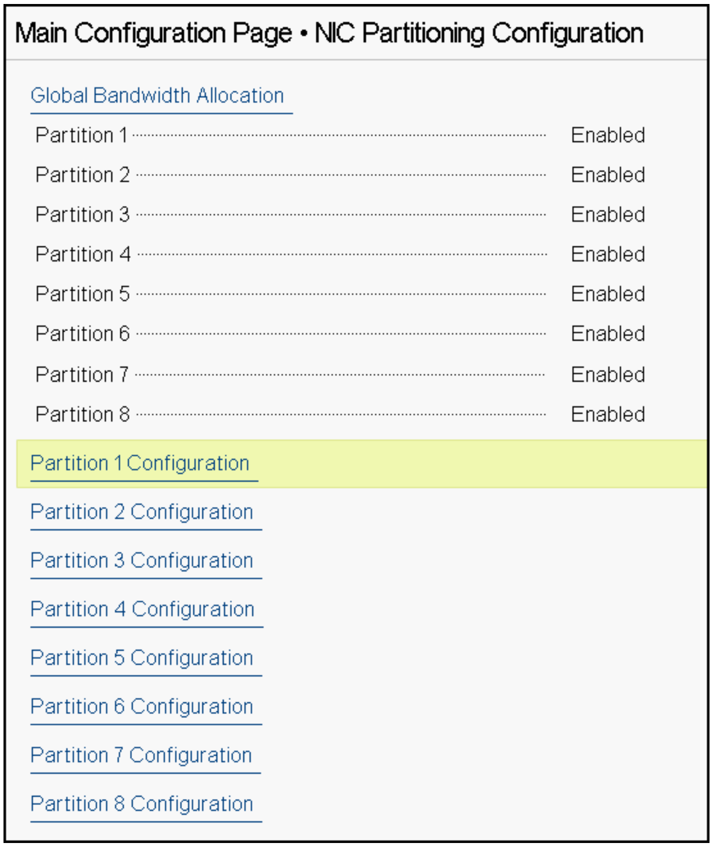
Figure 5-14. iSCSI Second Target Parameters

Configuring Partitions

You can configure bandwidth ranges for each partition on the adapter. For information specific to partition configuration on VMware ESXi 6.0/6.5, see Partitioning for VMware ESXi 6.0 and ESXi 6.5.

To configure the maximum and minimum bandwidth allocations:

1. On the Main Configuration Page, select **NIC Partitioning Configuration**, and then press ENTER.
2. On the Partitions Configuration page (Figure 5-15), select **Global Bandwidth Allocation**.



Main Configuration Page • NIC Partitioning Configuration

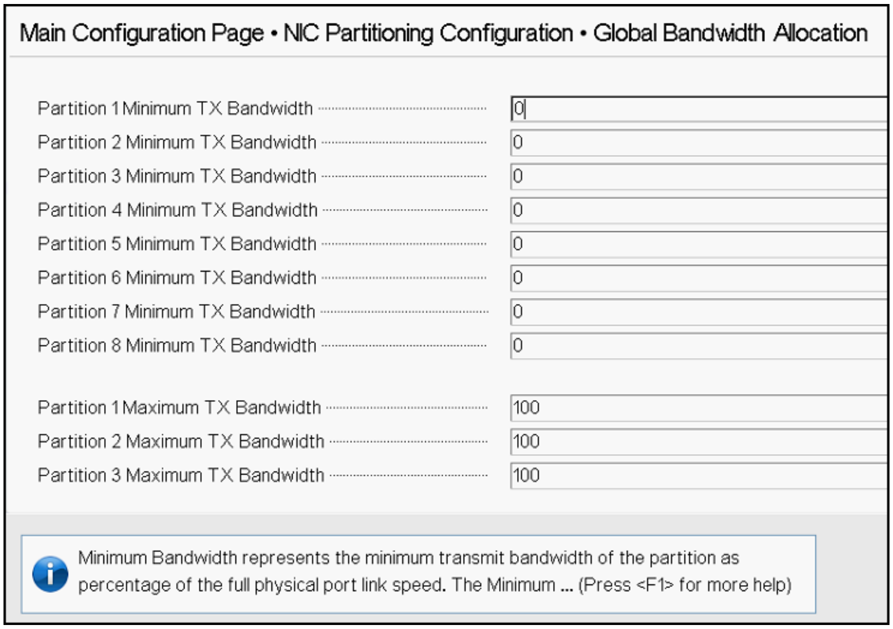
Global Bandwidth Allocation

Partition 1	Enabled
Partition 2	Enabled
Partition 3	Enabled
Partition 4	Enabled
Partition 5	Enabled
Partition 6	Enabled
Partition 7	Enabled
Partition 8	Enabled

Partition 1 Configuration
Partition 2 Configuration
Partition 3 Configuration
Partition 4 Configuration
Partition 5 Configuration
Partition 6 Configuration
Partition 7 Configuration
Partition 8 Configuration

Figure 5-15. NIC Partitioning Configuration, Global Bandwidth Allocation

3. On the Global Bandwidth Allocation page (Figure 5-16), click each partition minimum and maximum TX bandwidth field for which you want to allocate bandwidth. There are eight partitions per port in dual-port mode.



Main Configuration Page • NIC Partitioning Configuration • Global Bandwidth Allocation

Partition 1 Minimum TX Bandwidth	0
Partition 2 Minimum TX Bandwidth	0
Partition 3 Minimum TX Bandwidth	0
Partition 4 Minimum TX Bandwidth	0
Partition 5 Minimum TX Bandwidth	0
Partition 6 Minimum TX Bandwidth	0
Partition 7 Minimum TX Bandwidth	0
Partition 8 Minimum TX Bandwidth	0
Partition 1 Maximum TX Bandwidth	100
Partition 2 Maximum TX Bandwidth	100
Partition 3 Maximum TX Bandwidth	100

Minimum Bandwidth represents the minimum transmit bandwidth of the partition as percentage of the full physical port link speed. The Minimum ... (Press <F1> for more help)

Figure 5-16. Global Bandwidth Allocation Page

- **Partition *n* Minimum TX Bandwidth** is the minimum transmit bandwidth of the selected partition expressed as a percentage of the maximum physical port link speed. Values can be `0`–`100`. When DCBX ETS mode is enabled, the per-traffic class DCBX ETS minimum bandwidth value is used simultaneously with the per-partition minimum TX bandwidth value.The total of the minimum TX bandwidth values of all partitions on a single port must equal 100 or be all zeros.

  Setting the TX minimum bandwidth to all zeros is similar to equally dividing the available bandwidth over every active partition; however, the bandwidth is dynamically allocated over all actively sending partitions. A zero value (when one or more of the other values are set to a non-zero value) allocates a minimum of one percent to that partition, when congestion (from all of the partitions) is restricting TX bandwidth.

- **Partition *n* Maximum TX Bandwidth** is the maximum transmit bandwidth of the selected partition expressed as a percentage of the maximum physical port link speed. Values can be `1`–`100`. The per-partition maximum TX bandwidth value applies regardless of the DCBX ETS mode setting.

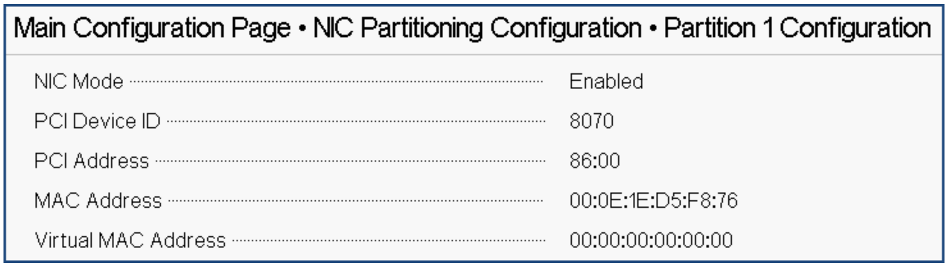
Type a value in each selected field, and then click **Back**.

4. When prompted, click **Yes** to save the changes. Changes take effect after a system reset.

To configure partitions:

1. To examine a specific partition configuration, on the NIC Partitioning Configuration page (Figure 5-15 on page 56), select **Partition *n* Configuration**. If NParEP is not enabled, only four partitions exist per port.
2. To configure the first partition, select **Partition 1 Configuration** to open the Partition 1 Configuration page (Figure 5-17), which shows the following parameters:
   - ❑ **NIC Mode** (always enabled)
   - ❑ **PCI Device ID**
   - ❑ **PCI** (bus) **Address**
   - ❑ **MAC Address**
   - ❑ **Virtual MAC Address**

   If NParEP is not enabled, only four partitions per port are available. On non-offload-capable adapters, the **FCoE Mode** and **iSCSI Mode** options and information are not displayed.



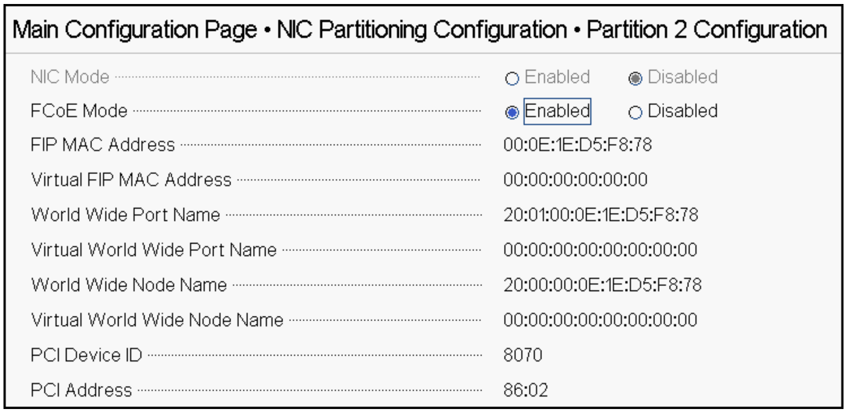
Main Configuration Page • NIC Partitioning Configuration • Partition 1 Configuration

```
NIC Mode ........................... Enabled
PCI Device ID ...................... 8070
PCI Address ........................ 86:00
MAC Address ........................ 00:0E:1E:D5:F8:76
Virtual MAC Address ................ 00:00:00:00:00:00
```

Figure 5-17. Partition 1 Configuration

3. To configure the second partition, select **Partition 2 Configuration** to open the Partition 2 Configuration page. If FCoE Offload is present, the Partition 2 Configuration (Figure 5-18) shows the following parameters:
   - ❑ **NIC Mode** enables or disables the L2 Ethernet NIC personality on Partitions 2 and greater. To disable any of the remaining partitions, set the **NIC Mode** to **Disabled**. To disable offload-capable partitions, disable both the **NIC Mode** and respective offload mode.

- **FCoE Mode** enables or disables the FCoE-Offload personality on the second partition. If you enable this mode on the second partition, you should disable **NIC Mod**e. Because only one offload is available per port, if FCoE-Offload is enabled on the port's second partition, iSCSI-Offload cannot be enabled on the third partition of that same NPAR mode port. Not all adapters support **FCoE Mode**.
- **iSCSI Mode** enables or disables the iSCSI-Offload personality on the third partition. If you enable this mode on the third partition, you should disable **NIC Mode**. Because only one offload is available per port, if iSCSI-Offload is enabled on the port's third partition, FCoE-Offload cannot be enabled on the second partition of that same NPAR mode port. Not all adapters support **iSCSI Mode**.
- **FIP MAC Address**[1]
- **Virtual FIP MAC Address**[1]
- **World Wide Port Name**[1]
- **Virtual World Wide Port Name**[1]
- **World Wide Node Name**[1]
- **Virtual World Wide Node Name**[1]
- **PCI Device ID**
- **PCI** (bus) **Address**



Main Configuration Page • NIC Partitioning Configuration • Partition 2 Configuration

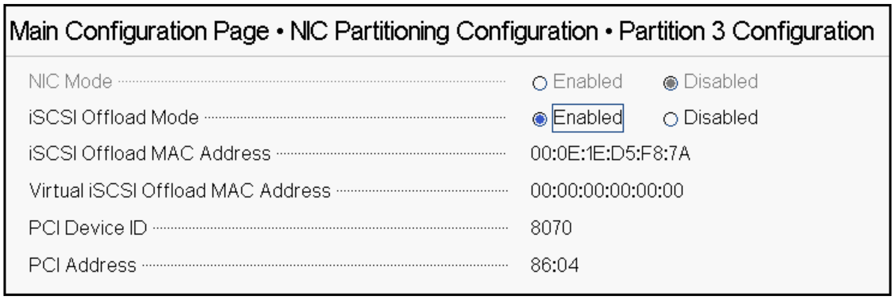
NIC Mode ○ Enabled ● Disabled
FCoE Mode ● Enabled ○ Disabled
FIP MAC Address 00:0E:1E:D5:F8:78
Virtual FIP MAC Address 00:00:00:00:00:00
World Wide Port Name 20:01:00:0E:1E:D5:F8:78
Virtual World Wide Port Name 00:00:00:00:00:00:00:00
World Wide Node Name 20:00:00:0E:1E:D5:F8:78
Virtual World Wide Node Name 00:00:00:00:00:00:00:00
PCI Device ID 8070
PCI Address 86:02

Figure 5-18. Partition 2 Configuration: FCoE Offload

[1] This parameter is only present on the second partition of an NPAR mode port of FCoE offload-capable adapters.

4. To configure the third partition, select **Partition 3 Configuration** to open the Partition 3 Configuration page (Figure 5-19). If iSCSI Offload is present, the Partition 3 Configuration shows the following parameters:

   - ❑ **NIC Mode (Disabled)**
   - ❑ **iSCSI Offload Mode (Enabled)**
   - ❑ **iSCSI Offload MAC Address**[2]
   - ❑ **Virtual iSCSI Offload MAC Address[2]**
   - ❑ **PCI Device ID**
   - ❑ **PCI Address**



Main Configuration Page • NIC Partitioning Configuration • Partition 3 Configuration

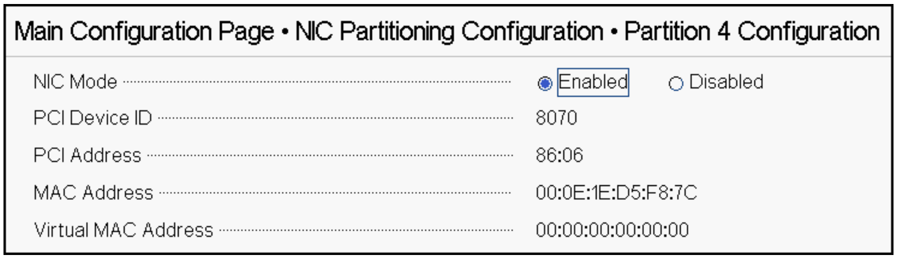
NIC Mode ○ Enabled ● Disabled
iSCSI Offload Mode ● Enabled ○ Disabled
iSCSI Offload MAC Address 00:0E:1E:D5:F8:7A
Virtual iSCSI Offload MAC Address 00:00:00:00:00:00
PCI Device ID 8070
PCI Address 86:04

Figure 5-19. Partition 3 Configuration: iSCSI Offload

5. To configure the remaining Ethernet partitions, including the previous (if not offload-enabled), open the page for a partition 2 or greater partition (see Figure 5-20).

   - ❑ **NIC Mode (Enabled or Disabled)**. When disabled, the partition is hidden such that it does not appear to the OS if fewer than the maximum quantity of partitions (or PCI PFs) are detected.
   - ❑ **PCI Device ID**
   - ❑ **PCI Address**
   - ❑ **MAC Address**
   - ❑ **Virtual MAC Address**

[2] This parameter is only present on the third partition of an NPAR mode port of iSCSI offload-capable adapters.



Main Configuration Page • NIC Partitioning Configuration • Partition 4 Configuration

NIC Mode ● Enabled ○ Disabled

PCI Device ID 8070

PCI Address 86:06

MAC Address 00:0E:1E:D5:F8:7C

Virtual MAC Address 00:00:00:00:00:00

Figure 5-20. Partition 4 Configuration

Partitioning for VMware ESXi 6.0 and ESXi 6.5

If the following conditions exist on a system running either VMware ESXi 6.0 or ESXi 6.5, you must uninstall and reinstall the drivers:

- The adapter is configured to enable NPAR with all NIC partitions.
- The adapter is in Single Function mode.
- The configuration is saved and the system is rebooted.
- Storage partitions are enabled (by converting one of the NIC partitions as storage) while drivers are already installed on the system.
- Partition 2 is changed to FCoE.
- The configuration is saved and the system is rebooted again.

Driver re-installation is required because the storage functions may keep the `vmnicX` enumeration rather than `vmhbaX`, as shown when you issue the following command on the system:

```
# esxcfg-scsidevs -a
vmnic4  qedf              link-up   fc.2000000e1ed6fa2a:2001000e1ed6fa2a
(0000:19:00.2) QLogic Corp. QLogic FastLinQ QL41xxx Series 10/25 GbE
Controller (FCoE)
vmhba0  lsi_mr3           link-n/a  sas.51866da071fa9100
(0000:18:00.0) Avago (LSI) PERC H330 Mini
vmnic10 qedf              link-up   fc.2000000e1ef249f8:2001000e1ef249f8
(0000:d8:00.2) QLogic Corp. QLogic FastLinQ QL41xxx Series 10/25 GbE
Controller (FCoE)
vmhba1  vmw_ahci          link-n/a  sata.vmhba1
(0000:00:11.5) Intel Corporation Lewisburg SSATA Controller [AHCI mode]
vmhba2  vmw_ahci          link-n/a  sata.vmhba2
(0000:00:17.0) Intel Corporation Lewisburg SATA Controller [AHCI mode]
vmhba32 qedil            online    iscsi.vmhba32                       QLogic
FastLinQ QL41xxx Series 10/25 GbE Controller (iSCSI)
vmhba33 qedil            online    iscsi.vmhba33                       QLogic
FastLinQ QL41xxx Series 10/25 GbE Controller (iSCSI)
```

In the preceding command output, notice that `vmnic4` and `vmnic10` are actually storage adapter ports. To prevent this behavior, you should enable storage functions at the same time that you configure the adapter for NPAR mode.

For example, assuming that the adapter is in Single Function mode by default, you should:

1. Enable NPAR mode.
2. Change Partition 2 to FCoE.
3. Save and reboot.

6 RoCE Configuration

This chapter describes RDMA over converged Ethernet (RoCE v1 and v2) configuration on the 41*xxx* Series Adapter, the Ethernet switch, and the Windows, Linux, or VMware host, including:

NOTE

Some RoCE features may not be fully enabled in the current release.

Supported Operating Systems and OFED

Table 6-1 shows the operating system support for RoCE v1, RoCE v2, iWARP, and OpenFabrics Enterprise Distribution (OFED). OFED is not supported on Windows or VMware ESXi.

Table 6-1. OS Support for RoCE v1, RoCE v2, iWARP, and OFED

Operating System	Inbox	OFED 3.18-3 GA	OFED-4.8-1 GA
Windows Server 2012 R2	No	N/A	N/A
Windows Server 2016	No	N/A	N/A
RHEL 6.8	RoCE v1, iWARP	RoCE v1, iWARP	No
RHEL 6.9	RoCE v1, iWARP	No	No

Table 6-1. OS Support for RoCE v1, RoCE v2, iWARP, and OFED (Continued)

Operating System	Inbox	OFED 3.18-3 GA	OFED-4.8-1 GA
RHEL 7.3	RoCE v1, RoCE v2, iWARP, iSER	No	RoCE v1, RoCE v2, iWARP
RHEL 7.4	RoCE v1, RoCE v2, iWARP, iSER	No	No
SLES 12 SP3	RoCE v1, RoCE v2, iWARP, iSER	No	No
CentOS 7.3	RoCE v1, RoCE v2, iWARP, iSER	No	RoCE v1, RoCE v2, iWARP
CentOS 7.4	RoCE v1, RoCE v2, iWARP, iSER	No	No
VMware ESXi 6.0 u3	No	N/A	N/A
VMware ESXi 6.5, 6.5U1	RoCE v1, RoCE v2	N/A	N/A
VMware ESXi 6.7	RoCE v1, RoCE v2	N/A	N/A

Planning for RoCE

As you prepare to implement RoCE, consider the following limitations:

- If you are using the inbox OFED, the operating system should be the same on the server and client systems. Some of the applications may work between different operating systems, but there is no guarantee. This is an OFED limitation.
- For OFED applications (most often perftest applications), server and client applications should use the same options and values. Problems can arise if the operating system and the perftest application have different versions. To confirm the perftest version, issue the following command:

```
# ib_send_bw --version
```

- Building libqedr in inbox OFED requires installing libibverbs-devel.
- Running user space applications in inbox OFED requires installing the InfiniBand® Support group, by yum groupinstall “InfiniBand Support” that contains libibcm, libibverbs, and more.

- OFED and RDMA applications that depend on libibverbs also require the QLogic RDMA user space library, libqedr. Install libqedr using the libqedr RPM or source packages.

  NOTE

  The user space library libqedr is part of the RDMA-Core. If your inbox OFED or out-of-box OFED has the RDMA-Core, you do not need to install libqedr. For example, SLES 12 SP3 and RHEL 7.4 and later do not require or support libqedr installation.

- RoCE supports only little endian.
- RoCE does not work over a VF in an SR-IOV environment.

Preparing the Adapter

Follow these steps to enable DCBX and specify the RoCE priority using the HII management application. For information about the HII application, see Chapter 5 Adapter Preboot Configuration.

To prepare the adapter:

1. In the Main Configuration Page, select **Data Center Bridging (DCB) Settings**, and then click **Finish**.
2. In the Data Center Bridging (DCB) Settings window, click the **DCBX Protocol** option. The 41*xxx* Series Adapter supports both CEE and IEEE protocols. This value should match the corresponding value on the DCB switch. In this example, select **CEE** or **Dynamic**.
3. In the **RoCE Priority** box, type a priority value. This value should match the corresponding value on the DCB switch. In this example, type 5. Typically, 0 is used for the default lossy traffic class, 3 is used for the FCoE traffic class, and 4 is used for lossless iSCSI-TLV over DCB traffic class.
4. Click **Back**.
5. When prompted, click **Yes** to save the changes. Changes will take effect after a system reset.

   For Windows, you can configure DCBX using the HII or QoS method. The configuration shown in this section is through HII. For QoS, refer to “Configuring QoS for RoCE” on page 235“.

Preparing the Ethernet Switch

This section describes how to configure a Cisco® Nexus® 6000 Ethernet Switch and a Dell Z9100 Ethernet Switch for RoCE.

- Configuring the Cisco Nexus 6000 Ethernet Switch
- Configuring the Dell Z9100 Ethernet Switch

Configuring the Cisco Nexus 6000 Ethernet Switch

Steps for configuring the Cisco Nexus 6000 Ethernet Switch for RoCE include configuring class maps, configuring policy maps, applying the policy, and assigning a VLAN ID to the switch port.

To configure the Cisco switch:

1. Open a config terminal session as follows:

```
Switch# config terminal
switch(config)#
```

2. Configure the quality of service (QoS) class map and set the RoCE priority (`cos`) to match the adapter (5) as follows:

```
switch(config)# class-map type qos class-roce
switch(config)# match cos 5
```

3. Configure queuing class maps as follows:

```
switch(config)# class-map type queuing class-roce
switch(config)# match qos-group 3
```

4. Configure network QoS class maps as follows:

```
switch(config)# class-map type network-qos class-roce
switch(config)# match qos-group 3
```

5. Configure QoS policy maps as follows:

```
switch(config)# policy-map type qos roce
switch(config)# class type qos class-roce
switch(config)# set qos-group 3
```

6. Configure queuing policy maps to assign network bandwidth. In this example, use a value of 50 percent:

```
switch(config)# policy-map type queuing roce
switch(config)# class type queuing class-roce
switch(config)# bandwidth percent 50
```

7. Configure network QoS policy maps to set priority flow control for no-drop traffic class as follows:

```
switch(config)# policy-map type network-qos roce
switch(config)# class type network-qos class-roce
switch(config)# pause no-drop
```

8. Apply the new policy at the system level as follows

```
switch(config)# system qos
switch(config)# service-policy type qos input roce
switch(config)# service-policy type queuing output roce
switch(config)# service-policy type queuing input roce
switch(config)# service-policy type network-qos roce
```

9. Assign a VLAN ID to the switch port to match the VLAN ID assigned to the adapter (5).

```
switch(config)# interface ethernet x/x
switch(config)# switchport mode trunk
switch(config)# switchport trunk allowed vlan 1,5
```

Configuring the Dell Z9100 Ethernet Switch

To configure the Dell Z9100 Ethernet Switch for RoCE, see the procedure in Appendix C Dell Z9100 Switch Configuration.

Configuring RoCE on the Adapter for Windows Server

Configuring RoCE on the adapter for Windows Server host comprises enabling RoCE on the adapter and verifying the Network Direct MTU size.

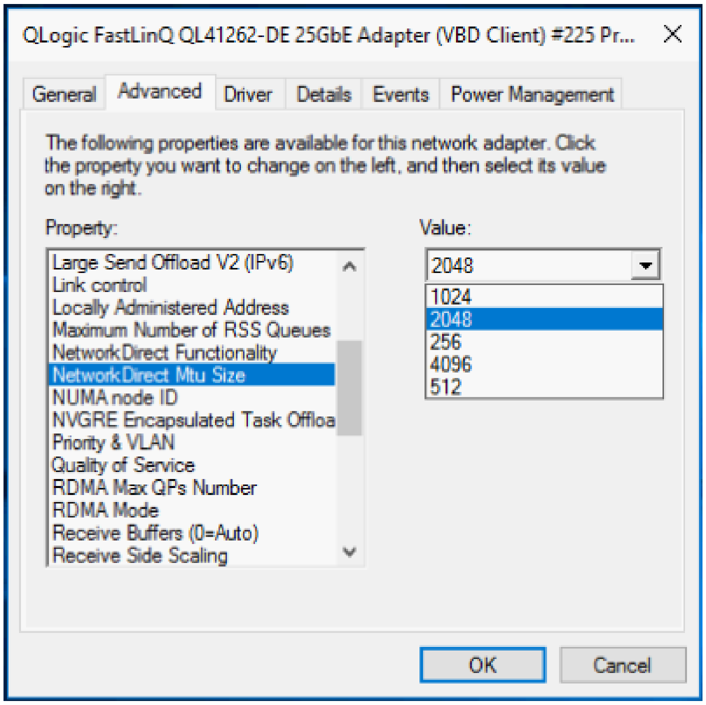
To configure RoCE on a Windows Server host:

1. Enable RoCE on the adapter.
   a. Open the Windows Device Manager, and then open the 41*xxx* Series Adapters NDIS Miniport Properties.
   b. On the QLogic FastLinQ Adapter Properties, click the **Advanced** tab.
   c. On the Advanced page, configure the properties listed in Table 6-2 by selecting each item under **Property** and choosing an appropriate **Value** for that item. Then click **OK**.

Table 6-2. Advanced Properties for RoCE

Property	Value or Description
`Network Direct Functionality`	**Enabled**
`Network Direct Mtu Size`	The network direct MTU size must be less than the jumbo packet size.
`Quality of Service`	For RoCE v1/v2, always select **Enabled** to allow Windows DCB-QoS service to control and monitor DCB. For more information, see "Configuring QoS by Disabling DCBX on the Adapter" on page 235 and "Configuring QoS by Enabling DCBX on the Adapter" on page 239.
`RDMA Mode`	**RoCE v1** or **RoCE v2**. The **iWARP** value applies only when configuring ports for iWARP as described in Chapter 7 iWARP Configuration.
`VLAN ID`	Assign any VLAN ID to the interface. The value must be the same as is assigned on the switch.

Figure 6-1 shows an example of configuring a property value.



QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

Large Send Offload V2 (IPv6)
Link control
Locally Administered Address
Maximum Number of RSS Queues
NetworkDirect Functionality
NetworkDirect Mtu Size
NUMA node ID
NVGRE Encapsulated Task Offloa
Priority & VLAN
Quality of Service
RDMA Max QPs Number
RDMA Mode
Receive Buffers (0=Auto)
Receive Side Scaling

Value:

2048

1024
2048
256
4096
512

Figure 6-1. Configuring RoCE Properties

2. Using Windows PowerShell, verify that RDMA is enabled on the adapter. The `Get-NetAdapterRdma` command lists the adapters that support RDMA—both ports are enabled.

NOTE

If you are configuring RoCE over Hyper-V, do not assign a VLAN ID to the physical interface.

```
PS C:\Users\Administrator> Get-NetAdapterRdma
Name                InterfaceDescription         Enabled
-----               --------------------         -------
SLOT 4 3 Port 1     QLogic FastLinQ QL41262...   True
SLOT 4 3 Port 2     QLogic FastLinQ QL41262...   True
```

3. Using Windows PowerShell, verify that `NetworkDirect` is enabled on the host operating system. The `Get-NetOffloadGlobalSetting` command shows `NetworkDirect` is enabled.

```
PS C:\Users\Administrators> Get-NetOffloadGlobalSetting
ReceiveSideScaling           : Enabled
```

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

4. Connect a server message block (SMB) drive, run RoCE traffic, and verify the results.

   To set up and connect to an SMB drive, view the information available online from Microsoft:

   https://technet.microsoft.com/en-us/library/hh831795(v=ws.11).aspx

5. By default, Microsoft's SMB Direct establishes two RDMA connections per port, which provides good performance, including line rate at a higher block size (for example, 64KB). To optimize performance, you can change the quantity of RDMA connections per RDMA interface to four (or greater).

   To increase the quantity of RDMA connections to four (or more), issue the following command in Windows PowerShell:

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

Viewing RDMA Counters

The following procedure also applies to iWARP.

To view RDMA counters for RoCE:

1. Launch Performance Monitor.
2. Open the Add Counters dialog box. Figure 6-2 shows an example.

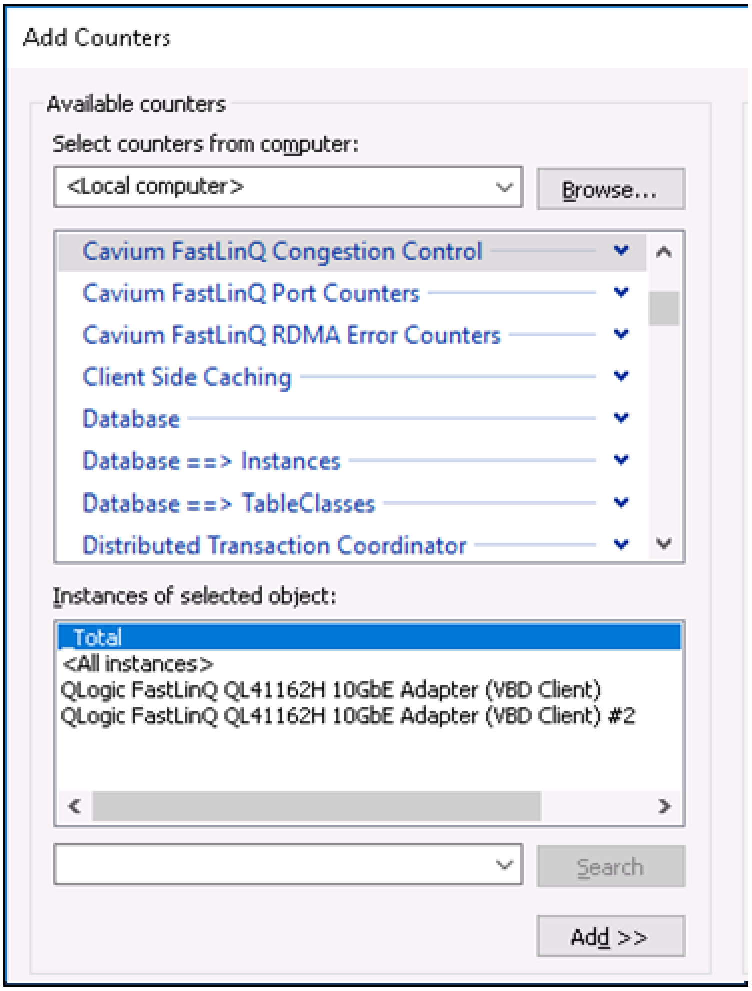

Add Counters
Available counters
Select counters from computer:
<Local computer>
Browse...
Cavium FastLinQ Congestion Control
Cavium FastLinQ Port Counters
Cavium FastLinQ RDMA Error Counters
Client Side Caching
Database
Database ==> Instances
Database ==> TableClasses
Distributed Transaction Coordinator
Instances of selected object:
Total
<All instances>
QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client)
QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client) #2
Search
Add >>

Figure 6-2. Add Counters Dialog Box

NOTE

If Cavium RDMA counters are not listed in the Performance Monitor Add Counters dialog box, manually add them by issuing the following command from the driver location:

```
Lodctr /M:qend.man
```

3. Select one of the following counter types:

   ❑ **Cavium FastLinQ Congestion Control**:

   - Increment when there is congestion in the network and ECN is enabled on the switch.
   - Describe RoCE v2 ECN Marked Packets and Congestion Notification Packets (CNPs) sent and received successfully.
   - Apply only to RoCE v2.

   ❑ **Cavium FastLinQ Port Counters**:

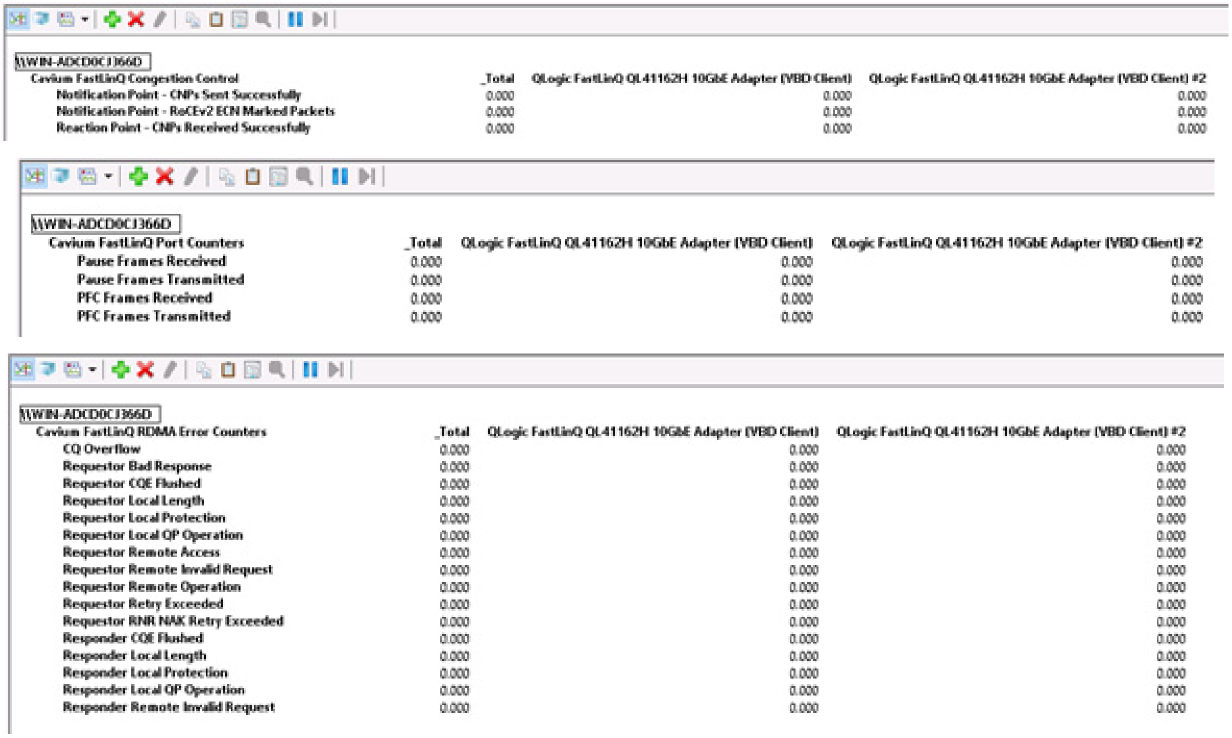
   - Increment when there is congestion in the network.
   - Pause counters increment when flow control or global pause is configured and there is a congestion in the network.
   - PFC counters increment when priority flow control is configured and there is a congestion in the network.

   ❑ **Cavium FastLinQ RDMA Error Counters**:

   - Increment if any error occurs in transport operations.
   - For details, see Table 6-3.

4. Under **Instances of selected object**, select **Total**, and then click **Add**.

Figure 6-3 shows three examples of the counter monitoring output.



```
\\WIN-ADCD0CJ366D
  Cavium FastLinQ Congestion Control               _Total  QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client)  QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client) #2
      Notification Point - CNPs Sent Successfully   0.000                                                0.000                                                     0.000
      Notification Point - RoCEv2 ECN Marked Packets 0.000                                               0.000                                                     0.000
      Reaction Point - CNPs Received Successfully   0.000                                                0.000                                                     0.000
```

```
\\WIN-ADCD0CJ366D
  Cavium FastLinQ Port Counters          _Total  QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client)  QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client) #2
      Pause Frames Received              0.000                                                0.000                                                     0.000
      Pause Frames Transmitted           0.000                                                0.000                                                     0.000
      PFC Frames Received                0.000                                                0.000                                                     0.000
      PFC Frames Transmitted             0.000                                                0.000                                                     0.000
```

```
\\WIN-ADCD0CJ366D
  Cavium FastLinQ RDMA Error Counters            _Total  QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client)  QLogic FastLinQ QL41162H 10GbE Adapter (VBD Client) #2
      CQ Overflow                                0.000                                                0.000                                                     0.000
      Requestor Bad Response                     0.000                                                0.000                                                     0.000
      Requestor CQE Flushed                      0.000                                                0.000                                                     0.000
      Requestor Local Length                     0.000                                                0.000                                                     0.000
      Requestor Local Protection                 0.000                                                0.000                                                     0.000
      Requestor Local QP Operation               0.000                                                0.000                                                     0.000
      Requestor Remote Access                    0.000                                                0.000                                                     0.000
      Requestor Remote Invalid Request           0.000                                                0.000                                                     0.000
      Requestor Remote Operation                 0.000                                                0.000                                                     0.000
      Requestor Retry Exceeded                   0.000                                                0.000                                                     0.000
      Requestor RNR NAK Retry Exceeded           0.000                                                0.000                                                     0.000
      Responder CQE Flushed                      0.000                                                0.000                                                     0.000
      Responder Local Length                     0.000                                                0.000                                                     0.000
      Responder Local Protection                 0.000                                                0.000                                                     0.000
      Responder Local QP Operation               0.000                                                0.000                                                     0.000
      Responder Remote Invalid Request           0.000                                                0.000                                                     0.000
```

Figure 6-3. Performance Monitor: Cavium FastLinQ Counters

Table 6-3 provides details about error counters.

Table 6-3. Cavium FastLinQ RDMA Error Counters

RDMA Error Counter	Description	Applies to RoCE?	Applies to iWARP?	Troubleshooting
`CQ overflow`	A completion queue on which an RDMA work request is posted. This counter specifies the quantity of instances where there was a completion for a work request on send or receive queue, but no space on the associated completion queue.	Yes	Yes	Indicates a software design issue causing an insufficient completion queue size.
`Requestor Bad response`	A malformed response was returned by the responder.	Yes	Yes	—

Table 6-3. Cavium FastLinQ RDMA Error Counters (Continued)

RDMA Error Counter	Description	Applies to RoCE?	Applies to iWARP?	Troubleshooting
`Requestor CQEs flushed with error`	Posted work requests may be flushed by sending completions with a flush status to the CQ (without completing the actual execution of work request) if the QP moves to an error state for any reason and pending work requests exist. If a work request completed with error status, then all other pending work requests for that QP are flushed.	Yes	Yes	Occurs when the RDMA connection is down.
`Requestor Local length`	The RDMA Read response message contained too much or too little payload data.	Yes	Yes	Usually indicates an issue with the host software components.
`Requestor local protection`	The locally posted work request's data segment does not reference a memory region that is valid for the requested operation.	Yes	Yes	Usually indicates an issue with the host software components.
`Requestor local QP operation`	An internal QP consistency error was detected while processing this work request.	Yes	Yes	—
`Requestor Remote access`	A protection error occurred on a remote data buffer to be read by an RDMA Read, written by an RDMA Write, or accessed by an atomic operation.	Yes	Yes	—
`Requestor Remote Invalid request`	The remote side received an invalid message on the channel. The invalid request may have been a Send message or an RDMA request.	Yes	Yes	Possible causes include the operation is not supported by this receive queue, insufficient buffering to receive a new RDMA or atomic operation request, or the length specified in an RDMA request is greater than 231 bytes.

Table 6-3. Cavium FastLinQ RDMA Error Counters (Continued)

RDMA Error Counter	Description	Applies to RoCE?	Applies to iWARP?	Troubleshooting
`Requestor remote operation`	Remote side could not complete the operation requested due to a local issue.	Yes	Yes	A software issue at the remote side (for example, one that caused a QP error or a malformed WQE on the RQ) prevented operation completion.
`Requestor retry exceeded`	Transport retries have exceeded the maximum limit.	Yes	Yes	The remote peer may have stopped responding, or a network issue is preventing messages acknowledgment.
`Requestor RNR Retries exceeded`	Retry due to RNR NAK received have been tried the maximum number of times without success.	Yes	No	The remote peer may have stopped responding, or a network issue is preventing messages acknowledgment.
`Responder CQE flushed`	Posted work requests (receive buffers on RQ) may be flushed by sending completions with a flush status to the CQ if the QP moves to an error state for any reason, and pending receive buffers exist on the RQ. If a work request completed with an error status, then all other pending work requests for that QP are flushed.	Yes	Yes	—
`Responder local length`	Invalid length in inbound messages.	Yes	Yes	Misbehaving remote peer. For example, the inbound send messages have lengths greater than the receive buffer size.
`Responder local protection`	The locally posted work request’s data segment does not reference a memory region that is valid for the requested operation.	Yes	Yes	Indicates a software issue with memory management.

Table 6-3. Cavium FastLinQ RDMA Error Counters (Continued)

RDMA Error Counter	Description	Applies to RoCE?	Applies to iWARP?	Troubleshooting
`Responder Local QP Operation error`	An internal QP consistency error was detected while processing this work request.	Yes	Yes	Indicates a software issue.
`Responder remote invalid request`	The responder detected an invalid inbound message on the channel.	Yes	Yes	Indicates possible misbehavior by a remote peer. Possible causes include: the operation is not supported by this receive queue, insufficient buffering to receive a new RDMA request, or the length specified in an RDMA request is greater than 2^{31} bytes.

Configuring RoCE on the Adapter for Linux

This section describes the RoCE configuration procedure for RHEL and SLES. It also describes how to verify the RoCE configuration and provides some guidance about using group IDs (GIDs) with VLAN interfaces.

- RoCE Configuration for RHEL
- RoCE Configuration for SLES
- Verifying the RoCE Configuration on Linux
- VLAN Interfaces and GID Index Values
- RoCE v2 Configuration for Linux

RoCE Configuration for RHEL

To configure RoCE on the adapter, the Open Fabrics Enterprise Distribution (OFED) must be installed and configured on the RHEL host.

To prepare inbox OFED for RHEL:

1. While installing or upgrading the operating system, select the InfiniBand and OFED support packages.
2. Install the following RPMs from the RHEL ISO image. These packages are required for earlier versions, including RHEL 7.3 and 7.2:

   `libibverbs-devel-x.x.x.x86_64.rpm`
   (required for libqedr library)

   `perftest-x.x.x.x86_64.rpm`
   (required for InfiniBand bandwidth and latency applications)

   or, using Yum, install the inbox OFED:

   ```
   yum groupinstall "Infiniband Support"
   yum install perftest
   yum install tcl tcl-devel tk zlib-devel libibverbs
   libibverbs-devel
   ```

   NOTE

   During installation, if you already selected the previously mentioned packages, you need not reinstall them. The inbox OFED and support packages may vary depending on the operating system version.

3. Install the new Linux drivers as described in "Installing the Linux Drivers with RDMA" on page 14.

RoCE Configuration for SLES

To configure RoCE on the adapter for an SLES host, OFED must be installed and configured on the SLES host.

To install inbox OFED for SLES:

1. While installing or upgrading the operating system, select the InfiniBand support packages.
2. Install the following RPMs from the corresponding SLES SDK kit image. These packages are required for earlier versions, including SLES 12 SP1/SP2:

   `libibverbs-devel-x.x.x.x86_64.rpm`
   (required for libqedr installation)

`perftest-x.x.x.x86_64.rpm`
(required for bandwidth and latency applications)

3. Install the Linux drivers, as described in “Installing the Linux Drivers with RDMA” on page 14.

Verifying the RoCE Configuration on Linux

After installing OFED, installing the Linux driver, and loading the RoCE drivers, verify that the RoCE devices were detected on all Linux operating systems.

To verify RoCE configuration on Linux:

1. Stop firewall tables using `service/systemctl` commands.
2. For RHEL only: If the RDMA service is installed (`yum install rdma`), verify that the RDMA service has started.

NOTE

For RHEL 6.*x* and SLES 11 SP4, you must start RDMA service after reboot. For RHEL 7.*x* and SLES 12 SP*x* and later, RDMA service starts itself after reboot.

On RHEL or CentOS: Use the `service rdma` status command to start service:

- If RDMA has not started, issue the following command:

```
# service rdma start
```

- If RDMA does not start, issue either of the following alternative commands:

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

or

```
# systemctl start rdma.service
```

3. Verify that the RoCE devices were detected by examining the dmesg logs:

```
# dmesg|grep qedr
  [87910.988411] qedr: discovered and registered 2 RoCE funcs
```

4. Verify that all of the modules have been loaded. For example:

```
# lsmod|grep qedr
  qedr                  89871   0
  qede                  96670   1 qedr
  qed                   2075255   2 qede,qedr
```

```
ib_core                  88311  16  qedr, rdma_cm, ib_cm,
                                    ib_sa,iw_cm,xprtrdma,ib_mad,ib_srp,
                                    ib_ucm,ib_iser,ib_srpt,ib_umad,
                                    ib_uverbs,rdma_ucm,ib_ipoib,ib_isert
```

5. Configure the IP address and enable the port using a configuration method such as ifconfig. For example:

   ```
   # ifconfig ethX 192.168.10.10/24 up
   ```

6. Issue the `ibv_devinfo` command. For each PCI function, you should see a separate `hca_id`, as shown in the following example:

   ```
   root@captain:~# ibv_devinfo
   hca_id: qedr0
           transport:                      InfiniBand (0)
           fw_ver:                         8.3.9.0
           node_guid:                      020e:1eff:fe50:c7c0
           sys_image_guid:                 020e:1eff:fe50:c7c0
           vendor_id:                      0x1077
           vendor_part_id:                 5684
           hw_ver:                         0x0
           phys_port_cnt:                  1
                   port:   1
                          state:                  PORT_ACTIVE (1)
                           max_mtu:                4096 (5)
                           active_mtu:             1024 (3)
                           sm_lid:                 0
                           port_lid:               0
                           port_lmc:               0x00
                           link_layer:             Ethernet
   ```

7. Verify the L2 and RoCE connectivity between all servers: one server acts as a server, another acts as a client.

   - Verify the L2 connection using a simple `ping` command.
   - Verify the RoCE connection by performing an RDMA ping on the server or client:

     On the server, issue the following command:

     ```
     ibv_rc_pingpong -d <ib-dev> -g 0
     ```

     On the client, issue the following command:

     ```
     ibv_rc_pingpong -d <ib-dev> -g 0 <server L2 IP address>
     ```

The following are examples of successful ping pong tests on the server and the client.

Server Ping:

```
root@captain:~# ibv_rc_pingpong -d qedr0 -g 0
local address:  LID 0x0000, QPN 0xff0000, PSN 0xb3e07e, GID
fe80::20e:1eff:fe50:c7c0
remote address: LID 0x0000, QPN 0xff0000, PSN 0x934d28, GID
fe80::20e:1eff:fe50:c570
8192000 bytes in 0.05 seconds = 1436.97 Mbit/sec
1000 iters in 0.05 seconds = 45.61 usec/iter
```

Client Ping:

```
root@lambodar:~# ibv_rc_pingpong -d qedr0 -g 0 192.168.10.165
local address:  LID 0x0000, QPN 0xff0000, PSN 0x934d28, GID
fe80::20e:1eff:fe50:c570
remote address: LID 0x0000, QPN 0xff0000, PSN 0xb3e07e, GID
fe80::20e:1eff:fe50:c7c0
8192000 bytes in 0.02 seconds = 4211.28 Mbit/sec
1000 iters in 0.02 seconds = 15.56 usec/iter
```

- To display RoCE statistics, issue the following commands, where ***X*** is the device number:

  ```
  > mount -t debugfs nodev /sys/kernel/debug
  > cat /sys/kernel/debug/qedr/qedrX/stats
  ```

VLAN Interfaces and GID Index Values

If you are using VLAN interfaces on both the server and the client, you must also configure the same VLAN ID on the switch. If you are running traffic through a switch, the InfiniBand applications must use the correct GID value, which is based on the VLAN ID and VLAN IP address.

Based on the following results, the GID value (-x 4 / -x 5) should be used for any perftest applications.

```
# ibv_devinfo -d qedr0 -v|grep GID
     GID[  0]:  fe80:0000:0000:0000:020e:1eff:fe50:c5b0
     GID[  1]:  0000:0000:0000:0000:0000:ffff:c0a8:0103
     GID[  2]:  2001:0db1:0000:0000:020e:1eff:fe50:c5b0
     GID[  3]:  2001:0db2:0000:0000:020e:1eff:fe50:c5b0
     GID[  4]:  0000:0000:0000:0000:0000:ffff:c0a8:0b03  IP address for VLAN interface
     GID[  5]:  fe80:0000:0000:0000:020e:1e00:0350:c5b0     VLAN ID 3
```

NOTE

The default GID value is zero (0) for back-to-back or pause settings. For server/switch configurations, you must identify the proper GID value. If you are using a switch, refer to the corresponding switch configuration documents for the correct settings.

RoCE v2 Configuration for Linux

To verify RoCE v2 functionality, you must use RoCE v2 supported kernels.

To configure RoCE v2 for Linux:

1. Ensure that you are using one of the following supported kernels:

   - □ SLES 12 SP2 GA
   - □ RHEL 7.3 GA

2. Configure RoCE v2 as follows:

   a. Identify the GID index for RoCE v2.

   b. Configure the routing address for the server and client.

   c. Enable L3 routing on the switch.

NOTE

You can configure RoCE v1 and RoCE v2 by using RoCE v2-supported kernels. These kernels allow you to run RoCE traffic over the same subnet, as well as over different subnets such as RoCE v2 and any routable environment. Only a few settings are required for RoCE v2, and all other switch and adapter settings are common for RoCE v1 and RoCE v2.

Identifying RoCE v2 GID Index or Address

To find RoCE v1- and RoCE v2-specific GIDs, use either sys or class parameters, or run RoCE scripts from the 41*xxx* FastLinQ source package. To check the default **RoCE GID Index** and address, issue the `ibv_devinfo` command and compare it with the sys or class parameters. For example:

```
#ibv_devinfo -d qedr0 -v|grep GID
           GID[  0]:         fe80:0000:0000:0000:020e:1eff:fec4:1b20
           GID[  1]:         fe80:0000:0000:0000:020e:1eff:fec4:1b20
           GID[  2]:         0000:0000:0000:0000:0000:ffff:1e01:010a
           GID[  3]:         0000:0000:0000:0000:0000:ffff:1e01:010a
           GID[  4]:         3ffe:ffff:0000:0f21:0000:0000:0000:0004
           GID[  5]:         3ffe:ffff:0000:0f21:0000:0000:0000:0004
```

```
GID[   6]:          0000:0000:0000:0000:0000:ffff:c0a8:6403
GID[   7]:          0000:0000:0000:0000:0000:ffff:c0a8:6403
```

Verifying RoCE v1 or RoCE v2 GID Index and Address from sys and class Parameters

Use one of the following options to verify the RoCE v1 or RoCE v2 GID Index and address from the sys and class parameters:

- **Option 1:**

```
# cat /sys/class/infiniband/qedr0/ports/1/gid_attrs/types/0
IB/RoCE v1
# cat /sys/class/infiniband/qedr0/ports/1/gid_attrs/types/1
RoCE v2

# cat /sys/class/infiniband/qedr0/ports/1/gids/0
fe80:0000:0000:0000:020e:1eff:fec4:1b20
# cat /sys/class/infiniband/qedr0/ports/1/gids/1
fe80:0000:0000:0000:020e:1eff:fec4:1b20
```

- **Option 2:**

  Use the scripts from the FastLinQ source package.

```
#/../fastlinq-8.x.x.x/add-ons/roce/show_gids.sh
DEV    PORT  INDEX  GID                                      IPv4           VER     DEV
---    ----  -----  ---                                      ------------   ---     ---
qedr0   1     0     fe80:0000:0000:0000:020e:1eff:fec4:1b20                 v1      p4p1
qedr0   1     1     fe80:0000:0000:0000:020e:1eff:fec4:1b20                 v2      p4p1
qedr0   1     2     0000:0000:0000:0000:0000:ffff:1e01:010a  30.1.1.10      v1      p4p1
qedr0   1     3     0000:0000:0000:0000:0000:ffff:1e01:010a  30.1.1.10      v2      p4p1
qedr0   1     4     3ffe:ffff:0000:0f21:0000:0000:0000:0004                 v1      p4p1
qedr0   1     5     3ffe:ffff:0000:0f21:0000:0000:0000:0004                 v2      p4p1
qedr0   1     6     0000:0000:0000:0000:0000:ffff:c0a8:6403  192.168.100.3  v1      p4p1.100
qedr0   1     7     0000:0000:0000:0000:0000:ffff:c0a8:6403  192.168.100.3  v2      p4p1.100
qedr1   1     0     fe80:0000:0000:0000:020e:1eff:fec4:1b21                 v1      p4p2
qedr1   1     1     fe80:0000:0000:0000:020e:1eff:fec4:1b21                 v2      p4p2
```

NOTE

You must specify the GID index values for RoCE v1- or RoCE v2-based on server or switch configuration (Pause/PFC). Use the GID index for the link local IPv6 address, IPv4 address, or IPv6 address. To use VLAN tagged frames for RoCE traffic, you must specify GID index values that are derived from the VLAN IPv4 or IPv6 address.

Verifying RoCE v1 or RoCE v2 Function Through perftest Applications

This section shows how to verify RoCE v1 or RoCE v2 function through perftest applications. In this example, the following server IP and client IP are used:

- Server IP: 192.168.100.3
- Client IP: 192.168.100.4

Verifying RoCE v1

Run over the same subnet and use the RoCE v1 GID index.

```
Server# ib_send_bw -d qedr0 -F -x 0
Client# ib_send_bw -d qedr0 -F -x 0  192.168.100.3
```

Verifying RoCE v2

Run over the same subnet and use the RoCE v2 GID index.

```
Server# ib_send_bw -d qedr0 -F -x 1
Client# ib_send_bw -d qedr0 -F -x 1  192.168.100.3
```

NOTE

If you are running through a switch PFC configuration, use VLAN GIDs for RoCE v1 or v2 through the same subnet.

Verifying RoCE v2 Through Different Subnets

NOTE

You must first configure the route settings for the switch and servers. On the adapter, set the RoCE priority and DCBX mode using either the HII, UEFI user interface, or one of the Cavium management utilities.

To verify RoCE v2 through different subnets:

1. Set the route configuration for the server and client using the DCBX-PFC configuration.

   - ❑ **System Settings:**

     Server VLAN IP : 192.168.100.3 and **Gateway :**192.168.100.1

     Client VLAN IP : 192.168.101.3 and **Gateway :**192.168.101.1

   - ❑ **Server Configuration:**

   ```
   #/sbin/ip link add link p4p1 name p4p1.100 type vlan id 100
   #ifconfig p4p1.100 192.168.100.3/24 up
   #ip route add 192.168.101.0/24 via 192.168.100.1 dev p4p1.100
   ```

   - ❑ **Client Configuration:**

   ```
   #/sbin/ip link add link p4p1 name p4p1.101 type vlan id 101
   #ifconfig p4p1.101 192.168.101.3/24 up
   #ip route add 192.168.100.0/24 via 192.168.101.1 dev p4p1.101
   ```

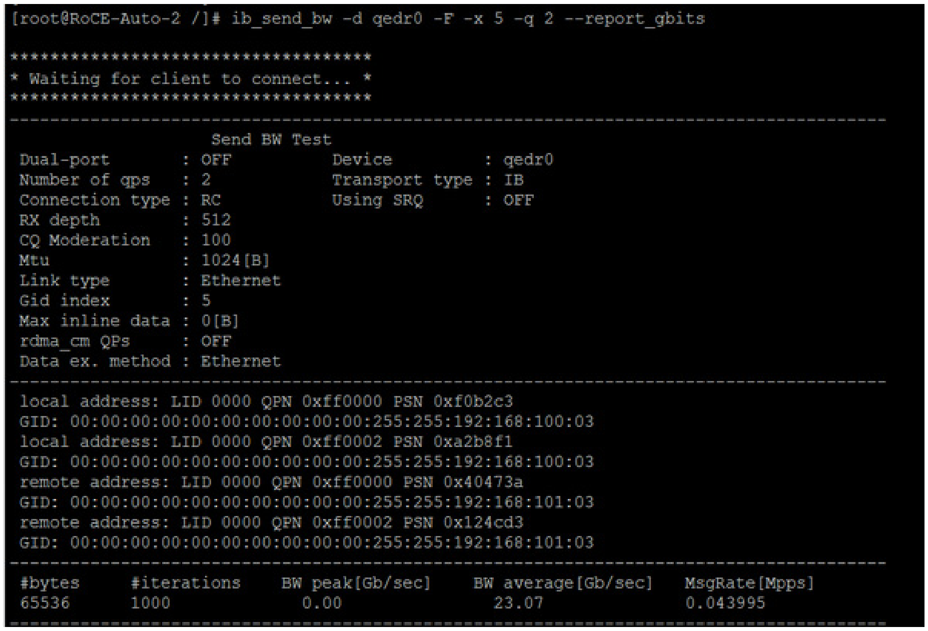
2. Set the switch settings using the following procedure.

   - ❑ Use any flow control method (Pause, DCBX-CEE, or DCBX-IEEE), and enable IP routing for RoCE v2. See “Preparing the Ethernet Switch” on page 66 for RoCE v2 configuration, or refer to the vendor switch documents.

   - ❑ If you are using PFC configuration and L3 routing, run RoCE v2 traffic over the VLAN using a different subnet, and use the RoCE v2 VLAN GID index.

     ```
     Server# ib_send_bw -d qedr0 -F -x 5
     Client# ib_send_bw -d qedr0 -F -x 5 192.168.100.3
     ```

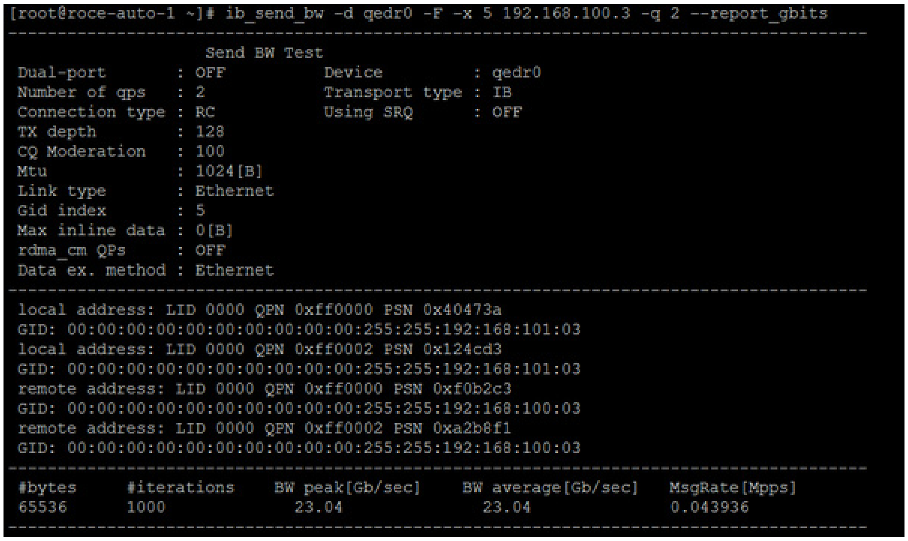
Server Switch Settings:



```
[root@RoCE-Auto-2 /]# ib_send_bw -d qedr0 -F -x 5 -q 2 --report_gbits

************************************
* Waiting for client to connect... *
************************************
---------------------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : qedr0
 Number of qps   : 2            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 RX depth        : 512
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 Gid index       : 5
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0xff0000 PSN 0xf0b2c3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
 local address: LID 0000 QPN 0xff0002 PSN 0xa2b8f1
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
 remote address: LID 0000 QPN 0xff0000 PSN 0x40473a
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
 remote address: LID 0000 QPN 0xff0002 PSN 0x124cd3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 65536      1000             0.00               23.07              0.043995
---------------------------------------------------------------------------------------
```

Figure 6-4. Switch Settings, Server

Client Switch Settings:



```
[root@roce-auto-1 ~]# ib_send_bw -d qedr0 -F -x 5 192.168.100.3 -q 2 --report_gbits
---------------------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : qedr0
 Number of qps   : 2            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 Gid index       : 5
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0xff0000 PSN 0x40473a
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
 local address: LID 0000 QPN 0xff0002 PSN 0x124cd3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
 remote address: LID 0000 QPN 0xff0000 PSN 0xf0b2c3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
 remote address: LID 0000 QPN 0xff0002 PSN 0xa2b8f1
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 65536      1000             23.04              23.04              0.043936
---------------------------------------------------------------------------------------
```

Figure 6-5. Switch Settings, Client

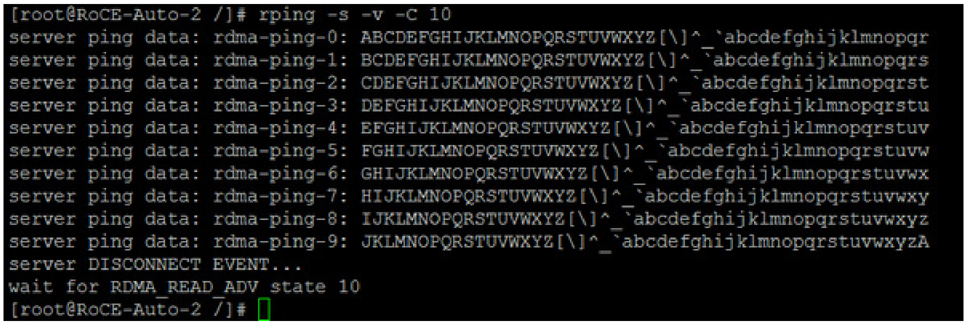
Configuring RoCE v1 or RoCE v2 Settings for RDMA_CM Applications

To configure RoCE, use the following scripts from the FastLinQ source package:

```
# ./show_rdma_cm_roce_ver.sh
qedr0 is configured to IB/RoCE v1
qedr1 is configured to IB/RoCE v1
```

```
# ./config_rdma_cm_roce_ver.sh v2
configured rdma_cm for qedr0 to RoCE v2
configured rdma_cm for qedr1 to RoCE v2
```

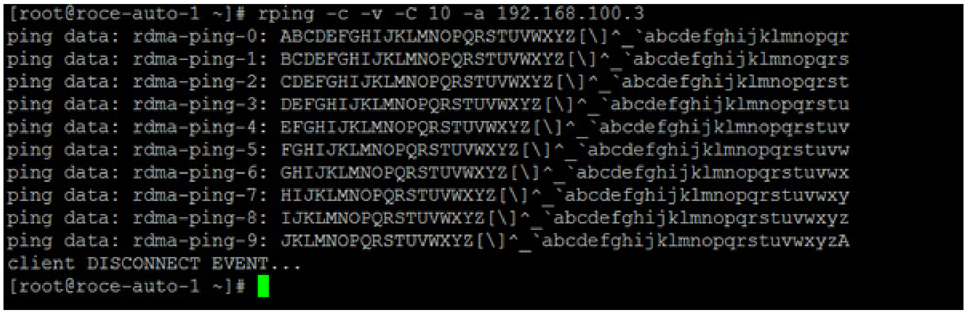
Server Settings:



```
[root@RoCE-Auto-2 /]# rping -s -v -C 10
server ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
server ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
server ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
server ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
server ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
server ping data: rdma-ping-5: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvw
server ping data: rdma-ping-6: GHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwx
server ping data: rdma-ping-7: HIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
server ping data: rdma-ping-8: IJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
server ping data: rdma-ping-9: JKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzA
server DISCONNECT EVENT...
wait for RDMA_READ_ADV state 10
[root@RoCE-Auto-2 /]#
```

Figure 6-6. Configuring RDMA_CM Applications: Server

Client Settings:



```
[root@roce-auto-1 ~]# rping -c -v -C 10 -a 192.168.100.3
ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
ping data: rdma-ping-5: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvw
ping data: rdma-ping-6: GHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwx
ping data: rdma-ping-7: HIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
ping data: rdma-ping-8: IJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
ping data: rdma-ping-9: JKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzA
client DISCONNECT EVENT...
[root@roce-auto-1 ~]#
```

Figure 6-7. Configuring RDMA_CM Applications: Client

Configuring RoCE on the Adapter for VMware ESX

This section provides the following procedures and information for RoCE configuration:

- Configuring RDMA Interfaces
- Configuring MTU
- RoCE Mode and Statistics
- Configuring a Paravirtual RDMA Device (PVRDMA)

Configuring RDMA Interfaces

To configure the RDMA interfaces:

1. Install both QLogic NIC and RoCE drivers.
2. Using the module parameter, enable the RoCE function from the NIC driver by issuing the following command:

   ```
   esxcfg-module -s 'enable_roce=1' qedentv
   ```

   To apply the change, reload the NIC driver or reboot the system.
3. To view a list of the NIC interfaces, issue the `esxcfg-nics -l` command. For example:

```
esxcfg-nics -l
Name    PCI          Driver     Link Speed     Duplex MAC Address       MTU    Description
Vmnic0  0000:01:00.2 qedentv    Up   25000Mbps  Full   a4:5d:36:2b:6c:92 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 1/10/25 GbE Ethernet Adapter
Vmnic1  0000:01:00.3 qedentv    Up   25000Mbps  Full   a4:5d:36:2b:6c:93 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 1/10/25 GbE Ethernet Adapter
```

4. To view a list of the RDMA devices, issue the `esxcli rdma device list` command. For example:

```
esxcli rdma device list
Name     Driver   State    MTU  Speed    Paired Uplink  Description
-------  -------  ------  ----  -------  -------------  -------------------------------
vmrdma0  qedrntv  Active  1024  25 Gbps  vmnic0     QLogic FastLinQ QL45xxx RDMA Interface
vmrdma1  qedrntv  Active  1024  25 Gbps  vmnic1     QLogic FastLinQ QL45xxx RDMA Interface
```

5. To create a new virtual switch, issue the following command:

   ```
   esxcli network vswitch standard add -v <new vswitch name>
   ```

   For example:

   ```
   # esxcli network vswitch standard add -v roce_vs
   ```

   This creates a new virtual switch named *roce_vs*.

6. To associate the QLogic NIC port to the vSwitch, issue the following command:

   ```
   # esxcli network vswitch standard uplink add -u <uplink device> -v <roce vswitch>
   ```

   For example:

   ```
   # esxcli network vswitch standard uplink add -u vmnic0 -v roce_vs
   ```

7. To create a new port group on this vSwitch, issue the following command:

   ```
   # esxcli network vswitch standard portgroup add -p roce_pg -v roce_vs
   ```

   For example:

   ```
   # esxcli network vswitch standard portgroup add -p roce_pg -v roce_vs
   ```

8. To create a vmknic interface on this port group and configure the IP, issue the following command:

   ```
   # esxcfg-vmknic -a -i <IP address> -n <subnet mask> <roce port group name>
   ```

   For example:

   ```
   # esxcfg-vmknic -a -i 192.168.10.20 -n 255.255.255.0 roce_pg
   ```

9. To configure the VLAN ID, issue the following command:

   ```
   # esxcfg-vswitch -v <VLAN ID> -p roce_pg
   ```

   To run RoCE traffic with VLAN ID, configure the VLAN ID on the corresponding VMkernel port group.

Configuring MTU

To modify MTU for an RoCE interface, change the MTU of the corresponding vSwitch. Set the MTU size of the RDMA interface based on the MTU of the vSwitch by issuing the following command:

```
# esxcfg-vswitch -m <new MTU> <RoCE vswitch name>
```

For example:

```
# esxcfg-vswitch -m 4000 roce_vs
# esxcli rdma device list
Name     Driver   State    MTU  Speed    Paired Uplink  Description
-------  -------  ------  ----  -------  -------------  -------------------------------
vmrdma0  qedrntv  Active     2048   25 Gbps  vmnic0    QLogic FastLinQ QL45xxx RDMA Interface
vmrdma1  qedrntv  Active     1024   25 Gbps  vmnic1    QLogic FastLinQ QL45xxx RDMA Interface
```

RoCE Mode and Statistics

For the RoCE mode, ESXi requires concurrent support of both RoCE v1 and v2. The decision regarding which RoCE mode to use is made during queue pair creation. The ESXi driver advertises both modes during registration and initialization. To view RoCE statistics, issue the following command:

```
# esxcli rdma device stats get -d vmrdma0
   Packets received: 0
   Packets sent: 0
   Bytes received: 0
   Bytes sent: 0
   Error packets received: 0
   Error packets sent: 0
   Error length packets received: 0
   Unicast packets received: 0
   Multicast packets received: 0
   Unicast bytes received: 0
   Multicast bytes received: 0
   Unicast packets sent: 0
   Multicast packets sent: 0
   Unicast bytes sent: 0
   Multicast bytes sent: 0
   Queue pairs allocated: 0
   Queue pairs in RESET state: 0
   Queue pairs in INIT state: 0
   Queue pairs in RTR state: 0
   Queue pairs in RTS state: 0
   Queue pairs in SQD state: 0
   Queue pairs in SQE state: 0
   Queue pairs in ERR state: 0
   Queue pair events: 0
   Completion queues allocated: 1
   Completion queue events: 0
   Shared receive queues allocated: 0
   Shared receive queue events: 0
   Protection domains allocated: 1
   Memory regions allocated: 3
```

```
Address handles allocated: 0
Memory windows allocated: 0
```

Configuring a Paravirtual RDMA Device (PVRDMA)

To configure PVRDMA using a vCenter interface:

1. Create and configure a new distributed virtual switch as follows:

   a. In the VMware vSphere Web Client, right-click the **RoCE** node in the left pane of the Navigator window.

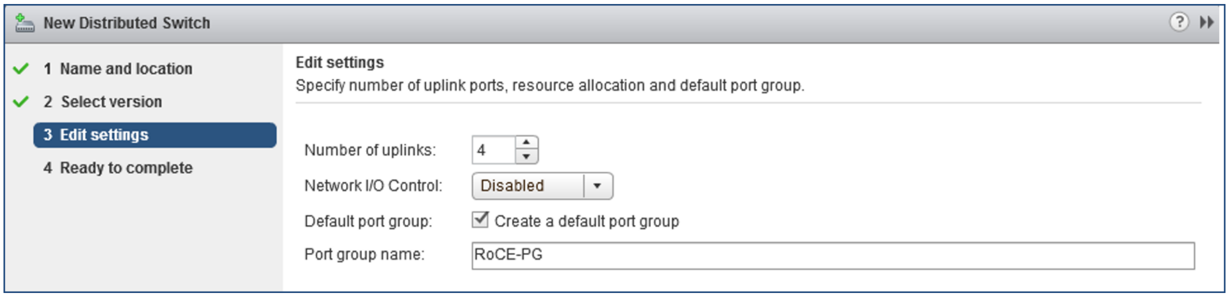
   b. On the Actions menu, point to **Distributed Switch**, and then click **New Distributed Switch**.

   c. Select version 6.5.0.

   d. Under **New Distributed Switch**, click **Edit settings**, and then configure the following:

      - **Number of uplinks**. Select an appropriate value.
      - **Network I/O Control**. Select **Disabled**.
      - **Default port group**. Select the **Create a default port group** check box.
      - **Port group name**. Type a name for the port group.

   Figure 6-8 shows an example.



Edit settings
Specify number of uplink ports, resource allocation and default port group.

Number of uplinks: 4
Network I/O Control: Disabled
Default port group: ☑ Create a default port group
Port group name: RoCE-PG

Figure 6-8. Configuring a New Distributed Switch

2. Configure a distributed virtual switch as follows:

   a. In the VMware vSphere Web Client, expand the **RoCE** node in the left pane of the Navigator window.

   b. Right-click **RoCE-VDS**, and then click **Add and Manage Hosts**.

   c. Under **Add and Manage Hosts**, configure the following:

      - **Assign uplinks**. Select from the list of available uplinks.

- **Manage VMkernel network adapters**. Accept the default, and then click **Next**.
- **Migrate VM networking**. Assign the port group created in Step 1.

3. Assign a vmknic for PVRDMA to use on ESX hosts:

   a. Right-click a host, and then click **Settings**.

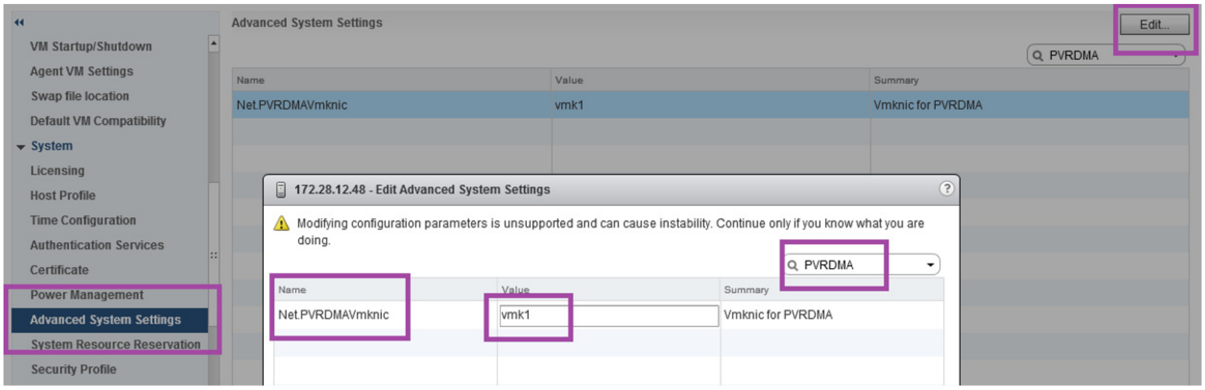
   b. On the Settings page, expand the **System** node, and then click **Advanced System Settings**.

   c. The Advanced System Settings page shows the key-pair value and its summary. Click **Edit**.

   d. On the Edit Advanced System Settings page, filter on **PVRDMA** to narrow all the settings to just Net.PVRDMAVmknic.

   e. Set the **Net.PVRDMAVmknic** value to **vmknic**; for example, **vmk1**.

   Figure 6-9 shows an example.



Advanced System Settings

Name	Value	Summary
Net.PVRDMAVmknic	vmk1	Vmknic for PVRDMA

172.28.12.48 - Edit Advanced System Settings

Modifying configuration parameters is unsupported and can cause instability. Continue only if you know what you are doing.

Name	Value	Summary
Net.PVRDMAVmknic	vmk1	Vmknic for PVRDMA

Figure 6-9. Assigning a vmknic for PVRDMA

4. Set the firewall rule for the PVRDMA:

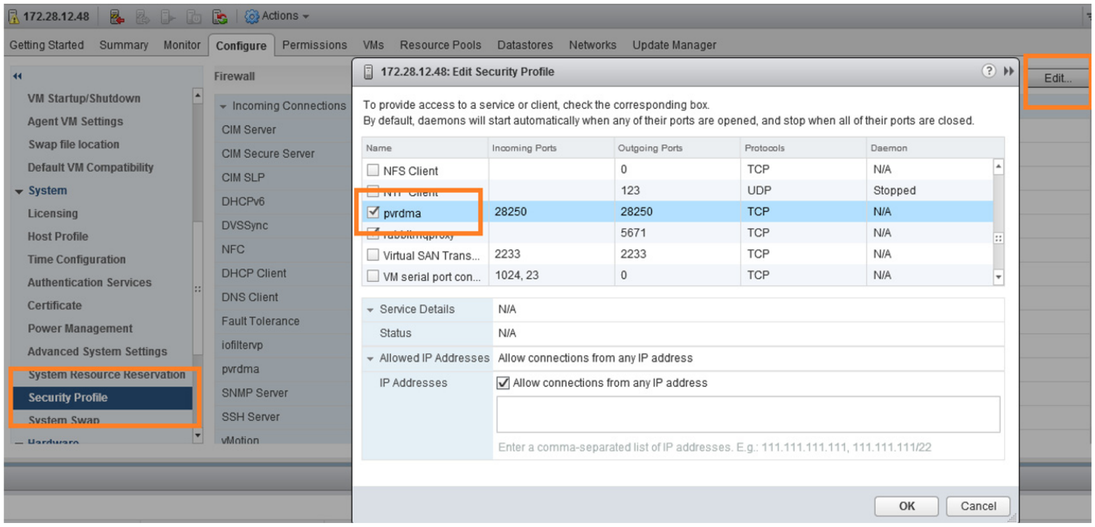
   a. Right-click a host, and then click **Settings**.

   b. On the Settings page, expand the **System** node, and then click **Security Profile**.

   c. On the Firewall Summary page, click **Edit**.

   d. In the Edit Security Profile dialog box under **Name**, scroll down, select the **pvrdma** check box, and then select the **Set Firewall** check box.

Figure 6-10 shows an example.



172.28.12.48: Edit Security Profile

To provide access to a service or client, check the corresponding box.
By default, daemons will start automatically when any of their ports are opened, and stop when all of their ports are closed.

Name	Incoming Ports	Outgoing Ports	Protocols	Daemon
□ NFS Client		0	TCP	N/A
□ NTP Client		123	UDP	Stopped
☑ pvrdma	28250	28250	TCP	N/A
☑ rabbitmqproxy		5671	TCP	N/A
□ Virtual SAN Trans...	2233	2233	TCP	N/A
□ VM serial port con...	1024, 23	0	TCP	N/A

Service Details	N/A
Status	N/A
Allowed IP Addresses	Allow connections from any IP address
IP Addresses	☑ Allow connections from any IP address Enter a comma-separated list of IP addresses. E.g.: 111.111.111.111, 111.111.111/22

OK Cancel

Figure 6-10. Setting the Firewall Rule

5. Set up the VM for PVRDMA as follows:

   a. Install one of the following supported guest OSs:

      - RHEL 7.2
      - Ubuntu 14.04 (kernel version 4.0)

   b. Install OFED-3.18.

   c. Compile and install the PVRDMA guest driver and library.

   d. Add a new PVRDMA network adapter to the VM as follows:

      - Edit the VM settings.
      - Add a new network adapter.
      - Select the newly added DVS port group as **Network**.
      - Select **PVRDMA** as the adapter type.

   e. After the VM is booted, ensure that the PVRDMA guest driver is loaded.

Configuring DCQCN

Data Center Quantized Congestion Notification (DCQCN) is a feature that determines how an RoCE receiver notifies a transmitter that a switch between them has provided an explicit congestion notification (notification point), and how a transmitter reacts to such notification (reaction point).

This section provides the following information about DCQCN configuration:

- DCQCN Terminology
- DCQCN Overview
- DCB-related Parameters
- Global Settings on RDMA Traffic
- Configuring DSCP-PFC
- Enabling DCQCN
- Configuring CNP
- DCQCN Algorithm Parameters
- MAC Statistics
- Script Example
- Limitations

DCQCN Terminology

The following terms describe DCQCN configuration:

- **TOS** (type of service) is a single-byte in the IPv4 header field. TOS comprises two ECN least significant bits (LSB) and six Differentiated Services Code Point (DSCP) most significant bits (MSB). For IPv6, traffic class is the equivalent of the IPv4 TOS.
- **ECN** (explicit congestion notification) is a mechanism where a switch adds to outgoing traffic an indication that congestion is imminent.
- **CNP** (congestion notification packet) is a packet used by the notification point to indicate that the ECN arrived from the switch back to the reaction point. CNP is defined in the Supplement to *InfiniBand Architecture Specification Volume 1 Release 1.2.1*, located here:

  https://cw.infinibandta.org/document/dl/7781
- **VLAN Priority** is a field in the L2 VLAN header. The field is the three MSBs in the VLAN tag.
- **PFC** (priority-based flow control) is a flow control mechanism that applies to traffic carrying a specific VLAN priority.

- **DSCP-PFC** is a feature that allows a receiver to interpret the priority of an incoming packet for PFC purposes, rather than according to the VLAN priority or the DSCP field in the IPv4 header. You may use an indirection table to indicate a specified DSCP value to a VLAN priority value. DSCP-PFC can work across L2 networks because it is an L3 (IPv4) feature.
- **Traffic classes**, also known as priority groups, are groups of VLAN priorities (or DSCP values if DSCP-PFC is used) that can have properties such as being lossy or lossless. Generally, 0 is used for the default common lossy traffic group, 3 is used for the FCoE traffic group, and 4 is used for the iSCSI-TLV traffic group. You may encounter DCB mismatch issues if you attempt to reuse these numbers on networks that also support FCoE or iSCSI-TLV traffic. Cavium recommends that you use numbers 1–2 or 5–7 for RoCE-related traffic groups.
- **ETS** (enhanced transition services) is an allocation of maximum bandwidth per traffic class.

DCQCN Overview

Some networking protocols (RoCE, for example) require droplessness. PFC is a mechanism for achieving droplessness in an L2 network, and DSCP-PFC is a mechanism for achieving it across distinct L2 networks. However, PFC is deficient in the following regards:

- When activated, PFC completely halts the traffic of the specified priority on the port, as opposed to reducing transmission rate.
- All traffic of the specified priority is affected, even if there is a subset of specific connections that are causing the congestion.
- PFC is a single-hop mechanism. That is, if a receiver experiences congestion and indicates the congestion through a PFC packet, only the nearest neighbor will react. When the neighbor experiences congestion (likely because it can no longer transmit), it also generates its own PFC. This generation is known as *pause propagation*. Pause propagation may cause inferior route utilization, because all buffers must congest before the transmitter is made aware of the problem.

DCQCN addresses all of these disadvantages. The ECN delivers congestion indication to the reaction point. The reaction point sends a CNP packet to the transmitter, which reacts by reducing its transmission rate and avoiding the congestion. DCQCN also specifies how the transmitter attempts to increase its transmission rate and use bandwidth effectively after congestion ceases. DCQCN is described in the 2015 SIGCOMM paper, *Congestion Control for Large-Scale RDMA Deployments*, located here:

http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf

DCB-related Parameters

Use DCB to map priorities to traffic classes (priority groups). DCB also controls which priority groups are subject to PFC (lossless traffic), and the related bandwidth allocation (ETS).

Global Settings on RDMA Traffic

Global settings on RDMA traffic include configuration of VLAN priority, ECN, and DSCP.

Setting VLAN Priority on RDMA Traffic

Use an application to set the VLAN priority used by a specified RDMA Queue Pair (QP) when creating a QP. For example, the `ib_write_bw` benchmark controls the priority using the `-sl` parameter. When RDMA-CM (RDMA Communication Manager) is present, you may be unable to set the priority.

Another method to control the VLAN priority is to use the `rdma_glob_vlan_pri` node. This method affects QPs that are created after setting the value. For example, to set the VLAN priority number to 5 for subsequently created QPs, issue the following command:

```
./debugfs.sh -n eth0 -t rdma_glob_vlan_pri 5
```

Setting ECN on RDMA Traffic

Use the `rdma_glob_ecn` node to enable ECN for a specified RoCE priority. For example, to enable ECN on RoCE traffic using priority 5, issue the following command:

```
./debugfs.sh -n eth0 -t rdma_glob_ecn 1
```

This command is typically required when DCQCN is enabled.

Setting DSCP on RDMA Traffic

Use the `rdma_glob_dscp` node to control DSCP. For example, to set DSCP on RoCE traffic using priority 5, issue the following command:

```
./debugfs.sh -n eth0 -t rdma_glob_dscp 6
```

This command is typically required when DCQCN is enabled.

Configuring DSCP-PFC

Use `dscp_pfc` nodes to configure the `dscp->priority` association for PFC. You must enable the feature before you can add entries to the map. For example, to map DSCP value 6 to priority 5, issue the following commands:

```
./debugfs.sh -n eth0 -t dscp_pfc_enable 1
./debugfs.sh -n eth0 -t dscp_pfc_set 6 5
```

Enabling DCQCN

To enable DCQCN for RoCE traffic, probe the qed driver with the `dcqcn_enable` module parameter. DCQCN requires enabled ECN indications (see "Setting ECN on RDMA Traffic" on page 95).

Configuring CNP

Congestion notification packets (CNPs) can have a separate configuration of VLAN priority and DSCP. Control these packets using the `dcqcn_cnp_dscp` and `dcqcn_cnp_vlan_priority` module parameters. For example:

```
modprobe qed dcqcn_cnp_dscp=10 dcqcn_cnp_vlan_priority=6
```

DCQCN Algorithm Parameters

Table 6-4 lists the algorithm parameters for DCQCN.

Table 6-4. DCQCN Algorithm Parameters

Parameter	Description and Values
`dcqcn_cnp_send_timeout`	Minimal difference of send time between CNPs. Units are in microseconds. Values range between 50..500000.
`dcqcn_cnp_dscp`	DSCP value to be used on CNPs. Values range between 0..63.
`dcqcn_cnp_vlan_priority`	VLAN priority to be used on CNPs. Values range between 0..7. FCoE-Offload uses `3` and iSCSI-Offload-TLV generally uses `4`. Cavium recommends that you specify a number from 1–2 or 5–7. Use this same value throughout the entire network.
`dcqcn_notification_point`	0 – Disable DCQCN notification point. 1 – Enable DCQCN notification point
`dcqcn_reaction_point`	0 – Disable DCQCN reaction point. 1 – Enable DCQCN reaction point
`dcqcn_rl_bc_rate`	Byte counter limit.
`dcqcn_rl_max_rate`	Maximum rate in Mbps.
`dcqcn_rl_r_ai`	Active increase rate in Mbps.
`dcqcn_rl_r_hai`	Hyperactive increase rate in Mbps.
`dcqcn_gd`	Alpha update gain denominator. Set to 32 for 1/32, and so on.

Table 6-4. DCQCN Algorithm Parameters (Continued)

Parameter	Description and Values
dcqcn_k_us	Alpha update interval.
dcqcn_timeout_us	DCQCN timeout.

MAC Statistics

To view MAC statistics, including per-priority PFC statistics, issue the phy_mac_stats command. For example, to view statistics on port 1 issue the following command:

```
./debugfs.sh -n eth0 -d phy_mac_stat -P 1
```

Script Example

The following example can be used as a script:

```
# probe the driver with both reaction point and notification point enabled
# with cnp dscp set to 10 and cnp vlan priority set to 6
modprobe qed dcqcn_enable=1 dcqcn_notification_point=1 dcqcn_reaction_point=1
dcqcn_cnp_dscp=10 dcqcn_cnp_vlan_priority=6
modprobe qede

# dscp-pfc configuration (associating dscp values to priorities)
# This example is using two DCBX traffic class priorities to better demonstrate
DCQCN in operation
debugfs.sh -n ens6f0 -t dscp_pfc_enable 1
debugfs.sh -n ens6f0 -t dscp_pfc_set 20 5
debugfs.sh -n ens6f0 -t dscp_pfc_set 22 6

# static DCB configurations. 0x10 is static mode. Mark priorities 5 and 6 as
# subject to pfc
debugfs.sh -n ens6f0 -t dcbx_set_mode 0x10
debugfs.sh -n ens6f0 -t dcbx_set_pfc 5 1
debugfs.sh -n ens6f0 -t dcbx_set_pfc 6 1

# set roce global overrides for qp params. enable exn and open QPs with dscp 20
debugfs.sh -n ens6f0 -t rdma_glob_ecn 1
debugfs.sh -n ens6f0 -t rdma_glob_dscp 20

# open some QPs (DSCP 20)
ib_write_bw -d qedr0 -q 16 -F -x 1 --run_infinitely

# change global dscp qp params
debugfs.sh -n ens6f0 -t rdma_glob_dscp 22

# open some more QPs (DSCP 22)
ib_write_bw -d qedr0 -q 16 -F -x 1 -p 8000 --run_infinitely
```

```
# observe PFCs being generated on multiple priorities
debugfs.sh -n ens6f0 -d phy_mac_stat -P 0 | grep "Class Based Flow Control"
```

Limitations

DCQCN has the following limitations:

- DCQCN mode currently supports only up to 64 QPs.
- Cavium adapters can determine VLAN priority for PFC purposes from VLAN priority or from DSCP bits in the TOS field. However, in the presence of both, VLAN takes precedence.

7 iWARP Configuration

Internet wide area RDMA protocol (iWARP) is a computer networking protocol that implements RDMA for efficient data transfer over IP networks. iWARP is designed for multiple environments, including LANs, storage networks, data center networks, and WANs.

This chapter provides instructions for:

- Preparing the Adapter for iWARP
- "Configuring iWARP on Windows" on page 100
- "Configuring iWARP on Linux" on page 104

NOTE

Some iWARP features may not be fully enabled in the current release. For details, refer to Appendix D Feature Constraints.

Preparing the Adapter for iWARP

This section provides instructions for preboot adapter iWARP configuration using HII. For more information about preboot adapter configuration, see Chapter 5 Adapter Preboot Configuration.

To configure iWARP through HII in Default mode:

1. Access the server BIOS System Setup, and then click **Device Settings**.
2. On the Device Settings page, select a port for the 25G 41*xxx* Series Adapter.
3. On the Main Configuration Page for the selected adapter, click **NIC Configuration**.
4. On the NIC Configuration page:
   a. Set the **NIC + RDMA Mode** to **Enabled**.
   b. Set the **RDMA Protocol Support** to **iWARP**.
   c. Click **Back**.
5. On the Main Configuration Page, click **Finish**.

6. In the Warning - Saving Changes message box, click **Yes** to save the configuration.
7. In the Success - Saving Changes message box, click **OK**.
8. Repeat Step 2 through Step 7 to configure the NIC and iWARP for the other ports.
9. To complete adapter preparation of both ports:
    a. On the Device Settings page, click **Finish**.
    b. On the main menu, click **Finish**.
    c. Exit to reboot the system.

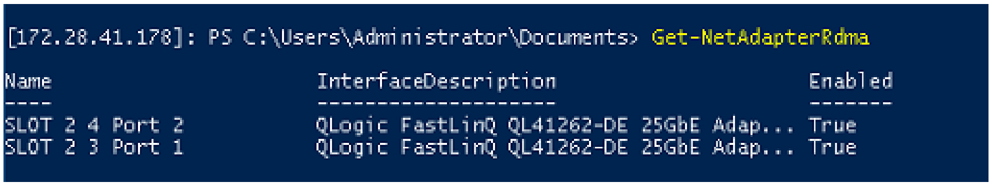
Proceed to “Configuring iWARP on Windows” on page 100 or “Configuring iWARP on Linux” on page 104.

Configuring iWARP on Windows

This section provides procedures for enabling iWARP, verifying RDMA, and verifying iWARP traffic on Windows. For a list of OSs that support iWARP, see Table 6-1 on page 63.

To enable iWARP on the Windows host and verify RDMA:

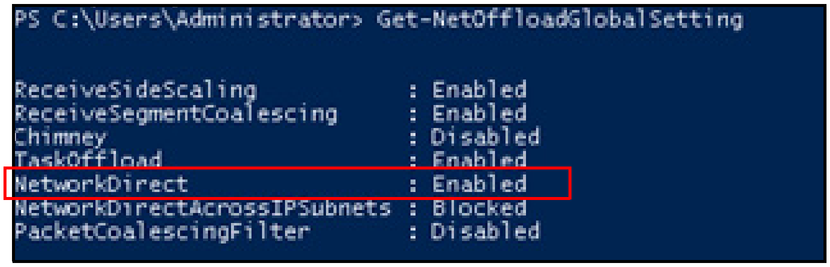
1. Enable iWARP on the Windows host.
    a. Open the Windows Device Manager, and then open the 41*xxx* Series Adapters NDIS Miniport Properties.
    b. On the FastLinQ Adapter properties, click the **Advanced** tab.
    c. On the Advanced page under **Property**, do the following:
        - Select **Network Direct Functionality**, and then select **Enabled** for the **Value**.
        - Select **RDMA Mode**, and then select **iWARP** for the **Value**.
    d. Click **OK** to save your changes and close the adapter properties.
2. Using Windows PowerShell, verify that RDMA is enabled. The `Get-NetAdapterRdma` command output (Figure 7-1) shows the adapters that support RDMA.



```
[172.28.41.178]: PS C:\Users\Administrator\Documents> Get-NetAdapterRdma

Name                      InterfaceDescription                     Enabled
----                      --------------------                     -------
SLOT 2 4 Port 2           QLogic FastLinQ QL41262-DE 25GbE Adap... True
SLOT 2 3 Port 1           QLogic FastLinQ QL41262-DE 25GbE Adap... True
```

Figure 7-1. Windows PowerShell Command: Get-NetAdapterRdma

3. Using Windows PowerShell, verify that `NetworkDirect` is enabled. The `Get-NetOffloadGlobalSetting` command output (Figure 7-2) shows `NetworkDirect` as `Enabled`.



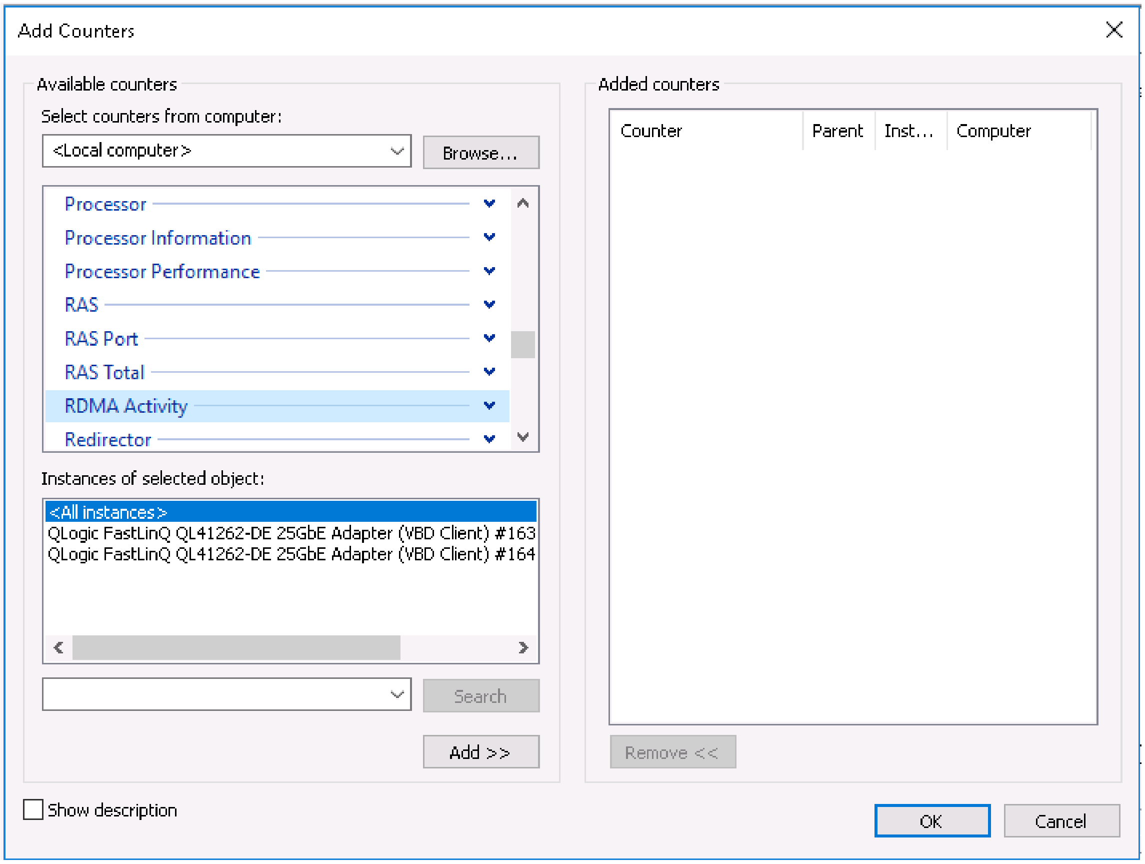
```
PS C:\Users\Administrator> Get-NetOffloadGlobalSetting


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

Figure 7-2. Windows PowerShell Command: Get-NetOffloadGlobalSetting

To verify iWARP traffic:

1. Map SMB drives and run iWARP traffic.
2. Launch Performance Monitor (Perfmon).
3. In the Add Counters dialog box, click **RDMA Activity**, and then select the adapter instances.

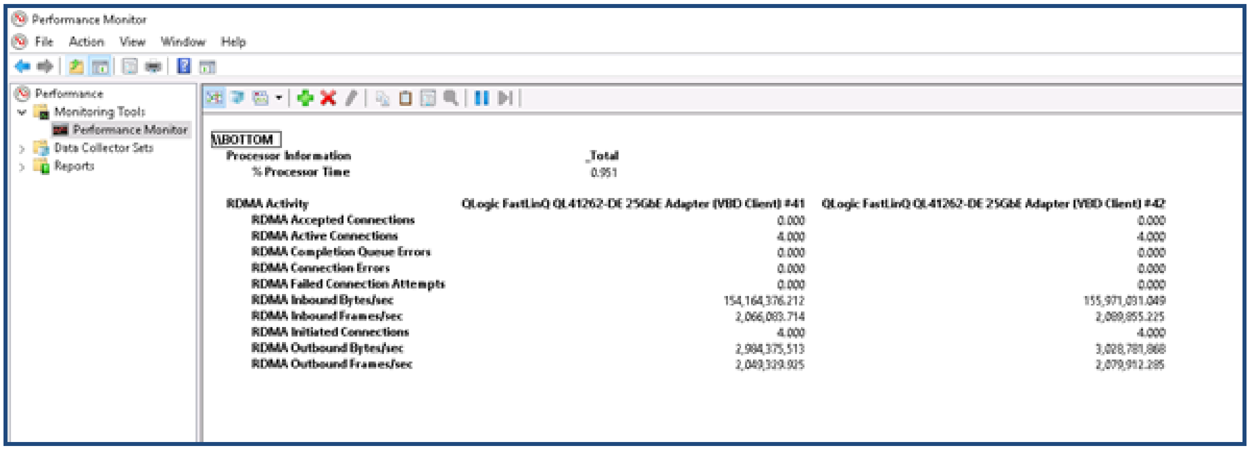
Figure 7-3 shows an example.



<All instances>
QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #163
QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #164

Figure 7-3. Perfmon: Add Counters

If iWARP traffic is running, counters appear as shown in the Figure 7-4 example.



```
\\BOTTOM
  Processor Information                         _Total
      % Processor Time                           0.951

  RDMA Activity                  QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #41   QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #42
      RDMA Accepted Connections                                                      0.000                                                          0.000
      RDMA Active Connections                                                        4.000                                                          4.000
      RDMA Completion Queue Errors                                                   0.000                                                          0.000
      RDMA Connection Errors                                                         0.000                                                          0.000
      RDMA Failed Connection Attempts                                                0.000                                                          0.000
      RDMA Inbound Bytes/sec                                               154,164,376.212                                                155,971,031.049
      RDMA Inbound Frames/sec                                                2,066,083.714                                                  2,089,855.225
      RDMA Initiated Connections                                                     4.000                                                          4.000
      RDMA Outbound Bytes/sec                                                2,984,375,513                                                  3,028,781,868
      RDMA Outbound Frames/sec                                               2,049,329.925                                                  2,079,912.285
```

Figure 7-4. Perfmon: Verifying iWARP Traffic

NOTE

For information on how to view Cavium RDMA counters in Windows, see “Viewing RDMA Counters” on page 71.

4. To verify the SMB connection:

   a. At a command prompt, issue the `net use` command as follows:

```
C:\Users\Administrator> net use
New connections will be remembered.

Status    Local      Remote                      Network
---------------------------------------------------------
OK        F:         \\192.168.10.10\Share1      Microsoft Windows Network
The command completed successfully.
```

   b. Issue the `net -xan` command as follows, where `Share1` is mapped as an SMB share:

```
C:\Users\Administrator> net -xan
Active NetworkDirect Connections, Listeners, ShareEndpoints

 Mode   IfIndex Type        Local Address        Foreign Address     PID

 Kernel       56 Connection 192.168.11.20:16159 192.168.11.10:445    0
 Kernel       56 Connection 192.168.11.20:15903 192.168.11.10:445    0
 Kernel       56 Connection 192.168.11.20:16159 192.168.11.10:445    0
 Kernel      56 Connection 192.168.11.20:15903 192.168.11.10:445    0
 Kernel      60 Listener   [fe80::e11d:9ab5:a47d:4f0a%56]:445 NA    0
 Kernel      60 Listener   192.168.11.20:445   NA                   0
 Kernel      60 Listener   [fe80::71ea:bdd2:ae41:b95f%60]:445 NA    0
 Kernel      60 Listener   192.168.11.20:16159 192.168.11.10:445    0
```

Configuring iWARP on Linux

Cavium 41*xxx* Series Adapters support iWARP on the Linux Open Fabric Enterprise Distributions (OFEDs) listed in Table 6-1 on page 63.

iWARP configuration on a Linux system includes the following:

- Installing the Driver
- Configuring iWARP and RoCE
- Detecting the Device
- Supported iWARP Applications
- Running Perftest for iWARP
- Configuring NFS-RDMA
- iWARP RDMA-Core Support on SLES 2 SP3, RHEL 7.4, and OFED 4.8x

Installing the Driver

Install the RDMA drivers as shown in Chapter 3 Driver Installation.

Configuring iWARP and RoCE

NOTE

This procedure applies only if you previously selected **iWARP+RoCE** as the value for the RDMA Protocol Support parameter during preboot configuration using HII (see Configuring NIC Parameters, Step 5 on page 47).

To enable iWARP and RoCE:

1. Unload all FastLinQ drivers as follows:

```
# modprobe -r qedr or modprobe -r qede
```

2. Use the following command syntax to change the RDMA protocol by loading the `qed` driver with a port interface PCI ID (`xx:xx.x`) and an RDMA protocol value (*p*).

   ```
   # modprobe -v qed rdma_protocol_map=<xx:xx.x-p>
   ```

   The RDMA protocol (`p`) values are as follows:

   - 0—Accept the default (RoCE)
   - 1—No RDMA
   - 2—RoCE
   - 3—iWARP

   For example, to change the interface on the port given by 04:00.0 from RoCE to iWARP, issue the following command:

   ```
   # modprobe -v qed rdma_protocol_map=04:00.0-3
   ```

3. Load the RDMA driver by issuing the following command:

   ```
   # modprobe -v qedr
   ```

The following example shows the command entries to change the RDMA protocol to iWARP on multiple NPAR interfaces:

```
# modprobe qed rdma_protocol_map=04:00.1-3,04:00.3-3,04:00.5-3,
04:00.7-3,04:01.1-3,04:01.3-3,04:01.5-3,04:01.7-3
# modprobe -v qedr
# ibv_devinfo |grep iWARP
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
```

Detecting the Device

To detect the device:

1. To verify whether RDMA devices are detected, view the `dmesg` logs:

   ```
   # dmesg |grep qedr
   [10500.191047] qedr 0000:04:00.0: registered qedr0
   [10500.221726] qedr 0000:04:00.1: registered qedr1
   ```

2. Issue the `ibv_devinfo` command, and then verify the transport type.

   If the command is successful, each PCI function will show a separate `hca_id`. For example (if checking the second port of the above dual-port adapter):

```
[root@localhost ~]# ibv_devinfo -d qedr1
hca_id: qedr1
        transport:                      iWARP (1)
        fw_ver:                         8.14.7.0
        node_guid:                      020e:1eff:fec4:c06e
        sys_image_guid:                 020e:1eff:fec4:c06e
        vendor_id:                      0x1077
        vendor_part_id:                 5718
        hw_ver:                         0x0
        phys_port_cnt:                  1
                port:   1
                       state:                  PORT_ACTIVE (4)
                        max_mtu:                4096 (5)
                        active_mtu:             1024 (3)
                        sm_lid:                 0
                        port_lid:               0
                        port_lmc:               0x00
                        link_layer:             Ethernet
```

Supported iWARP Applications

Linux-supported RDMA applications for iWARP include the following:

- ibv_devinfo, ib_devices
- ib_send_bw/lat, ib_write_bw/lat, ib_read_bw/lat, ib_atomic_bw/lat
  For iWARP, all applications must use the RDMA communication manager (rdma_cm) using the `-R` option.
- rdma_server, rdma_client
- rdma_xserver, rdma_xclient
- rping
- NFS over RDMA (NFSoRDMA)
- iSER (for details, see Chapter 8 iSER Configuration)
- NVMe-oF (for details, see Chapter 12 NVMe-oF Configuration with RDMA)

Running Perftest for iWARP

All perftest tools are supported over the iWARP transport type. You must run the tools using the RDMA connection manager (with the `-R` option).

Example:

1. On one server, issue the following command (using the second port in this example):

   ```
   # ib_send_bw -d qedr1 -F -R
   ```

2. On one client, issue the following command (using the second port in this example):

```
[root@localhost ~]# ib_send_bw -d qedr1 -F -R 192.168.11.3
----------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : qedr1
 Number of qps   : 1            Transport type : IW
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 GID index       : 0
 Max inline data : 0[B]
 rdma_cm QPs     : ON
 Data ex. method : rdma_cm
----------------------------------------------------------------------------
 local address: LID 0000 QPN 0x0192 PSN 0xcde932
 GID: 00:14:30:196:192:110:00:00:00:00:00:00:00:00:00:00
 remote address: LID 0000 QPN 0x0098 PSN 0x46fffc
 GID: 00:14:30:196:195:62:00:00:00:00:00:00:00:00:00:00
----------------------------------------------------------------------------
 #bytes   #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536     1000             2250.38            2250.36              0.036006
----------------------------------------------------------------------------
```

NOTE

For latency applications (send/write), if the perftest version is the latest (for example, `perftest-3.0-0.21.g21dc344.x86_64.rpm`), use the supported inline size value: `0-128`.

Configuring NFS-RDMA

NFS-RDMA for iWARP includes both server and client configuration steps.

To configure the NFS server:

1. In the `/etc/exports` file for the directories that you must export using NFS-RDMA on the server, make the following entry:

   ```
   /tmp/nfs-server *(fsid=0,async,insecure,no_root_squash)
   ```

   Ensure that you use a different file system identification (FSID) for each directory that you export.

2. Load the svcrdma module as follows:

   ```
   # modprobe svcrdma
   ```

3. Start the NFS service without any errors:

   ```
   # service nfs start
   ```

4. Include the default RDMA port 20049 into this file as follows:

   ```
   # echo rdma 20049 > /proc/fs/nfsd/portlist
   ```

5. To make local directories available for NFS clients to mount, issue the `exportfs` command as follows:

   ```
   # exportfs -v
   ```

To configure the NFS client:

NOTE

This procedure for NFS client configuration also applies to RoCE.

1. Load the xprtrdma module as follows:

   ```
   # modprobe xprtrdma
   ```

2. Mount the NFS file system as appropriate for your version:

   For NFS Version 3:

   ```
   #mount -o rdma,port=20049 192.168.2.4:/tmp/nfs-server /tmp/nfs-client
   ```

   For NFS Version 4:

   ```
   #mount -t nfs4 -o rdma,port=20049 192.168.2.4:/ /tmp/nfs-client
   ```

   NOTE

   The default port for NFSoRDMA is 20049. However, any other port that is aligned with the NFS client will also work.

3. Verify that the file system is mounted by issuing the `mount` command. Ensure that the RDMA port and file system versions are correct.

   ```
   #mount |grep rdma
   ```

iWARP RDMA-Core Support on SLES 2 SP3, RHEL 7.4, and OFED 4.8*x*

The user space library libqedr is part of the rdma-core. However, the out-of-box libqedr does not support SLES 12 SP3, RHEL 7.4, OFED 4.8*x*. Therefore, these OS versions require a patch to support iWARP RDMA-Core.

To apply the iWARP RDMA-Core patch:

1. To download the latest RDMA-core source, issue the following command:

   ```
   # git clone https://github.com/linux-rdma/rdma-core.git
   ```

2. Install all OS-dependent packages/libraries as described in the *RDMA-Core README*.

   For RHEL and CentOS, issue the following command:

   ```
   # yum install cmake gcc libnl3-devel libudev-devel make
   pkgconfig valgrind-devel
   ```

   For SLES 12 SP3 (ISO/SDK kit), install the following RPMs:

   `cmake-3.5.2-18.3.x86_64.rpm` (OS ISO)

   `libnl-1_1-devel-1.1.4-4.21.x86_64.rpm` (SDK ISO)

   `libnl3-devel-3.2.23-2.21.x86_64.rpm` (SDK ISO)

3. To build the RDMA-core, issue the following commands:

   ```
   # cd <rdma-core-path>/rdma-core-master/
   # ./build.sh
   ```

4. To run all OFED applications from the current RDMA-core-master location, issue the following command:

```
# ls <rdma-core-master>/build/bin
cmpost  ib_acme         ibv_devinfo      ibv_uc_pingpong
iwpmd  rdma_client  rdma_xclient  rping       ucmatose
umad_compile_test cmtime  ibv_asyncwatch  ibv_rc_pingpong
ibv_ud_pingpong   mckey  rdma-ndd    rdma_xserver  rstream
udaddy    umad_reg2 ibacm   ibv_devices     ibv_srq_pingpong
ibv_xsrq_pingpong  rcopy  rdma_server  riostream
srp_daemon  udpong    umad_register2
```

Run applications from the current RDMA-Core-master location. For example:

```
# ./rping -c -v -C 5 -a 192.168.21.3
ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
client DISCONNECT EVENT...
```

5. To run inbox OFED applications such as perftest and other InfiniBand applications, issue the following command to set the library path for iWARP:

```
# export
LD_LIBRARY_PATH=/builds/rdma-core-path-iwarp/rdma-core-master/build/lib
```

For example:

```
# /usr/bin/rping -c -v -C 5 -a 192.168.22.3 (or) rping -c -v -C 5 -a
192.168.22.3
ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
client DISCONNECT EVENT...
```

8 iSER Configuration

This chapter provides procedures for configuring iSCSI Extensions for RDMA (iSER) for Linux (RHEL and SLES) and ESXi 6.7, including:

Before You Begin

As you prepare to configure iSER, consider the following:

- iSER is supported only in inbox OFED for the following operating systems:
  - ❑ RHEL 7.1 and 7.2
  - ❑ SLES 12 and 12 SP1
- After logging into the targets or while running I/O traffic, unloading the Linux RoCE qedr driver may crash the system.
- While running I/O, performing interface down/up tests or performing cable pull-tests can cause driver or iSER module errors that may crash the system. If this happens, reboot the system.

Configuring iSER for RHEL

To configure iSER for RHEL:

1. Install inbox OFED as described in "RoCE Configuration for RHEL" on page 77. Out-of-box OFEDs are not supported for iSER because the ib_isert module is not available in the out-of-box OFED 3.18-2 GA/3.18-3 GA versions. The inbox ib_isert module does not work with any out-of-box OFED versions.

2. Unload any existing FastLinQ drivers as described in "Removing the Linux Drivers" on page 10.

3. Install the latest FastLinQ driver and libqedr packages as described in "Installing the Linux Drivers with RDMA" on page 14.

4. Load the RDMA services as follows;

   ```
   systemctl start rdma
   modprobe qedr
   modprobe ib_iser
   modprobe ib_isert
   ```

5. Verify that all RDMA and iSER modules loaded on the initiator and target devices by issuing the `lsmod | grep qed` and `lsmod | grep iser` commands.

6. Verify that there are separate `hca_id` instances by issuing the `ibv_devinfo` command, as shown in Step 6 on page 79.

7. Check the RDMA connection on the initiator device and the target device.

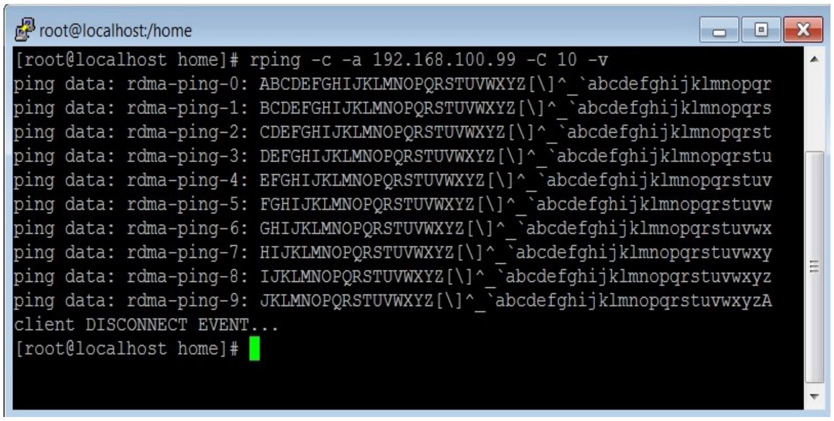
   a. On the initiator device, issue the following command:

      ```
      rping -s -C 10 -v
      ```

   b. On the target device, issue the following command:

      ```
      rping -c -a 192.168.100.99 -C 10 -v
      ```

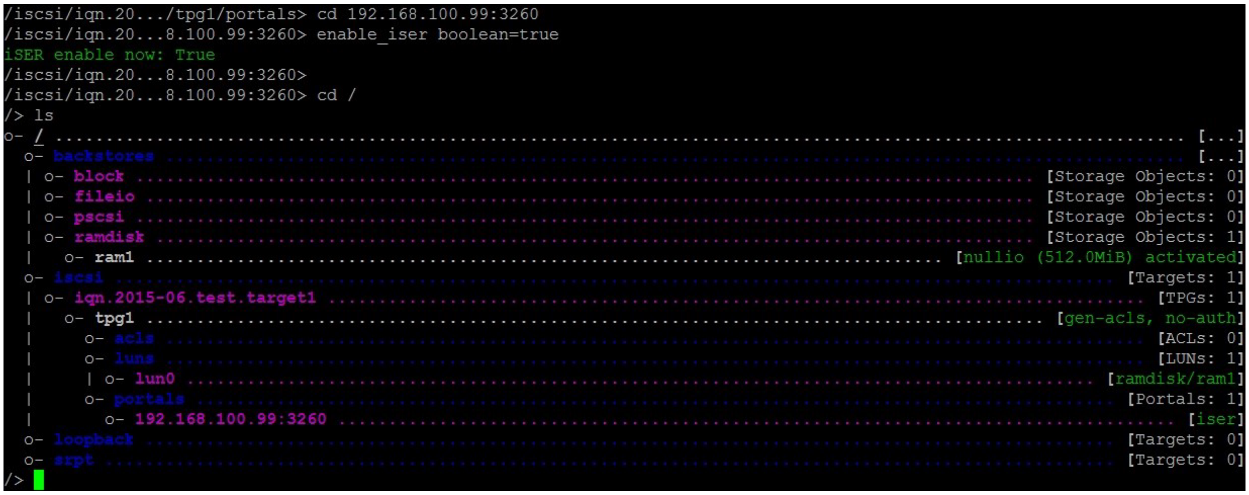
   Figure 8-1 shows an example of a successful RDMA ping.



```
[root@localhost home]# rping -c -a 192.168.100.99 -C 10 -v
ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
ping data: rdma-ping-5: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvw
ping data: rdma-ping-6: GHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwx
ping data: rdma-ping-7: HIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
ping data: rdma-ping-8: IJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
ping data: rdma-ping-9: JKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzA
client DISCONNECT EVENT...
[root@localhost home]#
```

Figure 8-1. RDMA Ping Successful

8. You can use a Linux TCM-LIO target to test iSER. The setup is the same for any iSCSI target, except that you issue the command `enable_iser Boolean=true` on the applicable portals. The portal instances are identified as **`iser`** in Figure 8-2.



```
/iscsi/iqn.20.../tpg1/portals> cd 192.168.100.99:3260
/iscsi/iqn.20...8.100.99:3260> enable_iser boolean=true
iSER enable now: True
/iscsi/iqn.20...8.100.99:3260>
/iscsi/iqn.20...8.100.99:3260> cd /
/> ls
o- / ......................................................................... [...]
  o- backstores .............................................................. [...]
  | o- block .................................................. [Storage Objects: 0]
  | o- fileio ................................................. [Storage Objects: 0]
  | o- pscsi .................................................. [Storage Objects: 0]
  | o- ramdisk ................................................ [Storage Objects: 1]
  |   o- ram1 ............................................ [nullio (512.0MiB) activated]
  o- iscsi ............................................................ [Targets: 1]
  | o- iqn.2015-06.test.target1 ............................................ [TPGs: 1]
  |   o- tpg1 ................................................... [gen-acls, no-auth]
  |     o- acls ........................................................ [ACLs: 0]
  |     o- luns ........................................................ [LUNs: 1]
  |     | o- lun0 ................................................ [ramdisk/ram1]
  |     o- portals .................................................. [Portals: 1]
  |       o- 192.168.100.99:3260 ............................................. [iser]
  o- loopback ......................................................... [Targets: 0]
  o- srpt ............................................................. [Targets: 0]
/>
```

Figure 8-2. iSER Portal Instances

9. Install Linux iSCSI Initiator Utilities using the `yum install iscsi-initiator-utils` commands.

   a. To discover the iSER target, issue the `iscsiadm` command. For example:

      `iscsiadm -m discovery -t st -p 192.168.100.99:3260`

   b. To change the transport mode to iSER, issue the `iscsiadm` command. For example:

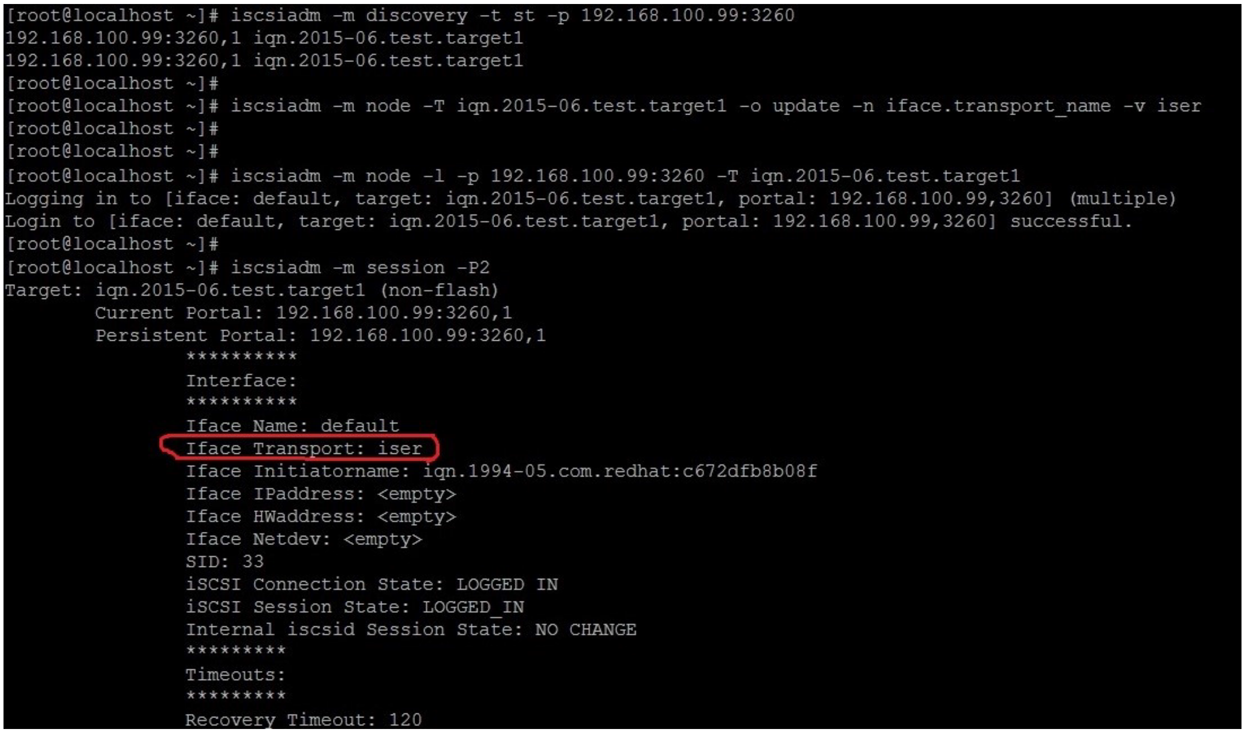
      ```
      iscsiadm -m node -T iqn.2015-06.test.target1 -o update -n
      iface.transport_name -v iser
      ```

   c. To connect to or log in to the iSER target, issue the `iscsiadm` command. For example:

      ```
      iscsiadm -m node -l -p 192.168.100.99:3260 -T
      iqn.2015-06.test.target1
      ```

   d. Confirm that the `Iface Transport` is **`iser`** in the target connection, as shown in Figure 8-3. Issue the `iscsiadm` command; for example:

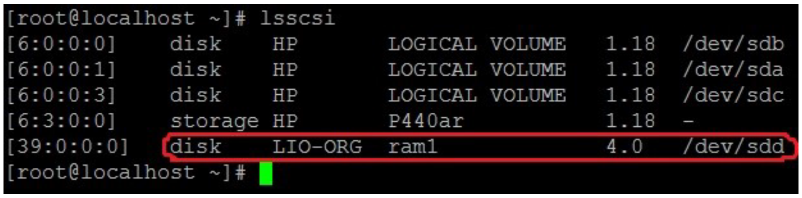
      `iscsiadm -m session -P2`



```
[root@localhost ~]# iscsiadm -m discovery -t st -p 192.168.100.99:3260
192.168.100.99:3260,1 iqn.2015-06.test.target1
192.168.100.99:3260,1 iqn.2015-06.test.target1
[root@localhost ~]#
[root@localhost ~]# iscsiadm -m node -T iqn.2015-06.test.target1 -o update -n iface.transport_name -v iser
[root@localhost ~]#
[root@localhost ~]#
[root@localhost ~]# iscsiadm -m node -l -p 192.168.100.99:3260 -T iqn.2015-06.test.target1
Logging in to [iface: default, target: iqn.2015-06.test.target1, portal: 192.168.100.99,3260] (multiple)
Login to [iface: default, target: iqn.2015-06.test.target1, portal: 192.168.100.99,3260] successful.
[root@localhost ~]#
[root@localhost ~]# iscsiadm -m session -P2
Target: iqn.2015-06.test.target1 (non-flash)
        Current Portal: 192.168.100.99:3260,1
        Persistent Portal: 192.168.100.99:3260,1
                **********
                Interface:
                **********
                Iface Name: default
                Iface Transport: iser
                Iface Initiatorname: iqn.1994-05.com.redhat:c672dfb8b08f
                Iface IPaddress: <empty>
                Iface HWaddress: <empty>
                Iface Netdev: <empty>
                SID: 33
                iSCSI Connection State: LOGGED IN
                iSCSI Session State: LOGGED_IN
                Internal iscsid Session State: NO CHANGE
                *********
                Timeouts:
                *********
                Recovery Timeout: 120
```

Figure 8-3. Iface Transport Confirmed

e. To check for a new iSCSI device, as shown in Figure 8-4, issue the `lsscsi` command.



```
[root@localhost ~]# lsscsi
[6:0:0:0]    disk    HP       LOGICAL VOLUME   1.18  /dev/sdb
[6:0:0:1]    disk    HP       LOGICAL VOLUME   1.18  /dev/sda
[6:0:0:3]    disk    HP       LOGICAL VOLUME   1.18  /dev/sdc
[6:3:0:0]    storage HP       P440ar           1.18  -
[39:0:0:0]   disk    LIO-ORG  ram1             4.0   /dev/sdd
[root@localhost ~]#
```

Figure 8-4. Checking for New iSCSI Device

Configuring iSER for SLES 12

Because the targetcli is not inbox on SLES 12.*x*, you must complete the following procedure.

To configure iSER for SLES 12:

1. To install targetcli, copy and install the following RPMs from the ISO image (x86_64 and noarch location):

```
lio-utils-4.1-14.6.x86_64.rpm
python-configobj-4.7.2-18.10.noarch.rpm
python-PrettyTable-0.7.2-8.5.noarch.rpm
python-configshell-1.5-1.44.noarch.rpm
python-pyparsing-2.0.1-4.10.noarch.rpm
python-netifaces-0.8-6.55.x86_64.rpm
python-rtslib-2.2-6.6.noarch.rpm
python-urwid-1.1.1-6.144.x86_64.rpm
targetcli-2.1-3.8.x86_64.rpm
```

2. Before starting the targetcli, load all RoCE device drivers and iSER modules as follows:

```
# modprobe qed
# modprobe qede
# modprobe qedr
# modprobe ib_iser   (Initiator)
# modprobe ib_isert  (Target)
```

3. Before configuring iSER targets, configure NIC interfaces and run L2 and RoCE traffic, as described in Step 7 on page 79.

4. Start the targetcli utility, and configure your targets on the iSER target system.

> **NOTE**
>
> targetcli versions are different in RHEL and SLES. Be sure to use the proper backstores to configure your targets:
>
> - RHEL uses *ramdisk*
> - SLES uses *rd_mcp*

Using iSER with iWARP on RHEL and SLES

Configure the iSER initiator and target similar to RoCE to work with iWARP. You can use different methods to create a Linux-IO Target (LIO™); one is listed in this section. You may encounter some difference in targetcli configuration in SLES 12 and RHEL 7.*x* because of the version.

To configure a target for LIO:

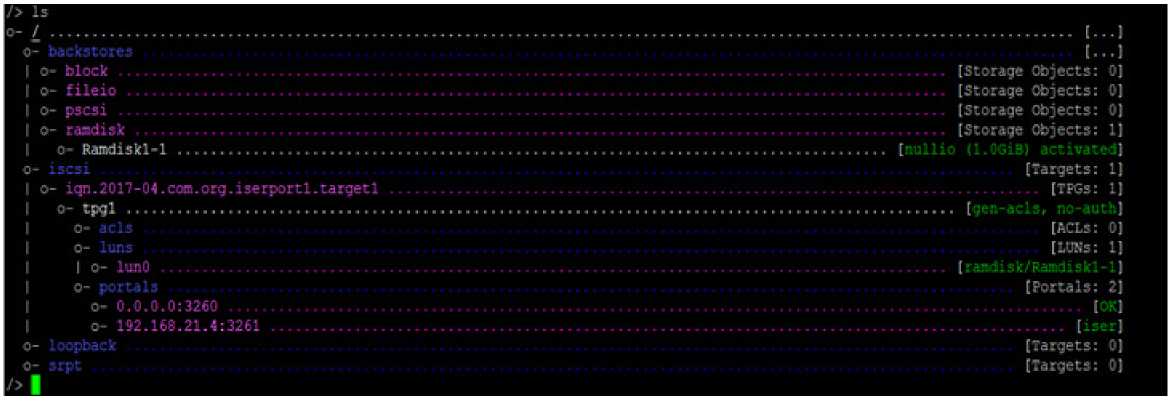
1. Create an LIO target using the targetcli utility. Issue the following command:

   ```
   # targetcli
   targetcli shell version 2.1.fb41
   Copyright 2011-2013 by Datera, Inc and others.
   For help on commands, type 'help'.
   ```

2. Issue the following commands:

```
/> /backstores/ramdisk create Ramdisk1-1 1g nullio=true
/> /iscsi create iqn.2017-04.com.org.iserport1.target1
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1/luns create /backstores/ramdisk/Ramdisk1-1
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1/portals/ create 192.168.21.4 ip_port=3261
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1/portals/192.168.21.4:3261 enable_iser
boolean=true
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1 set attribute authentication=0
demo_mode_write_protect=0 generate_node_acls=1 cache_dynamic_acls=1
/> saveconfig
```

Figure 8-5 shows the target configuration for LIO.



```
/> ls
o- / ......................................................................................... [...]
  o- backstores .............................................................................. [...]
  | o- block .................................................................. [Storage Objects: 0]
  | o- fileio ................................................................. [Storage Objects: 0]
  | o- pscsi .................................................................. [Storage Objects: 0]
  | o- ramdisk ................................................................ [Storage Objects: 1]
  |   o- Ramdisk1-1 ....................................................... [nullio (1.0GiB) activated]
  o- iscsi ............................................................................ [Targets: 1]
  | o- iqn.2017-04.com.org.iserport1.target1 ............................................... [TPGs: 1]
  |   o- tpg1 ................................................................... [gen-acls, no-auth]
  |     o- acls .............................................................................. [ACLs: 0]
  |     o- luns .............................................................................. [LUNs: 1]
  |     | o- lun0 ........................................................... [ramdisk/Ramdisk1-1]
  |     o- portals ....................................................................... [Portals: 2]
  |       o- 0.0.0.0:3260 ............................................................................ [OK]
  |       o- 192.168.21.4:3261 ..................................................................... [iser]
  o- loopback ......................................................................... [Targets: 0]
  o- srpt ............................................................................. [Targets: 0]
/>
```

Figure 8-5. LIO Target Configuration

To configure an initiator for iWARP:

1. To discover the iSER LIO target using port 3261, issue the `iscsiadm` command as follows:

   ```
   # iscsiadm -m discovery -t st -p 192.168.21.4:3261 -I iser
   192.168.21.4:3261,1 iqn.2017-04.com.org.iserport1.target1
   ```

2. Change the transport mode to `iser` as follows:

```
# iscsiadm -m node -o update -T iqn.2017-04.com.org.iserport1.target1 -n
iface.transport_name -v iser
```

3. Log into the target using port 3261:

```
# iscsiadm -m node -l -p 192.168.21.4:3261 -T iqn.2017-04.com.org.iserport1.target1
Logging in to [iface: iser, target: iqn.2017-04.com.org.iserport1.target1,
portal: 192.168.21.4,3261] (multiple)
Login to [iface: iser, target: iqn.2017-04.com.org.iserport1.target1, portal:
192.168.21.4,3261] successful.
```

4. Ensure that those LUNs are visible by issuing the following command:

   ```
   # lsscsi
   [1:0:0:0]    storage HP       P440ar           3.56  -
   [1:1:0:0]    disk    HP       LOGICAL VOLUME   3.56  /dev/sda
   [6:0:0:0]    cd/dvd  hp       DVD-ROM DUD0N    UMD0  /dev/sr0
   [7:0:0:0]    disk    LIO-ORG  Ramdisk1-1       4.0   /dev/sdb
   ```

Optimizing Linux Performance

Consider the following Linux performance configuration enhancements described in this section.

- Configuring CPUs to Maximum Performance Mode
- Configuring Kernel sysctl Settings
- Configuring IRQ Affinity Settings
- Configuring Block Device Staging

Configuring CPUs to Maximum Performance Mode

Configure the CPU scaling governor to performance by using the following script to set all CPUs to maximum performance mode:

```
for CPUFREQ in
/sys/devices/system/cpu/cpu*/cpufreq/scaling_governor; do [ -f
$CPUFREQ ] || continue; echo -n performance > $CPUFREQ; done
```

Verify that all CPU cores are set to maximum performance mode by issuing the following command:

```
cat /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor
```

Configuring Kernel sysctl Settings

Set the kernel sysctl settings as follows:

```
sysctl -w net.ipv4.tcp_mem="4194304 4194304 4194304"
sysctl -w net.ipv4.tcp_wmem="4096 65536 4194304"
sysctl -w net.ipv4.tcp_rmem="4096 87380 4194304"
sysctl -w net.core.wmem_max=4194304
sysctl -w net.core.rmem_max=4194304
sysctl -w net.core.wmem_default=4194304
sysctl -w net.core.rmem_default=4194304
sysctl -w net.core.netdev_max_backlog=250000
sysctl -w net.ipv4.tcp_timestamps=0
sysctl -w net.ipv4.tcp_sack=1
sysctl -w net.ipv4.tcp_low_latency=1
sysctl -w net.ipv4.tcp_adv_win_scale=1
echo 0 > /proc/sys/vm/nr_hugepages
```

Configuring IRQ Affinity Settings

The following example sets CPU core 0, 1, 2, and 3 to interrupt request (IRQ) XX, YY, ZZ, and XYZ respectively. Perform these steps for each IRQ assigned to a port (default is eight queues per port).

```
systemctl disable irqbalance
systemctl stop irqbalance
cat /proc/interrupts | grep qedr   Shows IRQ assigned to each port queue
echo 1 > /proc/irq/XX/smp_affinity_list
echo 2 > /proc/irq/YY/smp_affinity_list
echo 4 > /proc/irq/ZZ/smp_affinity_list
echo 8 > /proc/irq/XYZ/smp_affinity_list
```

Configuring Block Device Staging

Set the block device staging settings for each iSCSI device or target as follows:

```
echo noop > /sys/block/sdd/queue/scheduler
echo 2 > /sys/block/sdd/queue/nomerges
echo 0 > /sys/block/sdd/queue/add_random
echo 1 > /sys/block/sdd/queue/rq_affinity
```

Configuring iSER on ESXi 6.7

This section provides information for configuring iSER for VMware ESXi 6.7.

Before You Begin

Before you configure iSER for ESXi 6.7, ensure that the following is complete:

- The CNA package with NIC and RoCE drivers is installed on the ESXi 6.7 system and the devices are listed. To view RDMA devices, issue the following command:

```
esxcli rdma device list
Name     Driver   State    MTU  Speed    Paired Uplink  Description
-------  -------  ------  ----  -------  -------------  --------------------------------------
vmrdma0  qedrntv  Active  1024  40 Gbps  vmnic4         QLogic FastLinQ QL45xxx RDMA Interface
vmrdma1  qedrntv  Active  1024  40 Gbps  vmnic5         QLogic FastLinQ QL45xxx RDMA Interface
[root@localhost:~] esxcfg-vmknic -l
Interface  Port Group/DVPort/Opaque Network        IP Family IP Address
Netmask         Broadcast       MAC Address        MTU     TSO MSS   Enabled Type
NetStack
vmk0       Management Network                     IPv4      172.28.12.94
255.255.240.0   172.28.15.255   e0:db:55:0c:5f:94 1500     65535     true    DHCP
defaultTcpipStack
vmk0       Management Network                     IPv6      fe80::e2db:55ff:fe0c:5f94
64                              e0:db:55:0c:5f:94 1500     65535     true    STATIC, PREFERRED
defaultTcpipStack
```

- The iSER target is configured to communicate with the iSER initiator.

Configuring iSER for ESXi 6.7

To configure iSER for ESXi 6.7:

1. Add iSER devices by issuing the following commands:

```
esxcli rdma iser add
esxcli iscsi adapter list
Adapter  Driver  State    UID            Description
-------  ------  -------  -------------  -------------------------------------
vmhba64  iser    unbound  iscsi.vmhba64  VMware iSCSI over RDMA (iSER) Adapter
vmhba65  iser    unbound  iscsi.vmhba65  VMware iSCSI over RDMA (iSER) Adapter
```

2. Disable the firewall as follows.

```
esxcli network firewall set --enabled=false
esxcli network firewall unload
vsish -e set /system/modules/iscsi_trans/loglevels/iscsitrans 0
vsish -e set /system/modules/iser/loglevels/debug 4
```

3. Create a standard vSwitch VMkernel port group and assign the IP:

```
esxcli network vswitch standard add -v vSwitch_iser1
esxcfg-nics -l
Name    PCI          Driver      Link Speed      Duplex MAC Address       MTU    Description
vmnic0  0000:01:00.0 ntg3        Up   1000Mbps   Full   e0:db:55:0c:5f:94 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic1  0000:01:00.1 ntg3        Down 0Mbps      Half   e0:db:55:0c:5f:95 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic2  0000:02:00.0 ntg3        Down 0Mbps      Half   e0:db:55:0c:5f:96 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic3  0000:02:00.1 ntg3        Down 0Mbps      Half   e0:db:55:0c:5f:97 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic4  0000:42:00.0 qedentv     Up   40000Mbps  Full   00:0e:1e:d5:f6:a2 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 10/25/40/50/100 GbE Ethernet Adapter
vmnic5  0000:42:00.1 qedentv     Up   40000Mbps  Full   00:0e:1e:d5:f6:a3 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 10/25/40/50/100 GbE Ethernet Adapter

esxcli network vswitch standard uplink add -u vmnic5 -v vSwitch_iser1
esxcli network vswitch standard portgroup add -p "rdma_group1" -v vSwitch_iser1
esxcli network ip interface add -i vmk1 -p "rdma_group1"
esxcli network ip interface ipv4 set -i vmk1 -I 192.168.10.100 -N 255.255.255.0 -t static
esxcfg-vswitch -p "rdma_group1" -v 4095 vSwitch_iser1
esxcli iscsi networkportal add -A vmhba67 -n vmk1
esxcli iscsi networkportal list
esxcli iscsi adapter get -A vmhba65
vmhba65
   Name: iqn.1998-01.com.vmware:localhost.punelab.qlogic.com qlogic.org qlogic.com
mv.qlogic.com:1846573170:65
   Alias: iser-vmnic5
   Vendor: VMware
   Model: VMware iSCSI over RDMA (iSER) Adapter
   Description: VMware iSCSI over RDMA (iSER) Adapter
   Serial Number: vmnic5
   Hardware Version:
   Asic Version:
   Firmware Version:
   Option Rom Version:
   Driver Name: iser-vmnic5
   Driver Version:
   TCP Protocol Supported: false
   Bidirectional Transfers Supported: false
```

```
Maximum Cdb Length: 64
Can Be NIC: true
Is NIC: true
Is Initiator: true
Is Target: false
Using TCP Offload Engine: true
Using ISCSI Offload Engine: true
```

4. Add the target to the iSER initiator as follows:

```
esxcli iscsi adapter target list
esxcli iscsi adapter discovery sendtarget add -A vmhba65 -a 192.168.10.11
esxcli iscsi adapter target list
Adapter  Target                    Alias  Discovery Method  Last Error
-------  ------------------------  -----  ----------------  ----------
vmhba65  iqn.2015-06.test.target1         SENDTARGETS       No Error
esxcli storage core adapter rescan --adapter vmhba65
```

5. List the attached target as follows:

```
esxcfg-scsidevs -l
mpx.vmhba0:C0:T4:L0
   Device Type: CD-ROM
   Size: 0 MB
   Display Name: Local TSSTcorp CD-ROM (mpx.vmhba0:C0:T4:L0)
   Multipath Plugin: NMP
   Console Device: /vmfs/devices/cdrom/mpx.vmhba0:C0:T4:L0
   Devfs Path: /vmfs/devices/cdrom/mpx.vmhba0:C0:T4:L0
   Vendor: TSSTcorp  Model: DVD-ROM SN-108BB  Revis: D150
   SCSI Level: 5  Is Pseudo: false Status: on
   Is RDM Capable: false Is Removable: true
   Is Local: true  Is SSD: false
   Other Names:
      vml.0005000000766d686261303a343a30
   VAAI Status: unsupported
naa.6001405e81ae36b771c418b89c85dae0
   Device Type: Direct-Access
   Size: 512 MB
   Display Name: LIO-ORG iSCSI Disk (naa.6001405e81ae36b771c418b89c85dae0)
   Multipath Plugin: NMP
   Console Device: /vmfs/devices/disks/naa.6001405e81ae36b771c418b89c85dae0
   Devfs Path: /vmfs/devices/disks/naa.6001405e81ae36b771c418b89c85dae0
   Vendor: LIO-ORG   Model: ram1              Revis: 4.0
   SCSI Level: 5  Is Pseudo: false Status: degraded
   Is RDM Capable: true  Is Removable: false
   Is Local: false Is SSD: false
   Other Names:
      vml.02000000006001405e81ae36b771c418b89c85dae072616d312020
   VAAI Status: supported
naa.690b11c0159d050018255e2d1d59b612
```

9 iSCSI Configuration

This chapter provides the following iSCSI configuration information:

- iSCSI Boot
- "Configuring iSCSI Boot" on page 130
- "Configuring the DHCP Server to Support iSCSI Boot" on page 140
- "iSCSI Offload in Windows Server" on page 144
- "iSCSI Offload in Linux Environments" on page 153

NOTE

Some iSCSI features may not be fully enabled in the current release. For details, refer to Appendix D Feature Constraints.

iSCSI Boot

Cavium 4*xxxx* Series gigabit Ethernet (GbE) adapters support iSCSI boot to enable network boot of operating systems to diskless systems. iSCSI boot allows a Windows, Linux, or VMware operating system to boot from an iSCSI target machine located remotely over a standard IP network.

iSCSI boot information in this section includes:

- iSCSI Boot Setup
- Adapter UEFI Boot Mode Configuration

For both Windows and Linux operating systems, iSCSI boot can be configured with **UEFI iSCSI HBA** (offload path with the QLogic offload iSCSI driver). Set this option using Boot Protocol, under **Port Level Configuration**.

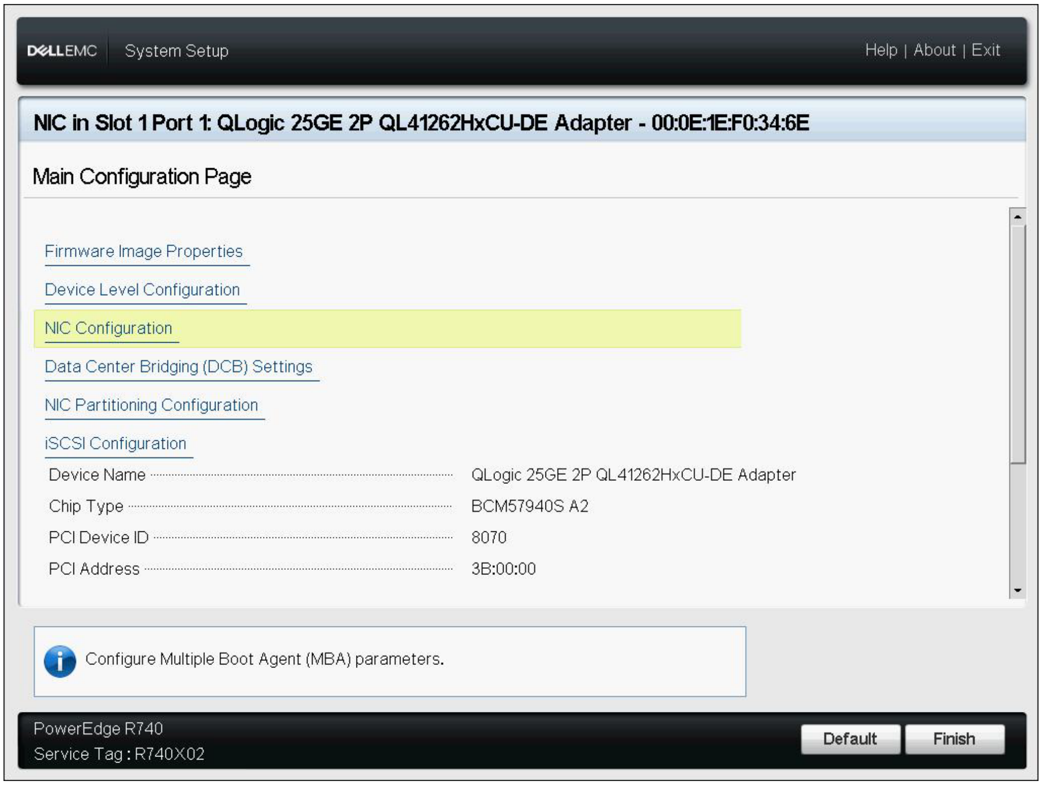
iSCSI Boot Setup

The iSCSI boot setup includes:

- Selecting the Preferred iSCSI Boot Mode
- Configuring the iSCSI Target
- Configuring iSCSI Boot Parameters

Selecting the Preferred iSCSI Boot Mode

The **Boot Mode** option is listed under **iSCSI Configuration** (Figure 9-1) of the adapter, and the setting is port specific. Refer to the OEM user manual for direction on accessing the device level configuration menu under UEFI HII.


DELLEMC
System Setup
Help | About | Exit

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page

Firmware Image Properties
Device Level Configuration
NIC Configuration
Data Center Bridging (DCB) Settings
NIC Partitioning Configuration
iSCSI Configuration

Device Name .... QLogic 25GE 2P QL41262HxCU-DE Adapter
Chip Type BCM57940S A2
PCI Device ID 8070
PCI Address 3B:00:00

Configure Multiple Boot Agent (MBA) parameters.

PowerEdge R740
Service Tag : R740X02

Figure 9-1. System Setup: NIC Configuration

NOTE

Boot from SAN boot is supported only in NPAR mode and is configured in UEFI, and not in legacy BIOS.

Configuring the iSCSI Target

Configuring the iSCSI target varies by target vendors. For information on configuring the iSCSI target, refer to the documentation provided by the vendor.

To configure the iSCSI target:

1. Select the appropriate procedure based on your iSCSI target, either:

   ❑ Create an iSCSI target for targets such as SANBlaze® or IET®.

   ❑ Create a vdisk or volume for targets such as EqualLogic® or EMC®.

2. Create a virtual disk.
3. Map the virtual disk to the iSCSI target created in Step 1.
4. Associate an iSCSI initiator with the iSCSI target. Record the following information:
   - ❑ iSCSI target name
   - ❑ TCP port number
   - ❑ iSCSI Logical Unit Number (LUN)
   - ❑ Initiator iSCSI qualified name (IQN)
   - ❑ CHAP authentication details
5. After configuring the iSCSI target, obtain the following:
   - ❑ Target IQN
   - ❑ Target IP address
   - ❑ Target TCP port number
   - ❑ Target LUN
   - ❑ Initiator IQN
   - ❑ CHAP ID and secret

Configuring iSCSI Boot Parameters

Configure the QLogic iSCSI boot software for either static or dynamic configuration. For configuration options available from the General Parameters window, see Table 9-1, which lists parameters for both IPv4 and IPv6. Parameters that are specific to either IPv4 or IPv6 are noted.

NOTE

The availability of the IPv6 iSCSI boot is platform- and device-dependent.

Table 9-1. Configuration Options

Option	Description
`TCP/IP parameters via DHCP`	This option is specific to IPv4. Controls whether the iSCSI boot host software acquires the IP address information using DHCP (`Enabled`) or use a static IP configuration (`Disabled`).
`iSCSI parameters via DHCP`	Controls whether the iSCSI boot host software acquires its iSCSI target parameters using DHCP (`Enabled`) or through a static configuration (`Disabled`). The static information is entered on the iSCSI Initiator Parameters Configuration page.

Table 9-1. Configuration Options (Continued)

Option	Description
`CHAP Authentication`	Controls whether the iSCSI boot host software uses CHAP authentication when connecting to the iSCSI target. If `CHAP Authentication` is enabled, configure the CHAP ID and CHAP Secret on the iSCSI Initiator Parameters Configuration page.
`IP Version`	This option is specific to IPv6. Toggles between `IPv4` and `IPv6`. All IP settings are lost if you switch from one protocol version to another.
`DHCP Request Timeout`	Allows you to specify a maximum wait time in seconds for a DHCP request, and response to complete.
`Target Login Timeout`	Allows you to specify a maximum wait time in seconds for the initiator to complete target login.
`DHCP Vendor ID`	Controls how the iSCSI boot host software interprets the `Vendor Class ID` field used during DHCP. If the `Vendor Class ID` field in the DHCP offer packet matches the value in the field, the iSCSI boot host software looks into the DHCP Option 43 fields for the required iSCSI boot extensions. If DHCP is disabled, this value does not need to be set.

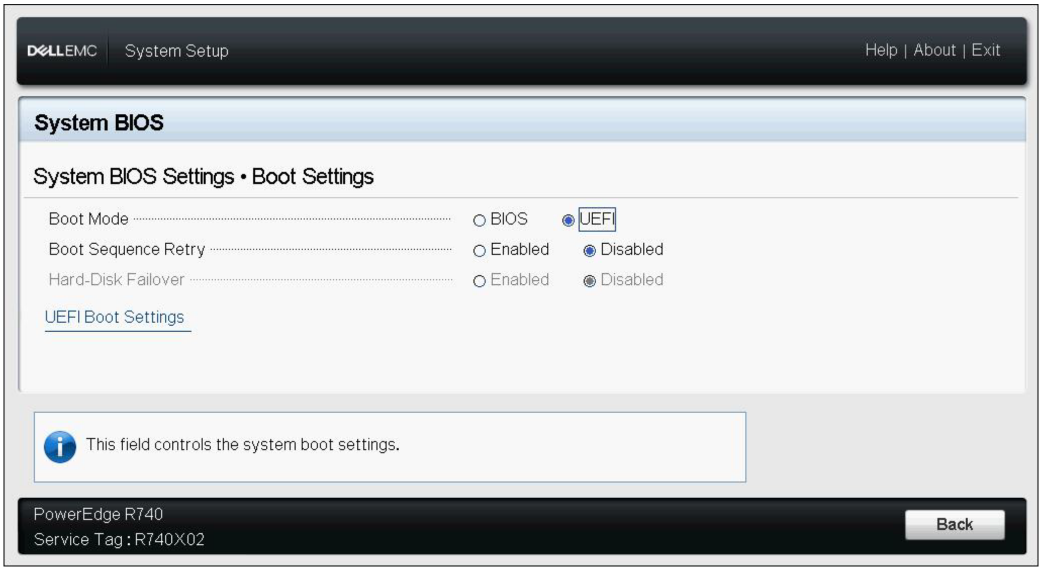
Adapter UEFI Boot Mode Configuration

To configure the boot mode:

1. Restart the system.
2. Access the System Utilities menu (Figure 9-2).

NOTE

SAN boot is supported in UEFI environment only. Make sure the system boot option is UEFI, and not legacy.



System BIOS

System BIOS Settings • Boot Settings

Boot Mode ○ BIOS ● UEFI

Boot Sequence Retry ○ Enabled ● Disabled

Hard-Disk Failover ○ Enabled ● Disabled

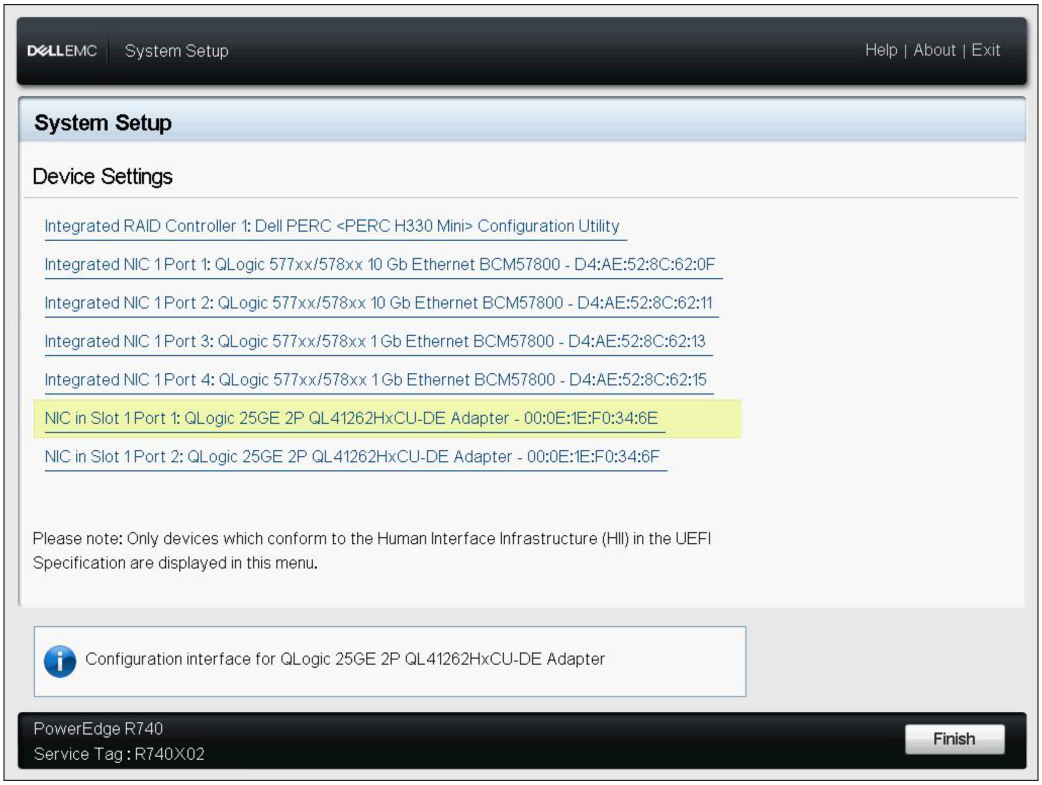
UEFI Boot Settings

This field controls the system boot settings.

PowerEdge R740
Service Tag : R740X02

Figure 9-2. System Setup: Boot Settings

3. In System Setup, Device Settings, select the QLogic device (Figure 9-3). Refer to the OEM user guide on accessing the PCI device configuration menu.



System Setup

Device Settings

Integrated RAID Controller 1: Dell PERC <PERC H330 Mini> Configuration Utility

Integrated NIC 1 Port 1: QLogic 577xx/578xx 10 Gb Ethernet BCM57800 - D4:AE:52:8C:62:0F

Integrated NIC 1 Port 2: QLogic 577xx/578xx 10 Gb Ethernet BCM57800 - D4:AE:52:8C:62:11

Integrated NIC 1 Port 3: QLogic 577xx/578xx 1 Gb Ethernet BCM57800 - D4:AE:52:8C:62:13

Integrated NIC 1 Port 4: QLogic 577xx/578xx 1 Gb Ethernet BCM57800 - D4:AE:52:8C:62:15

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

NIC in Slot 1 Port 2: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6F

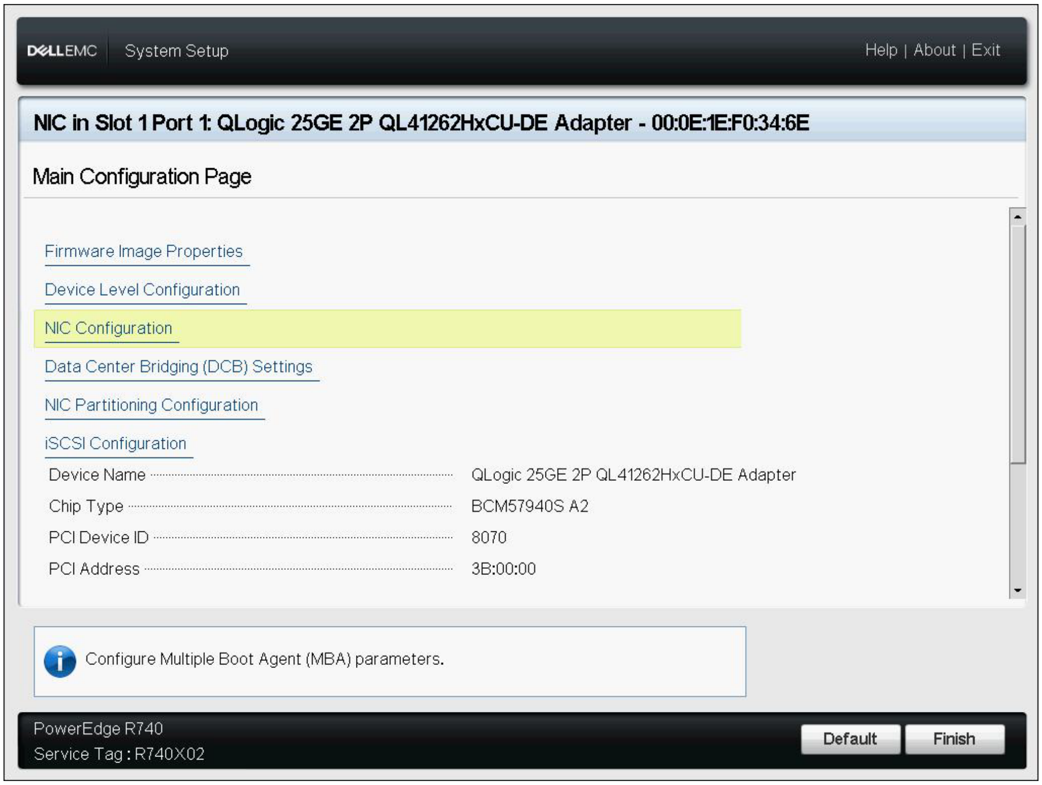
Please note: Only devices which conform to the Human Interface Infrastructure (HII) in the UEFI Specification are displayed in this menu.

Configuration interface for QLogic 25GE 2P QL41262HxCU-DE Adapter

PowerEdge R740
Service Tag : R740X02

Figure 9-3. System Setup: Device Settings Configuration Utility

4. On the Main Configuration Page, select **NIC Configuration** (Figure 9-4), and then press ENTER.



DELLEMC
System Setup
Help | About | Exit

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page

Firmware Image Properties
Device Level Configuration
NIC Configuration
Data Center Bridging (DCB) Settings
NIC Partitioning Configuration
iSCSI Configuration

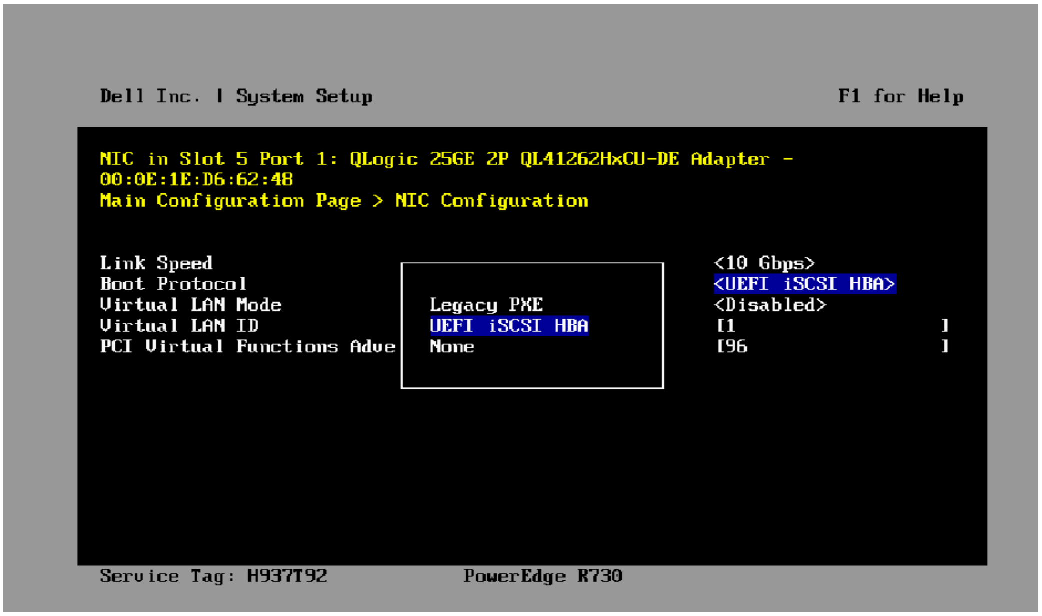
Device Name	QLogic 25GE 2P QL41262HxCU-DE Adapter
Chip Type	BCM57940S A2
PCI Device ID	8070
PCI Address	3B:00:00

Configure Multiple Boot Agent (MBA) parameters.

PowerEdge R740
Service Tag : R740X02

Figure 9-4. Selecting NIC Configuration

5. On the NIC Configuration page (Figure 9-5), select **Boot Protocol**, and then press ENTER to select **UEFI iSCSI HBA** (requires NPAR mode).



```
Dell Inc. | System Setup                                          F1 for Help

NIC in Slot 5 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter -
00:0E:1E:D6:62:48
Main Configuration Page > NIC Configuration

Link Speed                                          <10 Gbps>
Boot Protocol                                       <UEFI iSCSI HBA>
Virtual LAN Mode           Legacy PXE               <Disabled>
Virtual LAN ID             UEFI iSCSI HBA           [1            ]
PCI Virtual Functions Adve None                     [96           ]

Service Tag: H937T92               PowerEdge R730
```

Figure 9-5. System Setup: NIC Configuration, Boot Protocol

6. Proceed with one of the following configuration options:
   - ❑ "Static iSCSI Boot Configuration" on page 130
   - ❑ "Dynamic iSCSI Boot Configuration" on page 137

Configuring iSCSI Boot

iSCSI boot configuration options include:

- Static iSCSI Boot Configuration
- Dynamic iSCSI Boot Configuration
- Enabling CHAP Authentication

Static iSCSI Boot Configuration

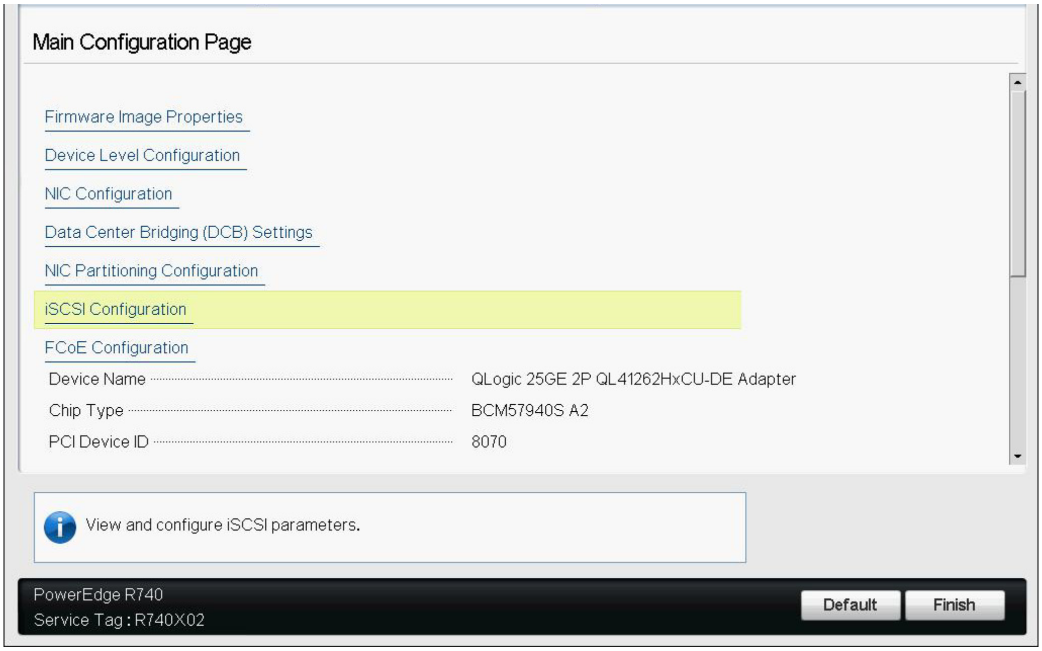
In a static configuration, you must enter data for the following:

- System's IP address
- System's initiator IQN
- Target parameters (obtained in "Configuring the iSCSI Target" on page 124)

For information on configuration options, see Table 9-1 on page 125.

To configure the iSCSI boot parameters using static configuration:

1. In the Device HII **Main Configuration Page**, select **iSCSI Configuration** (Figure 9-6), and then press ENTER.



Main Configuration Page

Firmware Image Properties
Device Level Configuration
NIC Configuration
Data Center Bridging (DCB) Settings
NIC Partitioning Configuration
iSCSI Configuration
FCoE Configuration

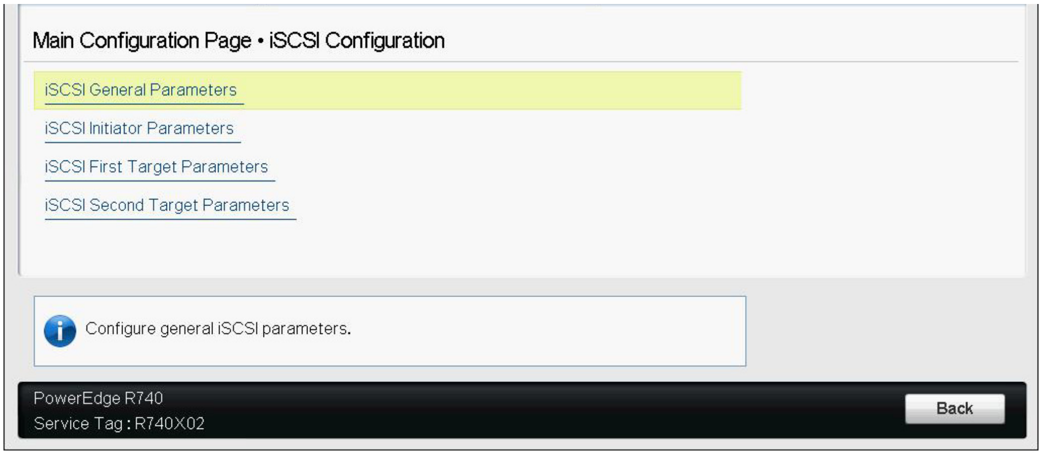
Device Name .... QLogic 25GE 2P QL41262HxCU-DE Adapter
Chip Type BCM57940S A2
PCI Device ID 8070

View and configure iSCSI parameters.

PowerEdge R740
Service Tag : R740X02

Figure 9-6. System Setup: iSCSI Configuration

2. On the **iSCSI Configuration** page, select **iSCSI General Parameters** (Figure 9-7), and then press ENTER.



Main Configuration Page • iSCSI Configuration

iSCSI General Parameters
iSCSI Initiator Parameters
iSCSI First Target Parameters
iSCSI Second Target Parameters

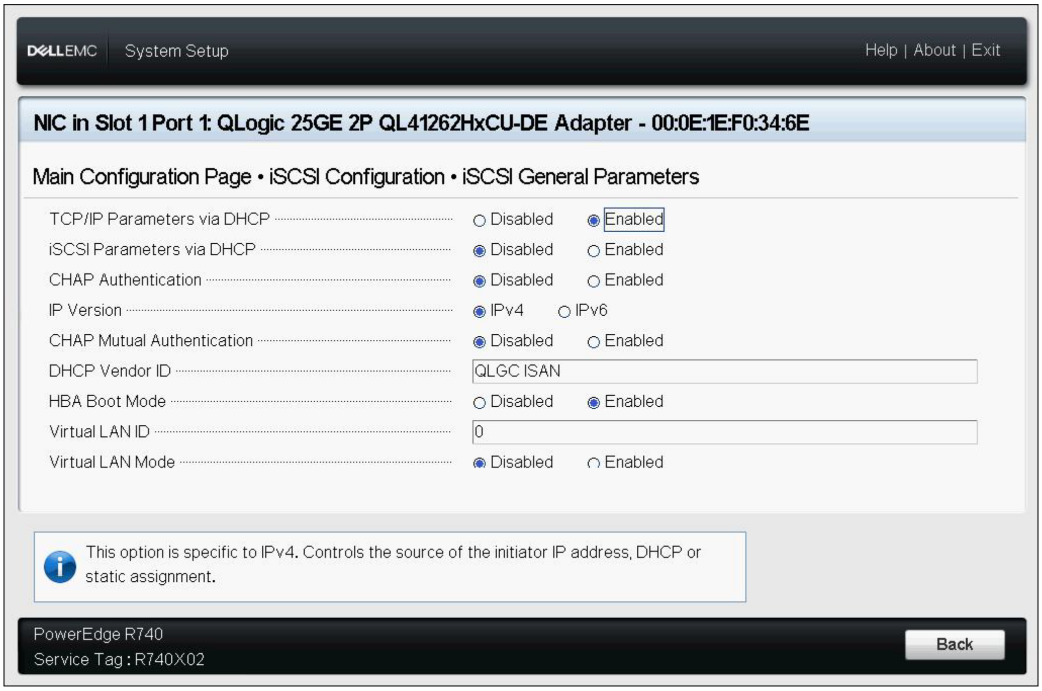
Configure general iSCSI parameters.

PowerEdge R740
Service Tag : R740X02

Figure 9-7. System Setup: Selecting General Parameters

3. On the iSCSI General Parameters page (Figure 9-8), press the DOWN ARROW key to select a parameter, and then press the ENTER key to input the following values:

   ❑ **TCP/IP Parameters via DHCP**: Disabled
   ❑ **iSCSI Parameters via DHCP**: Disabled
   ❑ **CHAP Authentication**: As required
   ❑ **IP Version**: As required (IPv4 or IPv6)
   ❑ **CHAP Mutual Authentication**: As required
   ❑ **DHCP Vendor ID**: Not applicable for static configuration
   ❑ **HBA Boot Mode**: Enabled
   ❑ **Virtual LAN ID**: Default value or as required
   ❑ **Virtual LAN Mode**: Disabled



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration • iSCSI General Parameters

TCP/IP Parameters via DHCP	○ Disabled	● Enabled
iSCSI Parameters via DHCP	● Disabled	○ Enabled
CHAP Authentication	● Disabled	○ Enabled
IP Version	● IPv4	○ IPv6
CHAP Mutual Authentication	● Disabled	○ Enabled
DHCP Vendor ID	QLGC ISAN	
HBA Boot Mode	○ Disabled	● Enabled
Virtual LAN ID	0	
Virtual LAN Mode	● Disabled	○ Enabled

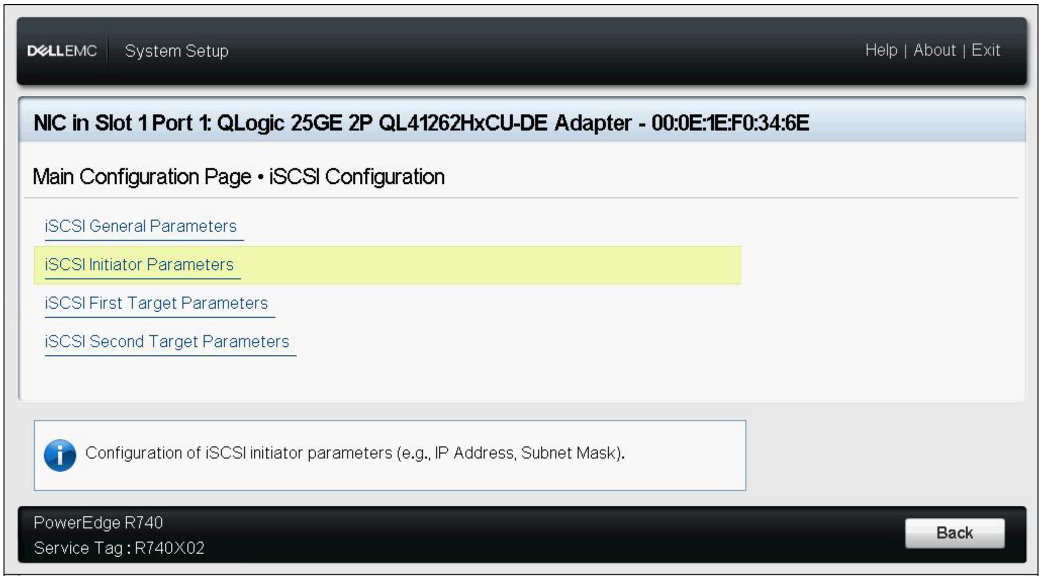
This option is specific to IPv4. Controls the source of the initiator IP address, DHCP or static assignment.

PowerEdge R740
Service Tag : R740X02

Figure 9-8. System Setup: iSCSI General Parameters

4. Return to the iSCSI Configuration page, and then press the ESC key.

5. Select **iSCSI Initiator Parameters** (Figure 9-9), and then press ENTER.



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration

iSCSI General Parameters

iSCSI Initiator Parameters

iSCSI First Target Parameters

iSCSI Second Target Parameters

Configuration of iSCSI initiator parameters (e.g., IP Address, Subnet Mask).

PowerEdge R740
Service Tag : R740X02

Figure 9-9. System Setup: Selecting iSCSI Initiator Parameters

6. On the iSCSI Initiator Parameters page (Figure 9-10), select the following parameters, and then type a value for each:

   □ **IPv4* Address**

   □ **Subnet Mask**

   □ **IPv4* Default Gateway**

   □ **IPv4* Primary DNS**

   □ **IPv4* Secondary DNS**

   □ **iSCSI Name**. Corresponds to the iSCSI initiator name to be used by the client system.

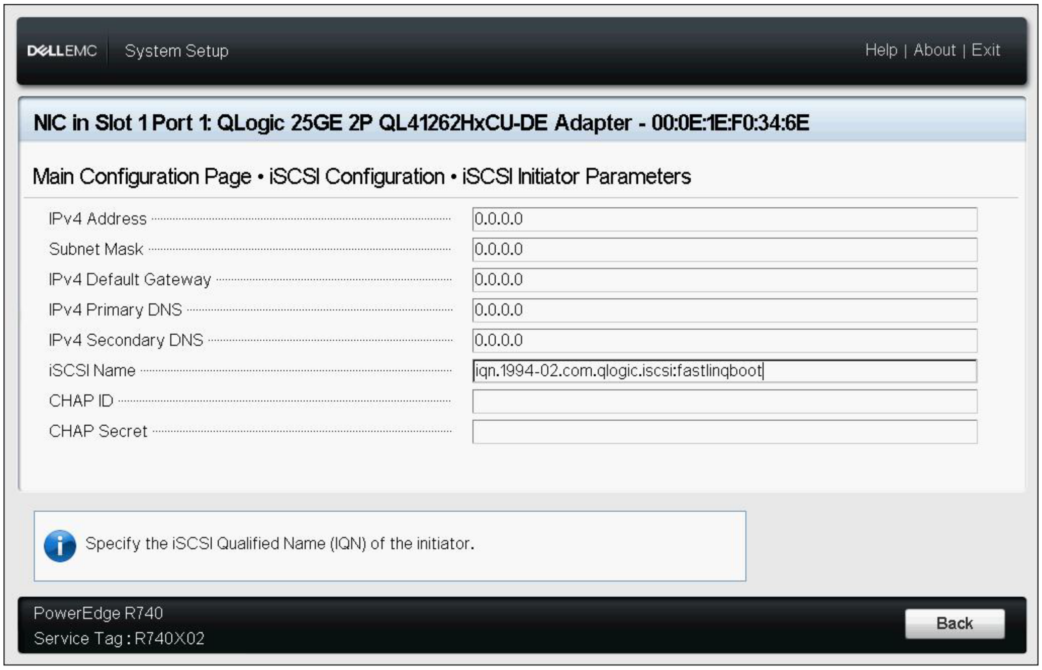
   □ **CHAP ID**

   □ **CHAP Secret**

NOTE

Note the following for the preceding items with asterisks (*):

- The label will change to **IPv6** or **IPv4** (default) based on the IP version set on the iSCSI General Parameters page (Figure 9-8 on page 132).
- Carefully enter the IP address. There is no error-checking performed against the IP address to check for duplicates, incorrect segment, or network assignment.



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration • iSCSI Initiator Parameters

IPv4 Address	0.0.0.0
Subnet Mask	0.0.0.0
IPv4 Default Gateway	0.0.0.0
IPv4 Primary DNS	0.0.0.0
IPv4 Secondary DNS	0.0.0.0
iSCSI Name	iqn.1994-02.com.qlogic.iscsi:fastlinqboot
CHAP ID	
CHAP Secret	

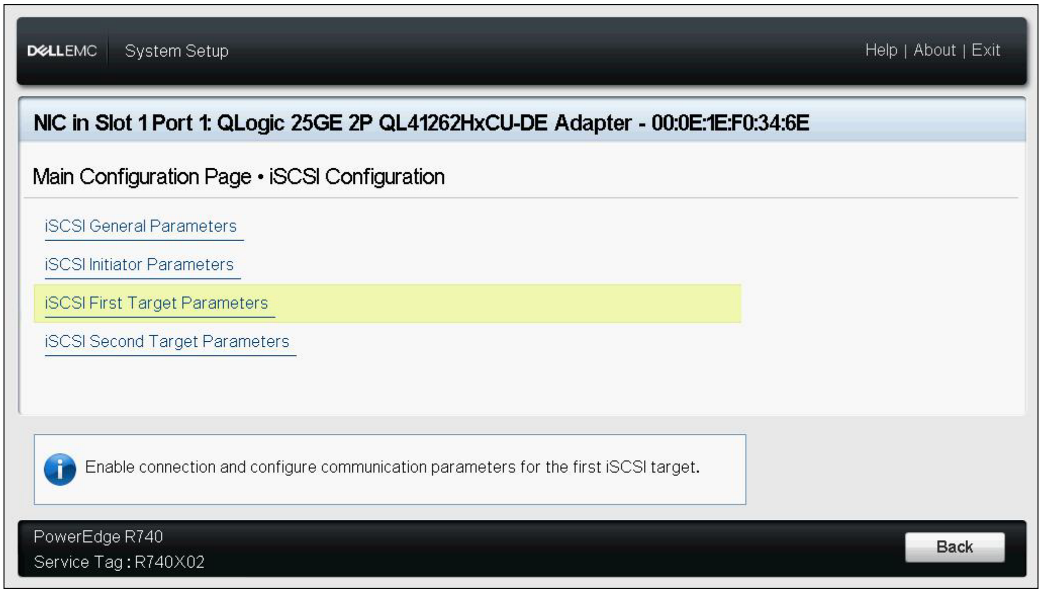
Specify the iSCSI Qualified Name (IQN) of the initiator.

PowerEdge R740
Service Tag : R740X02

Figure 9-10. System Setup: iSCSI Initiator Parameters

7. Return to the iSCSI Configuration page, and then press ESC.

8. Select **iSCSI First Target Parameters** (Figure 9-11), and then press ENTER.


DELLEMC
System Setup
Help | About | Exit
iSCSI General Parameters
iSCSI Initiator Parameters
iSCSI First Target Parameters
iSCSI Second Target Parameters
Back

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration

Enable connection and configure communication parameters for the first iSCSI target.

PowerEdge R740
Service Tag : R740X02

Figure 9-11. System Setup: Selecting iSCSI First Target Parameters

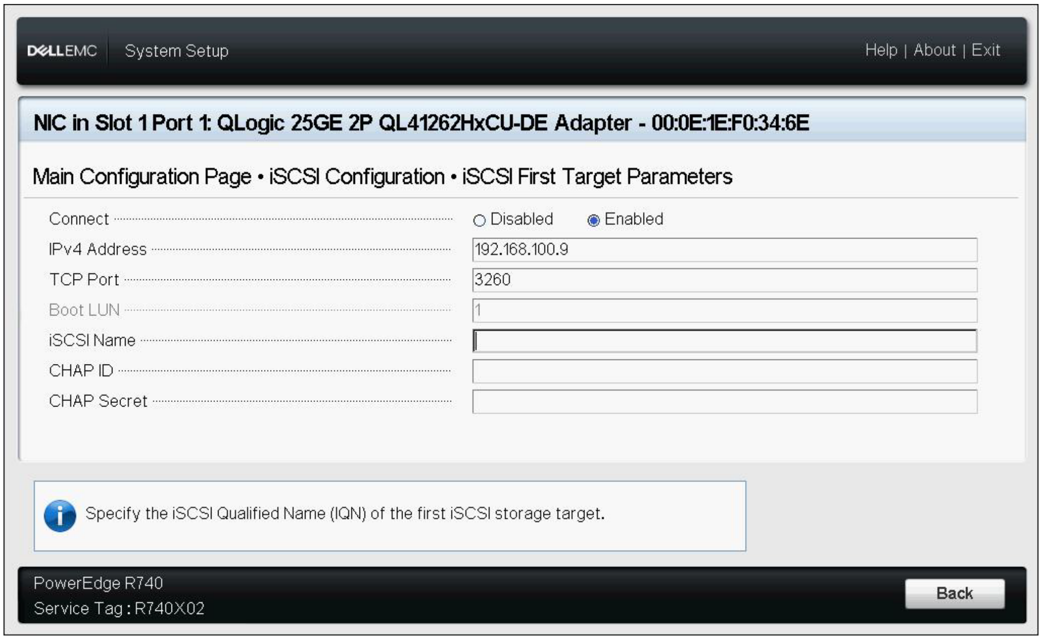
9. On the iSCSI First Target Parameters page, set the **Connect** option to **Enabled** to the iSCSI target.

10. Type values for the following parameters for the iSCSI target, and then press ENTER:

    ❑ **IPv4* Address**
    ❑ **TCP Port**
    ❑ **Boot LUN**
    ❑ **iSCSI Name**
    ❑ **CHAP ID**
    ❑ **CHAP Secret**

NOTE

For the preceding parameters with an asterisk (*), the label will change to **IPv6** or **IPv4** (default) based on IP version set on the iSCSI General Parameters page, as shown in Figure 9-12.



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration • iSCSI First Target Parameters

Connect	○ Disabled ● Enabled
IPv4 Address	192.168.100.9
TCP Port	3260
Boot LUN	1
iSCSI Name	
CHAP ID	
CHAP Secret	

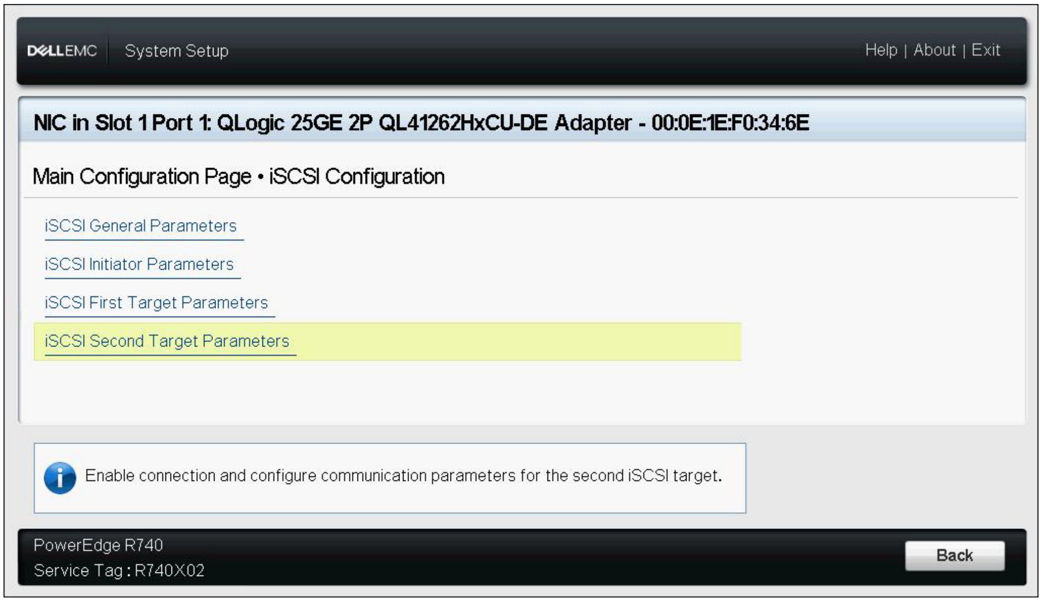
Specify the iSCSI Qualified Name (IQN) of the first iSCSI storage target.

PowerEdge R740
Service Tag : R740X02

Figure 9-12. System Setup: iSCSI First Target Parameters

11. Return to the iSCSI Boot Configuration page, and then press ESC.

12. If you want configure a second iSCSI target device, select **iSCSI Second Target Parameters** (Figure 9-13), and enter the parameter values as you did in Step 10. Otherwise, proceed to Step 13.


DELLEMC
System Setup
Help | About | Exit

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration

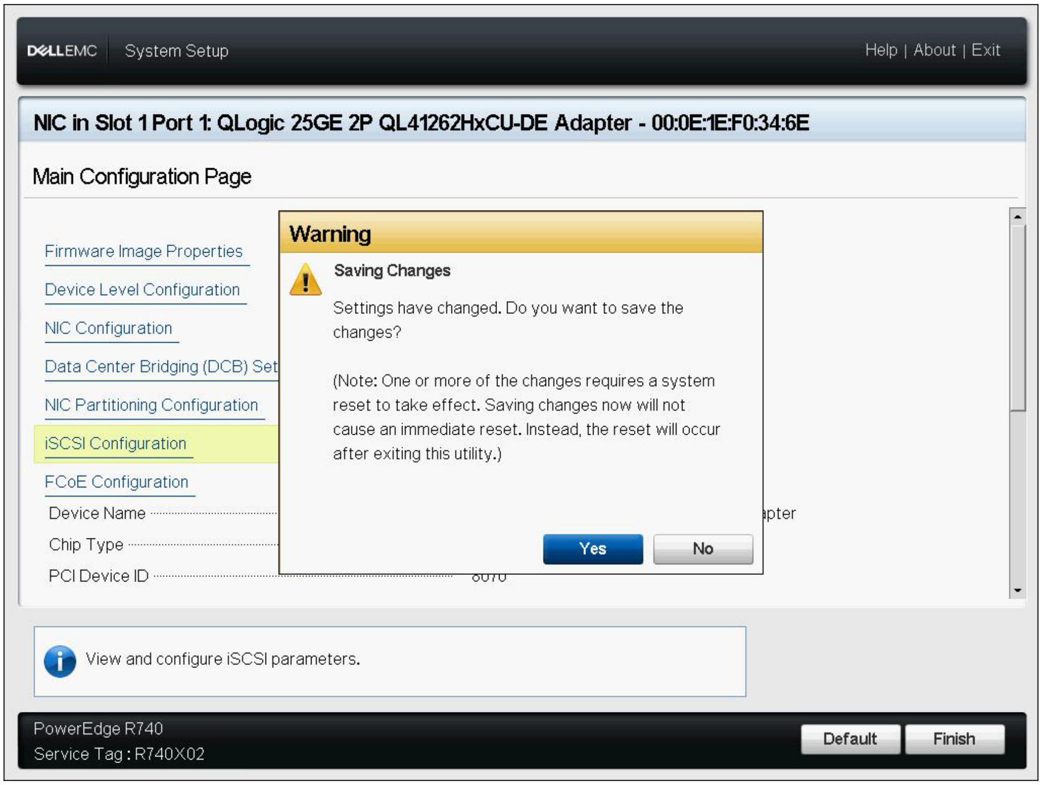
iSCSI General Parameters
iSCSI Initiator Parameters
iSCSI First Target Parameters
iSCSI Second Target Parameters

Enable connection and configure communication parameters for the second iSCSI target.

PowerEdge R740
Service Tag : R740X02

Figure 9-13. System Setup: iSCSI Second Target Parameters

13. Press ESC once, and a second time to exit.
14. Click **Yes** to save changes, or follow the OEM guidelines to save the device-level configuration. For example, click **Yes** to confirm the setting change (Figure 9-14).


DELLEMC
System Setup
Help | About | Exit
Firmware Image Properties
Device Level Configuration
NIC Configuration
Data Center Bridging (DCB) Set
NIC Partitioning Configuration
iSCSI Configuration
FCoE Configuration
Yes
No
Default
Finish

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page

Warning

Saving Changes

Settings have changed. Do you want to save the changes?

(Note: One or more of the changes requires a system reset to take effect. Saving changes now will not cause an immediate reset. Instead, the reset will occur after exiting this utility.)

Device Name .... pter

Chip Type ....

PCI Device ID 8070

View and configure iSCSI parameters.

PowerEdge R740
Service Tag : R740X02

Figure 9-14. System Setup: Saving iSCSI Changes

15. After all changes have been made, reboot the system to apply the changes to the adapter's running configuration.

Dynamic iSCSI Boot Configuration

In a dynamic configuration, ensure that the system's IP address and target (or initiator) information are provided by a DHCP server (see IPv4 and IPv6 configurations in "Configuring the DHCP Server to Support iSCSI Boot" on page 140).

Any settings for the following parameters are ignored and do not need to be cleared (with the exception of the initiator iSCSI name for IPv4, CHAP ID, and CHAP secret for IPv6):

- Initiator Parameters
- First Target Parameters or Second Target Parameters

For information on configuration options, see Table 9-1 on page 125.

NOTE

When using a DHCP server, the DNS server entries are overwritten by the values provided by the DHCP server. This override occurs even if the locally provided values are valid and the DHCP server provides no DNS server information. When the DHCP server provides no DNS server information, both the primary and secondary DNS server values are set to `0.0.0.0`. When the Windows OS takes over, the Microsoft iSCSI initiator retrieves the iSCSI initiator parameters and statically configures the appropriate registries. It will overwrite whatever is configured. Because the DHCP daemon runs in the Windows environment as a user process, all TCP/IP parameters must be statically configured before the stack comes up in the iSCSI boot environment.

If DHCP Option 17 is used, the target information is provided by the DHCP server, and the initiator iSCSI name is retrieved from the value programmed from the Initiator Parameters window. If no value was selected, the controller defaults to the following name:

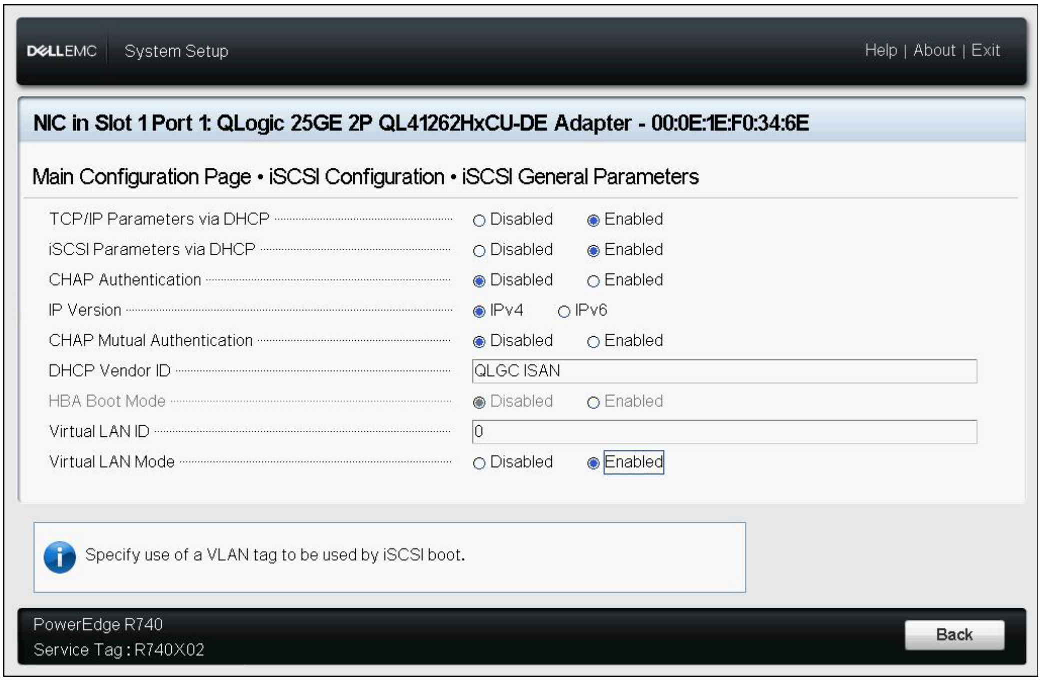
```
iqn.1995-05.com.qlogic.<11.22.33.44.55.66>.iscsiboot
```

The string `11.22.33.44.55.66` corresponds to the controller's MAC address. If DHCP Option 43 (IPv4 only) is used, any settings on the following windows are ignored and do not need to be cleared:

- Initiator Parameters
- First Target Parameters, or Second Target Parameters

To configure the iSCSI boot parameters using dynamic configuration:

- On the iSCSI General Parameters page, set the following options, as shown in Figure 9-15:
  - ❑ **TCP/IP Parameters via DHCP**: Enabled
  - ❑ **iSCSI Parameters via DHCP**: Enabled
  - ❑ **CHAP Authentication**: As required
  - ❑ **IP Version**: As required (IPv4 or IPv6)
  - ❑ **CHAP Mutual Authentication**: As required
  - ❑ **DHCP Vendor ID**: As required
  - ❑ **HBA Boot Mode**: Disabled
  - ❑ **Virtual LAN ID**: As required
  - ❑ **Virtual LAN Boot Mode**: Enabled



NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page • iSCSI Configuration • iSCSI General Parameters

TCP/IP Parameters via DHCP	○ Disabled	● Enabled
iSCSI Parameters via DHCP	○ Disabled	● Enabled
CHAP Authentication	● Disabled	○ Enabled
IP Version	● IPv4	○ IPv6
CHAP Mutual Authentication	● Disabled	○ Enabled
DHCP Vendor ID	QLGC ISAN	
HBA Boot Mode	● Disabled	○ Enabled
Virtual LAN ID	0	
Virtual LAN Mode	○ Disabled	● Enabled

Specify use of a VLAN tag to be used by iSCSI boot.

PowerEdge R740
Service Tag : R740X02

Figure 9-15. System Setup: iSCSI General Parameters

Enabling CHAP Authentication

Ensure that the CHAP authentication is enabled on the target.

To enable CHAP authentication:

1. Go to the iSCSI General Parameters page.
2. Set **CHAP Authentication** to **Enabled**.
3. In the Initiator Parameters window, type values for the following:
   - ❑ **CHAP ID** (up to 255 characters)
   - ❑ **CHAP Secret** (if authentication is required; must be 12 to 16 characters in length)
4. Press ESC to return to the iSCSI Boot Configuration page.
5. On the **iSCSI Boot Configuration Menu**, select **iSCSI First Target Parameters**.
6. In the iSCSI First Target Parameters window, type values used when configuring the iSCSI target:
   - ❑ **CHAP ID** (optional if two-way CHAP)
   - ❑ **CHAP Secret** (optional if two-way CHAP; must be 12 to 16 characters in length or longer)

7. Press ESC to return to the iSCSI Boot Configuration Menu.
8. Press ESC, and then select confirm **Save Configuration**.

Configuring the DHCP Server to Support iSCSI Boot

The DHCP server is an optional component, and is only necessary if you will be doing a dynamic iSCSI boot configuration setup (see “Dynamic iSCSI Boot Configuration” on page 137).

Configuring the DHCP server to support iSCSI boot differs for IPv4 and IPv6:

- DHCP iSCSI Boot Configurations for IPv4
- Configuring the DHCP Server
- Configuring DHCP iSCSI Boot for IPv6
- Configuring VLANs for iSCSI Boot

DHCP iSCSI Boot Configurations for IPv4

DHCP includes several options that provide configuration information to the DHCP client. For iSCSI boot, Cavium QLogic adapters support the following DHCP configurations:

- DHCP Option 17, Root Path
- DHCP Option 43, Vendor-specific Information

DHCP Option 17, Root Path

Option 17 is used to pass the iSCSI target information to the iSCSI client.

The format of the root path, as defined in IETC RFC 4173, is:

```
"iscsi:"<servername>":"<protocol>":"<port>":"<LUN>":"<targetname>"
```

Table 9-2 lists the DHCP Option 17 parameters.

Table 9-2. DHCP Option 17 Parameter Definitions

Parameter	Definition
`"iscsi:"`	A literal string
`<servername>`	IP address or fully qualified domain name (FQDN) of the iSCSI target
`":"`	Separator
`<protocol>`	IP protocol used to access the iSCSI target. Because only TCP is currently supported, the protocol is 6.

Table 9-2. DHCP Option 17 Parameter Definitions (Continued)

Parameter	Definition
`<port>`	Port number associated with the protocol. The standard port number for iSCSI is 3260.
`<LUN>`	Logical unit number to use on the iSCSI target. The value of the LUN must be represented in hexadecimal format. A LUN with an ID of 64 must be configured as 40 within the Option 17 parameter on the DHCP server.
`<targetname>`	Target name in either IQN or EUI format. For details on both IQN and EUI formats, refer to RFC 3720. An example IQN name is `iqn.1995-05.com.QLogic:iscsi-target`.

DHCP Option 43, Vendor-specific Information

DHCP Option 43 (vendor-specific information) provides more configuration options to the iSCSI client than does DHCP Option 17. In this configuration, three additional sub-options are provided that assign the initiator IQN to the iSCSI boot client, along with two iSCSI target IQNs that can be used for booting. The format for the iSCSI target IQN is the same as that of DHCP Option 17, while the iSCSI initiator IQN is simply the initiator’s IQN.

NOTE

DHCP Option 43 is supported on IPv4 only.

Table 9-3 lists the DHCP Option 43 sub-options.

Table 9-3. DHCP Option 43 Sub-option Definitions

Sub-option	Definition
201	First iSCSI target information in the standard root path format: `"iscsi:"<servername>":"<protocol>":"<port>":"<LUN>":"<targetname>"`
202	Second iSCSI target information in the standard root path format: `"iscsi:”<servername>":"<protocol>":"<port>":"<LUN>":"<targetname>"`
203	iSCSI initiator IQN

Using DHCP Option 43 requires more configuration than DHCP Option 17, but it provides a richer environment and more configuration options. You should use DHCP Option 43 when performing dynamic iSCSI boot configuration.

Configuring the DHCP Server

Configure the DHCP server to support either Option 16, 17, or 43.

NOTE

The format of DHCPv6 Option 16 and Option 17 are fully defined in RFC 3315.

If you use Option 43, you must also configure Option 60. The value of Option 60 must match the DHCP Vendor ID value, QLGC ISAN, as shown in **iSCSI General Parameters** of the iSCSI Boot Configuration page.

Configuring DHCP iSCSI Boot for IPv6

The DHCPv6 server can provide several options, including stateless or stateful IP configuration, as well as information for the DHCPv6 client. For iSCSI boot, Cavium QLogic adapters support the following DHCP configurations:

- DHCPv6 Option 16, Vendor Class Option
- DHCPv6 Option 17, Vendor-Specific Information

NOTE

The DHCPv6 standard Root Path option is not yet available. Cavium suggests using Option 16 or Option 17 for dynamic iSCSI boot IPv6 support.

DHCPv6 Option 16, Vendor Class Option

DHCPv6 Option 16 (vendor class option) must be present and must contain a string that matches your configured DHCP Vendor ID parameter. The DHCP Vendor ID value is QLGC ISAN, as shown in the **General Parameters** of the iSCSI Boot Configuration menu.

The content of Option 16 should be `<2-byte length> <DHCP Vendor ID>`.

DHCPv6 Option 17, Vendor-Specific Information

DHCPv6 Option 17 (vendor-specific information) provides more configuration options to the iSCSI client. In this configuration, three additional sub-options are provided that assign the initiator IQN to the iSCSI boot client, along with two iSCSI target IQNs that can be used for booting.

Table 9-4 lists the DHCP Option 17 sub-options.

Table 9-4. DHCP Option 17 Sub-option Definitions

Sub-option	Definition
201	First iSCSI target information in the standard root path format: `"iscsi:"[<servername>]":"<protocol>":"<port>":"<LUN> ": "<targetname>"`
202	Second iSCSI target information in the standard root path format: `"iscsi:"[<servername>]":"<protocol>":"<port>":"<LUN> ": "<targetname>"`
203	iSCSI initiator IQN

Table Notes:
Brackets [ ] are required for the IPv6 addresses.

The format of Option 17 should be:

```
<2-byte Option Number 201|202|203> <2-byte length> <data>
```

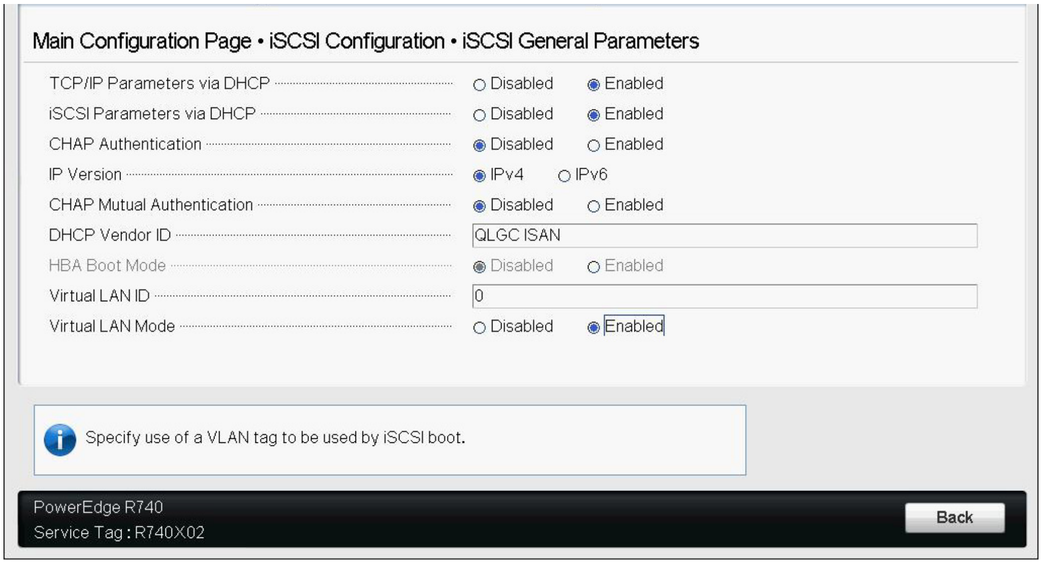
Configuring VLANs for iSCSI Boot

iSCSI traffic on the network may be isolated in a Layer 2 VLAN to segregate it from general traffic. If this is the case, make the iSCSI interface on the adapter a member of that VLAN.

To configure VLAN for iSCSI boot:

1. Go to the **iSCSI Configuration Page** for the port.
2. Select **iSCSI General Parameters**.

3. Select **VLAN ID** to enter and set the VLAN value, as shown in Figure 9-16.



Main Configuration Page • iSCSI Configuration • iSCSI General Parameters

TCP/IP Parameters via DHCP	○ Disabled ● Enabled
iSCSI Parameters via DHCP	○ Disabled ● Enabled
CHAP Authentication	● Disabled ○ Enabled
IP Version	● IPv4 ○ IPv6
CHAP Mutual Authentication	● Disabled ○ Enabled
DHCP Vendor ID	QLGC ISAN
HBA Boot Mode	● Disabled ○ Enabled
Virtual LAN ID	0
Virtual LAN Mode	○ Disabled ● Enabled

Specify use of a VLAN tag to be used by iSCSI boot.

Figure 9-16. System Setup: iSCSI General Parameters, VLAN ID

iSCSI Offload in Windows Server

iSCSI offload is a technology that offloads iSCSI protocol processing overhead from host processors to the iSCSI HBA. iSCSI offload increases network performance and throughput while helping to optimize server processor use. This section covers how to configure the Windows iSCSI offload feature for the Cavium 41*xxx* Series Adapters.

With the proper iSCSI offload licensing, you can configure your iSCSI-capable 41*xxx* Series Adapter to offload iSCSI processing from the host processor. The following sections describe how to enable the system to take advantage of Cavium’s iSCSI offload feature:

- Installing Cavium QLogic Drivers
- Installing the Microsoft iSCSI Initiator
- Configuring Microsoft Initiator to Use Cavium’s iSCSI Offload
- iSCSI Offload FAQs
- Windows Server 2012 R2 and 2016 iSCSI Boot Installation
- iSCSI Crash Dump

Installing Cavium QLogic Drivers

Install the Windows drivers as described in “Installing Windows Driver Software” on page 17.

Installing the Microsoft iSCSI Initiator

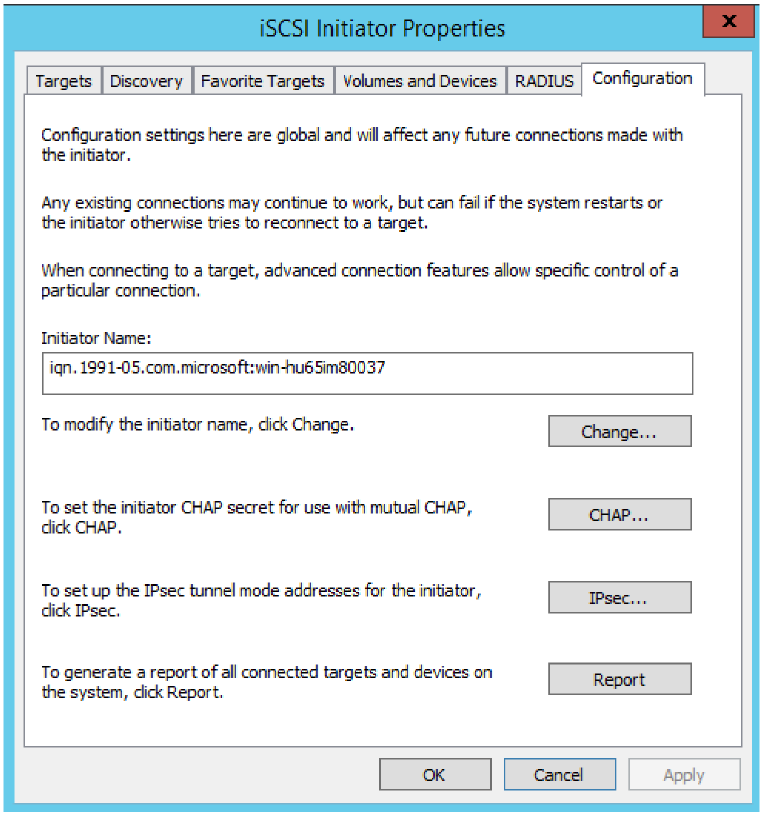
Launch the Microsoft iSCSI initiator applet. At the first launch, the system prompts for an automatic service start. Confirm the selection for the applet to launch.

Configuring Microsoft Initiator to Use Cavium’s iSCSI Offload

After the IP address is configured for the iSCSI adapter, you must use Microsoft Initiator to configure and add a connection to the iSCSI target using the Cavium QLogic iSCSI adapter. For more details on Microsoft Initiator, see the Microsoft user guide.

To configure Microsoft Initiator:

1. Open Microsoft Initiator.
2. To configure the initiator IQN name according to your setup, follow these steps:
   a. On the iSCSI Initiator Properties, click the **Configuration** tab.
   b. On the Configuration page (Figure 9-17), click **Change** next to **To modify the initiator name**.



Configuration settings here are global and will affect any future connections made with the initiator.

Any existing connections may continue to work, but can fail if the system restarts or the initiator otherwise tries to reconnect to a target.

When connecting to a target, advanced connection features allow specific control of a particular connection.

Initiator Name:

iqn.1991-05.com.microsoft:win-hu65im80037

To modify the initiator name, click Change.

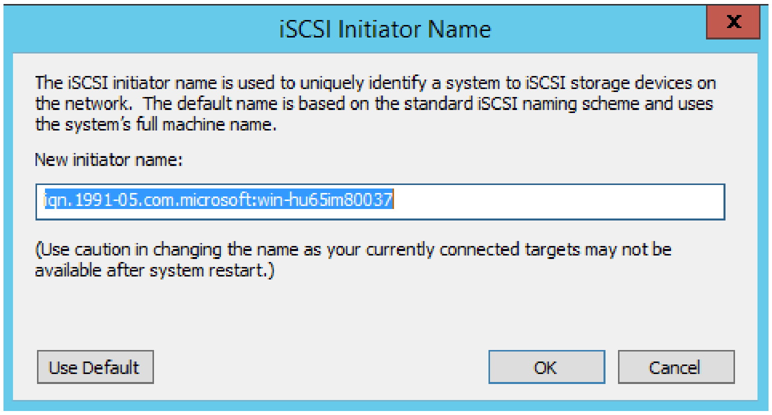
To set the initiator CHAP secret for use with mutual CHAP, click CHAP.

To set up the IPsec tunnel mode addresses for the initiator, click IPsec.

To generate a report of all connected targets and devices on the system, click Report.

Figure 9-17. iSCSI Initiator Properties, Configuration Page

c. In the iSCSI Initiator Name dialog box, type the new initiator IQN name, and then click **OK**. (Figure 9-18)



iSCSI Initiator Name

The iSCSI initiator name is used to uniquely identify a system to iSCSI storage devices on the network. The default name is based on the standard iSCSI naming scheme and uses the system's full machine name.

New initiator name:

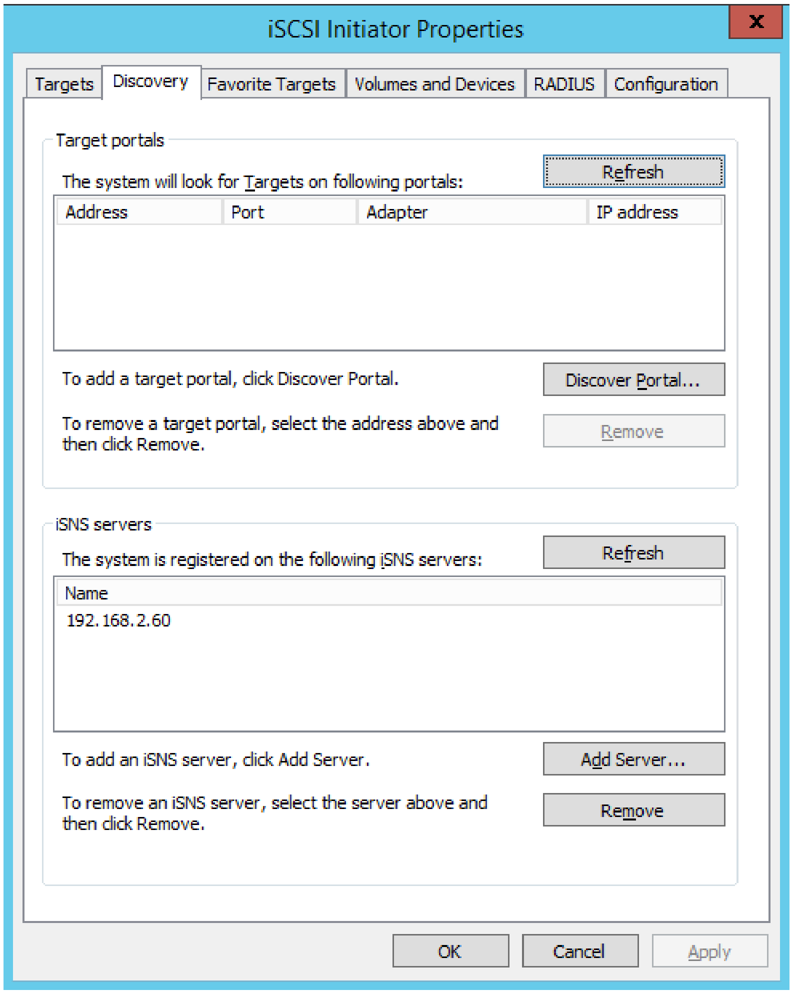
iqn.1991-05.com.microsoft:win-hu65im80037

(Use caution in changing the name as your currently connected targets may not be available after system restart.)

Figure 9-18. iSCSI Initiator Node Name Change

3. On the iSCSI Initiator Properties, click the **Discovery** tab.

4. On the Discovery page (Figure 9-19) under **Target portals**, click **Discover Portal**.



The system will look for Targets on following portals:

Address	Port	Adapter	IP address

To add a target portal, click Discover Portal.

To remove a target portal, select the address above and then click Remove.

The system is registered on the following iSNS servers:

Name
192.168.2.60

To add an iSNS server, click Add Server.

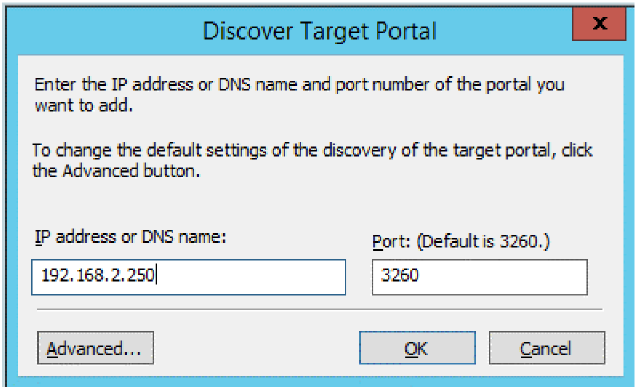
To remove an iSNS server, select the server above and then click Remove.

Figure 9-19. iSCSI Initiator—Discover Target Portal

5. In the Discover Target Portal dialog box (Figure 9-20):

   a. In the **IP address or DNS name** box, type the IP address of the target.

   b. Click **Advanced**.



Enter the IP address or DNS name and port number of the portal you want to add.

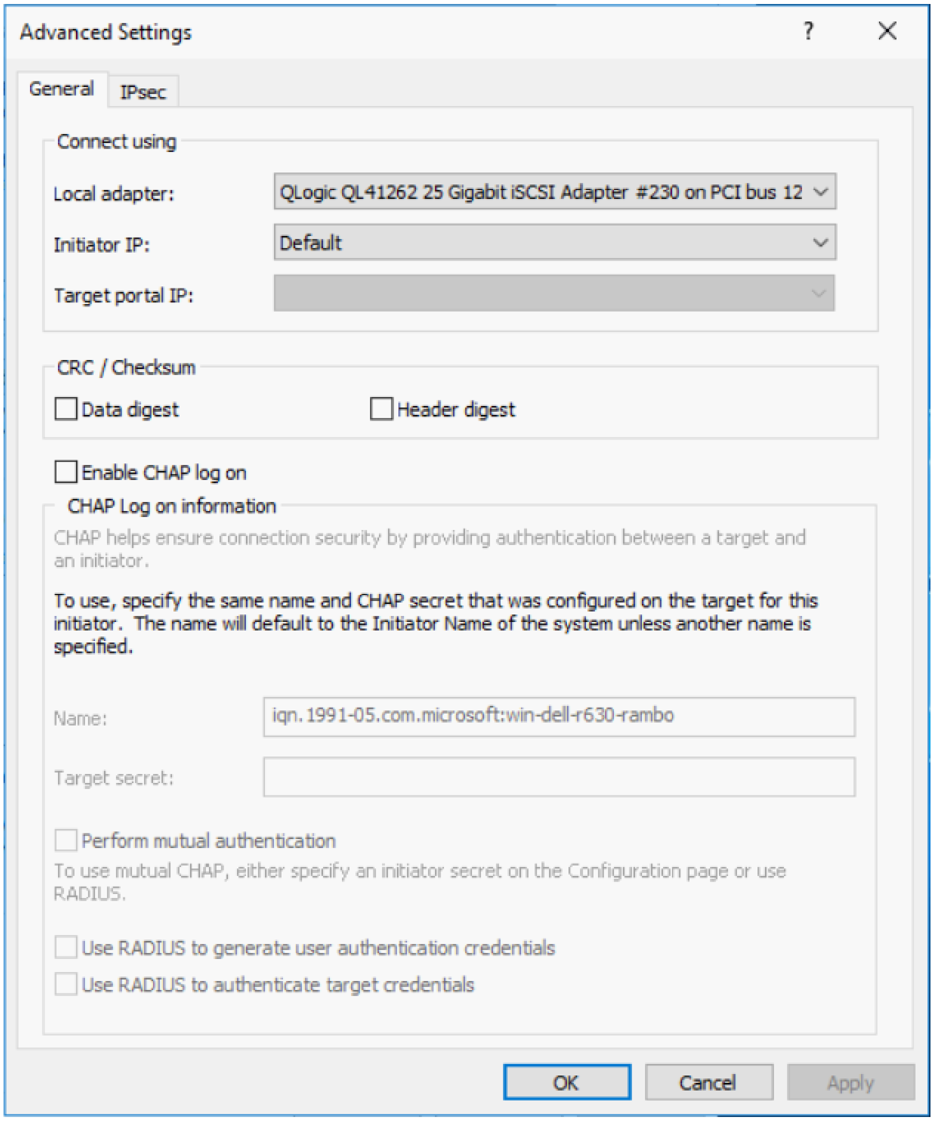
To change the default settings of the discovery of the target portal, click the Advanced button.

IP address or DNS name: 192.168.2.250

Port: (Default is 3260.) 3260

Figure 9-20. Target Portal IP Address

6. In the Advanced Settings dialog box (Figure 9-21), complete the following under **Connect using**:
   a. For **Local adapter**, select the **QLogic <name or model> Adapter**.
   b. For **Initiator IP**, select the adapter IP address.
   c. Click **OK**.



Advanced Settings

Connect using

Local adapter: QLogic QL41262 25 Gigabit iSCSI Adapter #230 on PCI bus 12

Initiator IP: Default

Target portal IP:

CRC / Checksum

□ Data digest □ Header digest

□ Enable CHAP log on

CHAP Log on information

CHAP helps ensure connection security by providing authentication between a target and an initiator.

To use, specify the same name and CHAP secret that was configured on the target for this initiator. The name will default to the Initiator Name of the system unless another name is specified.

Name: iqn.1991-05.com.microsoft:win-dell-r630-rambo

Target secret:

□ Perform mutual authentication

To use mutual CHAP, either specify an initiator secret on the Configuration page or use RADIUS.

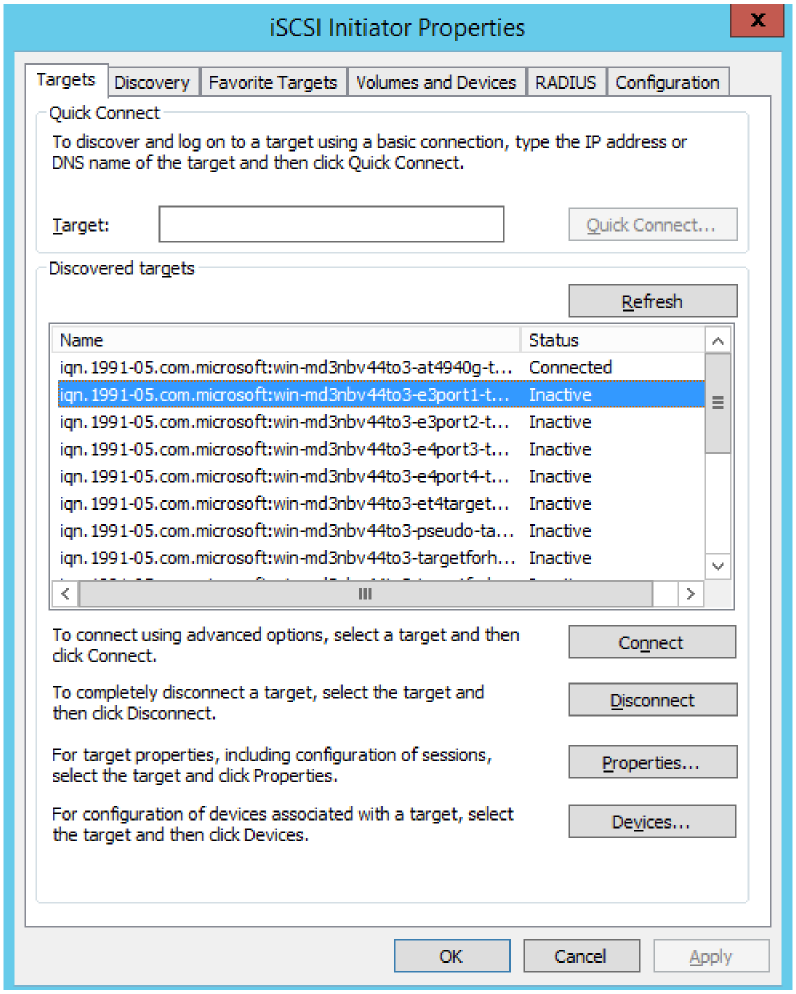
□ Use RADIUS to generate user authentication credentials

□ Use RADIUS to authenticate target credentials

Figure 9-21. Selecting the Initiator IP Address

7. On the iSCSI Initiator Properties, Discovery page, click **OK**.

8. Click the **Targets** tab, and then on the Targets page (Figure 9-22), click **Connect**.



To discover and log on to a target using a basic connection, type the IP address or DNS name of the target and then click Quick Connect.

Name	Status
iqn.1991-05.com.microsoft:win-md3nbv44to3-at4940g-t...	Connected
iqn.1991-05.com.microsoft:win-md3nbv44to3-e3port1-t...	Inactive
iqn.1991-05.com.microsoft:win-md3nbv44to3-e3port2-t...	Inactive
iqn.1991-05.com.microsoft:win-md3nbv44to3-e4port3-t...	Inactive
iqn.1991-05.com.microsoft:win-md3nbv44to3-e4port4-t...	Inactive
iqn.1991-05.com.microsoft:win-md3nbv44to3-et4target...	Inactive
iqn.1991-05.com.microsoft:win-md3nbv44to3-pseudo-ta...	Inactive
iqn.1991-05.com.microsoft:win-md3nbv44to3-targetforh...	Inactive

To connect using advanced options, select a target and then click Connect.

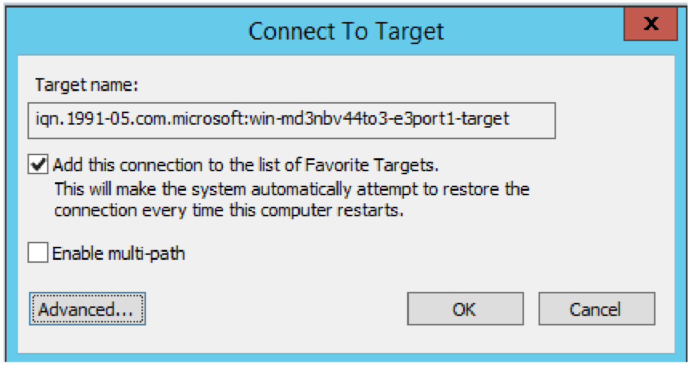
To completely disconnect a target, select the target and then click Disconnect.

For target properties, including configuration of sessions, select the target and click Properties.

For configuration of devices associated with a target, select the target and then click Devices.

Figure 9-22. Connecting to the iSCSI Target

9. On the Connect To Target dialog box (Figure 9-23), click **Advanced**.



Connect To Target

Target name:

iqn.1991-05.com.microsoft:win-md3nbv44to3-e3port1-target

☑ Add this connection to the list of Favorite Targets.
This will make the system automatically attempt to restore the connection every time this computer restarts.

□ Enable multi-path

Figure 9-23. Connect To Target Dialog Box

10. In the Local Adapter dialog box, select the **QLogic <name or model> Adapter**, and then click **OK**.
11. Click **OK** again to close Microsoft Initiator.
12. To format the iSCSI partition, use Disk Manager.

NOTE

Some limitations of the teaming functionality include:

- Teaming does not support iSCSI adapters.
- Teaming does not support NDIS adapters that are in the boot path.
- Teaming supports NDIS adapters that are not in the iSCSI boot path, but only for the SLB team type.

iSCSI Offload FAQs

Some of the frequently asked questions about iSCSI offload include:

Question: How do I assign an IP address for iSCSI offload?

Answer: Use the Configurations page in QConvergeConsole GUI.

Question: What tools should I use to create the connection to the target?

Answer: Use Microsoft iSCSI Software Initiator (version 2.08 or later).

Question: How do I know that the connection is offloaded?

Answer: Use Microsoft iSCSI Software Initiator. From a command line, type `oiscsicli sessionlist`. From **Initiator Name**, an iSCSI offloaded connection will display an entry beginning with `B06BDRV`. A non-offloaded connection displays an entry beginning with `Root`.

Question: What configurations should be avoided?

Answer: The IP address should not be the same as the LAN.

Windows Server 2012 R2 and 2016 iSCSI Boot Installation

Windows Server 2012 R2 and 2016 support booting and installing in either the offload or non-offload paths. Cavium requires that you use a slipstream DVD with the latest Cavium QLogic drivers injected. See "Injecting (Slipstreaming) Adapter Drivers into Windows Image Files" on page 184.

The following procedure prepares the image for installation and booting in either the offload or non-offload path.

To set up Windows Server 2012R2/2016 iSCSI boot:

1. Remove any local hard drives on the system to be booted (remote system).
2. Prepare the Windows OS installation media by following the slipstreaming steps in "Injecting (Slipstreaming) Adapter Drivers into Windows Image Files" on page 184.
3. Load the latest Cavium QLogic iSCSI boot images into the NVRAM of the adapter.
4. Configure the iSCSI target to allow a connection from the remote device. Ensure that the target has sufficient disk space to hold the new OS installation.
5. Configure the UEFI HII to set the iSCSI boot type (offload or non-offload), correct initiator, and target parameters for iSCSI boot.
6. Save the settings and reboot the system. The remote system should connect to the iSCSI target and then boot from the DVD-ROM device.
7. Boot from DVD and begin installation.
8. Follow the on-screen instructions.

   At the window that shows the list of disks available for the installation, the iSCSI target disk should be visible. This target is a disk connected through the iSCSI boot protocol and located in the remote iSCSI target.
9. To proceed with Windows Server 2012R2/2016 installation, click **Next**, and then follow the on-screen instructions. The server will undergo a reboot multiple times as part of the installation process.
10. After the server boots to the OS, you should run the driver installer to complete the Cavium QLogic drivers and application installation.

iSCSI Crash Dump

Crash dump functionality is supported for both non-offload and offload iSCSI boot for the 41*xxx* Series Adapters. No additional configurations are required to configure iSCSI crash dump generation.

iSCSI Offload in Linux Environments

The Cavium QLogic FastLinQ 41*xxx* iSCSI software consists of a single kernel module called `qedi.ko` (qedi). The qedi module is dependent on additional parts of the Linux kernel for specific functionality:

- **`qed.ko`** is the Linux eCore kernel module used for common Cavium FastLinQ 41*xxx* hardware initialization routines.
- **`scsi_transport_iscsi.ko`** is the Linux iSCSI transport library used for upcall and downcall for session management.
- **`libiscsi.ko`** is the Linux iSCSI library function needed for protocol data unit (PDU) and task processing, as well as session memory management.
- **`iscsi_boot_sysfs.ko`** is the Linux iSCSI sysfs interface that provides helpers to export iSCSI boot information.
- **`uio.ko`** is the Linux Userspace I/O interface, used for light L2 memory mapping for iscsiuio.

These modules must be loaded before qedi can be functional. Otherwise, you might encounter an “unresolved symbol” error. If the qedi module is installed in the distribution update path, the requisite is automatically loaded by modprobe.

This section provides the following information about iSCSI offload in Linux:

- Differences from bnx2i
- Configuring qedi.ko
- Verifying iSCSI Interfaces in Linux
- Configuring iSCSI Boot from SAN

Differences from bnx2i

Some key differences exist between qedi—the driver for the Cavium FastLinQ 41*xxx* Series Adapter (iSCSI)—and the previous QLogic iSCSI offload driver—bnx2i for the Cavium 8400 Series Adapters. Some of these differences include:

- qedi directly binds to a PCI function exposed by the CNA.
- qedi does not sit on top of the net_device.
- qedi is not dependent on a network driver such as bnx2x and cnic.
- qedi is not dependent on cnic, but it has dependency on qed.

- qedi is responsible for exporting boot information in sysfs using `iscsi_boot_sysfs.ko`, whereas bnx2i boot from SAN relies on the `iscsi_ibft.ko` module for exporting boot information.

Configuring qedi.ko

The qedi driver automatically binds to the exposed iSCSI functions of the CNA, and the target discovery and binding is done through the Open-iSCSI tools. This functionality and operation is similar to that of the bnx2i driver.

To load the `qedi.ko` kernel module, issue the following commands:

```
# modprobe qed
# modprobe libiscsi
# modprobe uio
# modprobe iscsi_boot_sysfs
# modprobe qedi
```

Verifying iSCSI Interfaces in Linux

After installing and loading the qedi kernel module, you must verify that the iSCSI interfaces were detected correctly.

To verify iSCSI interfaces in Linux:

1. To verify that the qedi and associated kernel modules are actively loaded, issue the following command:

```
# lsmod | grep qedi
qedi                    114578   2
qed                     697989   1 qedi
uio                     19259    4 cnic,qedi
libiscsi                57233    2 qedi,bnx2i
scsi_transport_iscsi    99909    5 qedi,bnx2i,libiscsi
iscsi_boot_sysfs        16000    1 qedi
```

2. To verify that the iSCSI interfaces were detected properly, issue the following command. In this example, two iSCSI CNA devices are detected with SCSI host numbers 4 and 5.

```
# dmesg | grep qedi
[0000:00:00.0]:[qedi_init:3696]: QLogic iSCSI Offload Driver v8.15.6.0.
....
[0000:42:00.4]:[__qedi_probe:3563]:59: QLogic FastLinQ iSCSI Module qedi 8.15.6.0, FW 8.15.3.0
....
[0000:42:00.4]:[qedi_link_update:928]:59: Link Up event.
....
[0000:42:00.5]:[__qedi_probe:3563]:60: QLogic FastLinQ iSCSI Module qedi 8.15.6.0, FW 8.15.3.0
```

```
....
[0000:42:00.5]:[qedi_link_update:928]:59: Link Up event
```

3. Use Open-iSCSI tools to verify that IP is configured properly. Issue the following command:

```
# iscsiadm -m iface | grep qedi
qedi.00:0e:1e:c4:e1:6d
qedi,00:0e:1e:c4:e1:6d,192.168.101.227,<empty>,iqn.1994-05.com.redhat:534ca9b6adf
qedi.00:0e:1e:c4:e1:6c
qedi,00:0e:1e:c4:e1:6c,192.168.25.91,<empty>,iqn.1994-05.com.redhat:534ca9b6adf
```

4. To ensure that the iscsiuio service is running, issue the following command:

```
# systemctl status iscsiuio.service
iscsiuio.service - iSCSI UserSpace I/O driver
Loaded: loaded (/usr/lib/systemd/system/iscsiuio.service; disabled; vendor preset: disabled)
Active: active (running) since Fri 2017-01-27 16:33:58 IST; 6 days ago
Docs: man:iscsiuio(8)
Process: 3745  ExecStart=/usr/sbin/iscsiuio (code=exited, status=0/SUCCESS)
    Main PID: 3747 (iscsiuio)
    CGroup: /system.slice/iscsiuio.service !--3747 /usr/sbin/iscsiuio
    Jan 27 16:33:58 localhost.localdomain systemd[1]: Starting iSCSI
UserSpace I/O driver...
    Jan 27 16:33:58 localhost.localdomain systemd[1]: Started iSCSI UserSpace I/O driver.
```

5. To discover the iSCSI target, issue the `iscsiadm` command:

```
#iscsiadm -m discovery -t st -p 192.168.25.100 -I qedi.00:0e:1e:c4:e1:6c
192.168.25.100:3260,1 iqn.2003-
04.com.sanblaze:virtualun.virtualun.target-05000007
192.168.25.100:3260,1 iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000012
192.168.25.100:3260,1 iqn.2003-04.com.sanblaze:virtualun.virtualun.target-0500000c
192.168.25.100:3260,1 iqn.2003-
04.com.sanblaze:virtualun.virtualun.target-05000001
192.168.25.100:3260,1 iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000002
```

6. Log into the iSCSI target using the IQN obtained in Step 5. To initiate the login procedure, issue the following command (where the last character in the command is a lowercase letter “L”):

```
#iscsiadm -m node -p 192.168.25.100 -T
iqn.2003-04.com.sanblaze:virtualun.virtualun.target-0000007 -l
Logging in to [iface: qedi.00:0e:1e:c4:e1:6c,
target:iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000007, portal:192.168.25.100,3260]
(multiple)
Login to [iface: qedi.00:0e:1e:c4:e1:6c, target:iqn.2003-
04.com.sanblaze:virtualun.virtualun.target-05000007, portal:192.168.25.100,3260] successful.
```

7. To verify that the iSCSI session was created, issue the following command:

```
# iscsiadm -m session
qedi: [297] 192.168.25.100:3260,1
iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000007 (non-flash)
```

8. To check for iSCSI devices, issue the `iscsiadm` command:

```
# iscsiadm -m session -P3
...
************************
Attached SCSI devices:
************************
Host Number: 59 State: running
scsi59 Channel 00 Id 0 Lun: 0
Attached scsi disk sdb   State: running scsi59 Channel 00 Id 0 Lun: 1
Attached scsi disk sdc   State: running scsi59 Channel 00 Id 0 Lun: 2
Attached scsi disk sdd   State: running scsi59 Channel 00 Id 0 Lun: 3
Attached scsi disk sde   State: running scsi59 Channel 00 Id 0 Lun: 4
Attached scsi disk sdf   State: running
```

For advanced target configurations, refer to the Open-iSCSI README at:

https://github.com/open-iscsi/open-iscsi/blob/master/README

Configuring iSCSI Boot from SAN

This section provides iSCSI boot from SAN procedures for the following Linux distributions:

- Configuring iSCSI Boot from SAN for RHEL 7.4
- Configuring iSCSI Boot from SAN for RHEL 7.5
- Configuring iSCSI Boot from SAN for SLES 12 SP3 and Later
- Configuring iSCSI Boot from SAN for Other Linux Distributions

Configuring iSCSI Boot from SAN for RHEL 7.4

To install RHEL 7.4:

1. Boot from the RHEL 7.*x* installation media with the iSCSI target already connected in UEFI.

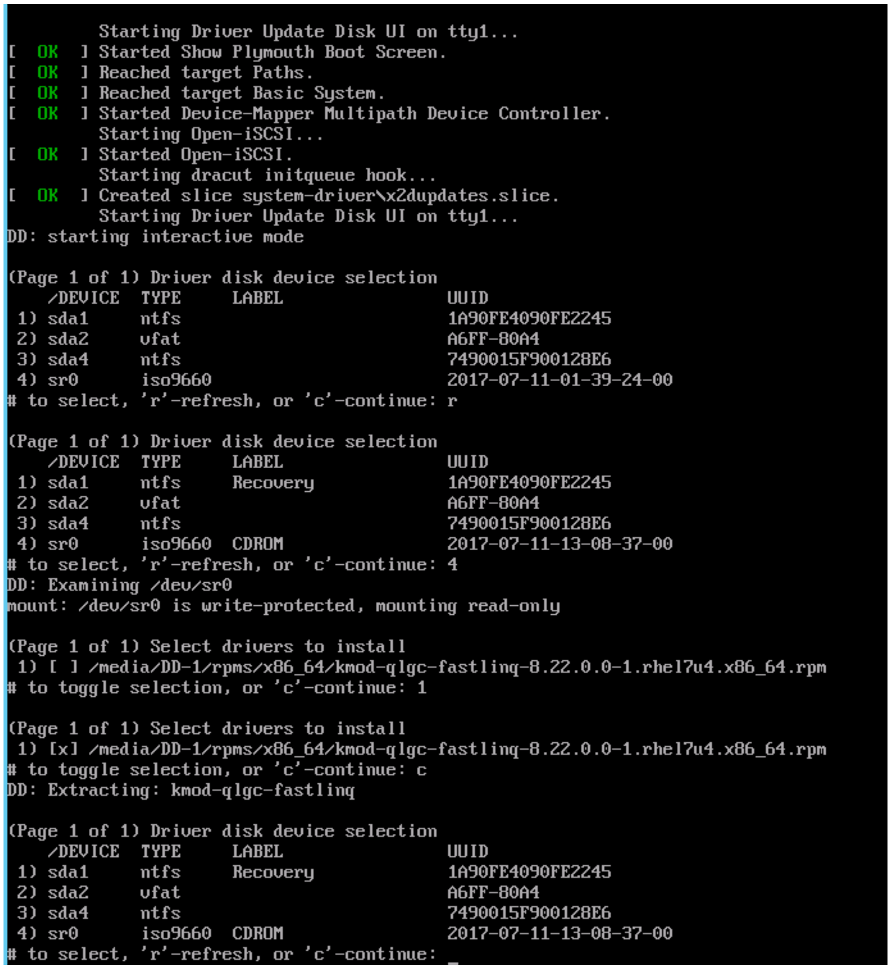
```
Install Red Hat Enterprise Linux 7.x
Test this media & install Red Hat Enterprise 7.x
Troubleshooting -->

Use the UP and DOWN keys to change the selection
```

```
Press 'e' to edit the selected item or 'c' for a command
prompt
```

2. To install an out-of-box driver, press the E key. Otherwise, proceed to Step 6.
3. Select the kernel line, and then press the E key.
4. Issue the following command, and then press ENTER.

   ```
   inst.dd modprobe.blacklist=qed,qede,qedr,qedi,qedf
   ```

   The installation process prompts you to install the out-of-box driver as shown in the Figure 9-24 example.



```
         Starting Driver Update Disk UI on tty1...
[  OK  ] Started Show Plymouth Boot Screen.
[  OK  ] Reached target Paths.
[  OK  ] Reached target Basic System.
[  OK  ] Started Device-Mapper Multipath Device Controller.
         Starting Open-iSCSI...
[  OK  ] Started Open-iSCSI.
         Starting dracut initqueue hook...
[  OK  ] Created slice system-driver\x2dupdates.slice.
         Starting Driver Update Disk UI on tty1...
DD: starting interactive mode

(Page 1 of 1) Driver disk device selection
    /DEVICE  TYPE     LABEL                UUID
 1) sda1     ntfs                          1A90FE4090FE2245
 2) sda2     vfat                          A6FF-80A4
 3) sda4     ntfs                          7490015F900128E6
 4) sr0      iso9660                       2017-07-11-01-39-24-00
# to select, 'r'-refresh, or 'c'-continue: r

(Page 1 of 1) Driver disk device selection
    /DEVICE  TYPE     LABEL                UUID
 1) sda1     ntfs     Recovery             1A90FE4090FE2245
 2) sda2     vfat                          A6FF-80A4
 3) sda4     ntfs                          7490015F900128E6
 4) sr0      iso9660  CDROM                2017-07-11-13-08-37-00
# to select, 'r'-refresh, or 'c'-continue: 4
DD: Examining /dev/sr0
mount: /dev/sr0 is write-protected, mounting read-only

(Page 1 of 1) Select drivers to install
 1) [ ] /media/DD-1/rpms/x86_64/kmod-qlgc-fastlinq-8.22.0.0-1.rhel7u4.x86_64.rpm
# to toggle selection, or 'c'-continue: 1

(Page 1 of 1) Select drivers to install
 1) [x] /media/DD-1/rpms/x86_64/kmod-qlgc-fastlinq-8.22.0.0-1.rhel7u4.x86_64.rpm
# to toggle selection, or 'c'-continue: c
DD: Extracting: kmod-qlgc-fastlinq

(Page 1 of 1) Driver disk device selection
    /DEVICE  TYPE     LABEL                UUID
 1) sda1     ntfs     Recovery             1A90FE4090FE2245
 2) sda2     vfat                          A6FF-80A4
 3) sda4     ntfs                          7490015F900128E6
 4) sr0      iso9660  CDROM                2017-07-11-13-08-37-00
# to select, 'r'-refresh, or 'c'-continue: _
```

Figure 9-24. Prompt for Out-of-Box Installation

5. If required for your setup, load the FastLinQ driver update disk when prompted for additional driver disks. Otherwise, if you have no other driver update disks to install, press the C key.

6. Continue with the installation. You can skip the media test. Click **Next** to continue with the installation.

7. In the Configuration window (Figure 9-25), select the language to use during the installation process, and then click **Continue**.

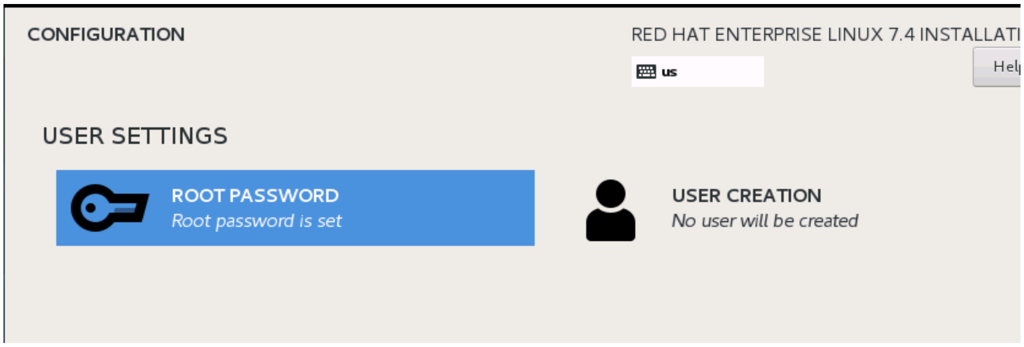


Figure 9-25. Red Hat Enterprise Linux 7.4 Configuration

8. In the Installation Summary window, click **Installation Destination**. The disk label is *sda*, indicating a single-path installation. If you configured multipath, the disk has a device mapper label.

9. In the **Specialized & Network Disks** section, select the iSCSI LUN.

10. Type the root user's password, and then click **Next** to complete the installation.

11. During the first boot, add the following kernel command line to fall into shell.

```
rd.iscsi.firmware rd.break=pre-pivot rd.driver.pre=qed,qede,qedr,qedf,qedi selinux=0
```

12. Issue the following commands:

```
# umount /sysroot/boot/efi
# umount /sysroot/boot/
# umount /sysroot/home/
# umount /sysroot
# mount /dev/mapper/rhel-root /sysroot
```

13. Open the `/sysroot/usr/libexec/iscsi-mark-root-nodes` file and locate the following statement:

```
if [ "$transport" = bnx2i ]; then
```

Change the statement to:

```
if [ "$transport" = bnx2i ] || [ "$transport" = qedi ]; then
```

14. Unmount file system by issuing the following command:

```
# umount /sysroot
```

15. Reboot the server, and then add the following parameters in the command line:

    `rd.iscsi.firmware`

    `rd.driver.pre=qed,qedi` (to load all drivers: `pre=qed,qedi,qede,qedf)`

    `selinux=0`

16. After a successful system boot, edit the `/etc/modprobe.d/anaconda-blacklist.conf` file to remove the blacklist entry for the selected driver.

17. Edit the `/etc/default/grub` file as follows:

    a. Locate the string in double quotes as shown in the following example. The command line is a specific reference to help find the string.

    ```
    GRUB_CMDLINE_LINUX="iscsi_firmware"
    ```

    b. The command line may contain other parameters that can remain. Change only the `iscsi_firmware` string as follows:

    ```
    GRUB_CMDLINE_LINUX="rd.iscsi.firmware selinux=0"
    ```

18. Create a new `grub.cfg` file by issuing the following command:

    ```
    # grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
    ```

19. Rebuild the ramdisk by issuing the **`dracut -f`** command, and then reboot.

NOTE

If the iSCSI target is on a different subnet than the initiator, update the `open-iscsi rpm` in the iSCSI BFS router setup:

1. Boot the RHEL 7.4 BFS with the same subnet.
2. Update the `open-iscsi rpm` by issuing the following commands:

   ```
   # rpm -ivh qlgc-open-iscsi-2.0_873.112.rhel7u4-1.x86_64.rpm
   -force
   # systemctl daemon-reload
   # systemctl restart iscsid
   ```

3. Verify the `iscsid` version as follows:

   ```
   Iscsid -version
   ```

4. Rebuild the `initrd` image as follows:

   ```
   dracut -f
   ```

5. Reboot, and then change the preboot IP to a different subnet.

Configuring iSCSI Boot from SAN for RHEL 7.5

To install RHEL 7.5:

1. Boot from the RHEL 7.*x* installation media with the iSCSI target already connected in UEFI.

   ```
   Install Red Hat Enterprise Linux 7.x
   Test this media & install Red Hat Enterprise 7.x
   Troubleshooting -->

   Use the UP and DOWN keys to change the selection
   Press 'e' to edit the selected item or 'c' for a command
   prompt
   ```

2. To install an out-of-box driver, press the E key. Otherwise, proceed to Step 6.
3. Select the kernel line, and then press the E key.
4. Issue the following command, and then press ENTER.

   ```
   inst.dd modprobe.blacklist=qed,qede,qedr,qedi,qedf
   ```

   The installation process prompts you to install the out-of-box driver.

5. If required for your setup, load the FastLinQ driver update disk when prompted for additional driver disks. Otherwise, if you have no other driver update disks to install, press the C key.
6. Continue with the installation. You can skip the media test. Click **Next** to continue with the installation.
7. In the Configuration window, select the language to use during the installation process, and then click **Continue**.
8. In the Installation Summary window, click **Installation Destination**. The disk label is *sda*, indicating a single-path installation. If you configured multipath, the disk has a device mapper label.
9. In the **Specialized & Network Disks** section, select the iSCSI LUN.
10. Type the root user's password, and then click **Next** to complete the installation.
11. Reboot the server, and then add the following parameters in the command line:

    `rd.iscsi.firmware`

    `rd.driver.pre=qed,qedi` (to load all drivers `pre=qed,qedi,qede,qedf)`

    `selinux=0`

12. After a successful system boot, edit the `/etc/modprobe.d/anaconda-blacklist.conf` file to remove the blacklist entry for the selected driver.

13. Edit the `/etc/default/grub` file as follows:

    a. Locate the string in double quotes as shown in the following example. The command line is a specific reference to help find the string.

    ```
    GRUB_CMDLINE_LINUX="iscsi_firmware"
    ```

    b. The command line may contain other parameters that can remain. Change only the `iscsi_firmware` string as follows:

    ```
    GRUB_CMDLINE_LINUX="rd.iscsi.firmware selinux=0"
    ```

14. Create a new `grub.cfg` file by issuing the following command:

    ```
    # grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
    ```

15. Rebuild the ramdisk by issuing the **`dracut -f`** command, and then reboot.

Configuring iSCSI Boot from SAN for SLES 12 SP3 and Later

To install SLES 12 SP3 and later:

1. Boot from the SLES 12 SP3 installation media with the iSCSI target pre-configured and connected in UEFI.

2. Update the latest driver package by adding the `dud=1` parameter in the installer command parameter. The driver update disk is required because the necessary iSCSI drivers are not inbox.

    NOTE

    For SLES 12 SP3 only: If the server is configured for Multi-Function mode (NPAR), you must provide the following additional parameters as part of this step:

    ```
    dud=1 brokenmodules=qed,qedi,qedf,qede withiscsi=1
    [BOOT_IMAGE=/boot/x86_64/loader/linux dud=1
    brokenmodules=qed,qedi,qedf,qede withiscsi=1]
    ```

3. Complete the installation steps specified by the SLES 12 SP3 OS.

Known Issue in DHCP Configuration

In DHCP configuration for SLES 12 SP3 and later, the first boot after OS installation fails if the initiator IP address acquired from the DHCP server is in different range than the target IP address. To resolve this issue, boot the OS using static configuration, update the latest iscsiuio out-of-box RPM, rebuild the initrd, and then reboot the OS using DHCP configuration. The OS should now boot successfully.

Configuring iSCSI Boot from SAN for Other Linux Distributions

For distributions such as RHEL 6.9/6.10/7.2/7.3, SLES 11 SP4, and SLES 12 SP1/2, the inbox iSCSI user space utility (Open-iSCSI tools) lacks support for qedi iSCSI transport and cannot perform user space-initiated iSCSI functionality. During boot from SAN installation, you can update the qedi driver using a driver update disk (DUD). However, no interface or process exists to update userspace inbox utilities, which causes the iSCSI target login and boot from SAN installation to fail.

To overcome this limitation, perform the initial boot from SAN with the pure L2 interface (do not use hardware-offloaded iSCSI) using the following procedure during the boot from SAN.

iSCSI offload for other distributions of Linux includes the following information:

- Booting from SAN Using a Software Initiator
- Migrating from Software iSCSI Installation to Offload iSCSI
- Linux Multipath Considerations

Booting from SAN Using a Software Initiator

To boot from SAN using a software initiator with Dell OEM Solutions:

1. Access the Dell EMC System BIOS settings.
2. Configure the initiator and target entries. For more information, refer to the Dell BIOS user guide.
3. At the beginning of the installation, pass the following boot parameter with the DUD option:
   - ❑ For RHEL 6.*x*, 7.*x*, and older:

     ```
     rd.iscsi.ibft dd
     ```

     No separate options are required for older distributions of RHEL.
   - ❑ For SLES 11 SP4 and SLES 12 SP1/SP2:

     ```
     ip=ibft dud=1
     ```

   - ❑ For the FastLinQ DUD package (for example, on RHEL 7):

     ```
     fastlinq-8.18.10.0-dd-rhel7u3-3.10.0_514.el7-x86_64.iso
     ```

     Where the DUD parameter is `dd` for RHEL 7.*x* and `dud=1` for SLES 12.*x*.
4. Install the OS on the target LUN.

Migrating from Software iSCSI Installation to Offload iSCSI

Migrate from the non-offload interface to an offload interface by following the instructions for your operating system.

- Migrating to Offload iSCSI for RHEL 6.9/6.10
- Migrating to Offload iSCSI for RHEL 7.2/7.3
- Migrating to Offload iSCSI for SLES 11 SP4
- Migrating to Offload iSCSI for SLES 12 SP1/SP2

Migrating to Offload iSCSI for RHEL 6.9/6.10

To migrate from a software iSCSI installation to an offload iSCSI for RHEL 6.9 or 6.10:

1. Boot into the iSCSI non-offload/L2 boot-from-SAN operating system. Issue the following commands to install the Open-iSCSI and iscsiuio RPMs:

```
# rpm -ivh --force qlgc-open-iscsi-2.0_873.111-1.x86_64.rpm
# rpm -ivh --force iscsiuio-2.11.5.2-1.rhel6u9.x86_64.rpm
```

NOTE

To retain the original contents of the inbox RPM during installation, you must use the `-ivh` option (instead of the `-Uvh` option), followed by the `--force` option.

2. Edit the `/etc/init.d/iscsid` file, add the following command, and then save the file:

```
modprobe -q qedi
```

For example:

```
echo -n $"Starting $prog: "
modprobe -q iscsi_tcp
modprobe -q ib_iser
modprobe -q cxgb3i
modprobe -q cxgb4i
modprobe -q bnx2i
modprobe -q be2iscsi
modprobe -q qedi
daemon iscsiuio
```

3. Open the `/boot/efi/EFI/redhat/grub.conf` file, make the following changes, and save the file:

   - ❑ Remove `ifname=eth5:14:02:ec:ce:dc:6d`
   - ❑ Remove `ip=ibft`
   - ❑ Add `selinux=0`

   For example:

   ```
   kernel /vmlinuz-2.6.32-696.el6.x86_64 ro
   root=/dev/mapper/vg_prebooteit-lv_root rd_NO_LUKS
   iscsi_firmware LANG=en_US.UTF-8 ~~ifname=eth5:14:02:ec:ce:dc:6d~~
   rd_NO_MD SYSFONT=latarcyrheb-sun16 crashkernel=auto rd_NO_DM
   rd_LVM_LV=vg_prebooteit/lv_swap ~~ip=ibft~~  KEYBOARDTYPE=pc
   KEYTABLE=us rd_LVM_LV=vg_prebooteit/lv_root rhgb quiet
           initrd /initramfs-2.6.32-696.el6.x86_64.img

   kernel /vmlinuz-2.6.32-696.el6.x86_64 ro
   root=/dev/mapper/vg_prebooteit-lv_root rd_NO_LUKS
   iscsi_firmware LANG=en_US.UTF-8 rd_NO_MD
   SYSFONT=latarcyrheb-sun16 crashkernel=auto rd_NO_DM
   rd_LVM_LV=vg_prebooteit/lv_swap KEYBOARDTYPE=pc KEYTABLE=us
   rd_LVM_LV=vg_prebooteit/lv_root selinux=0
           initrd /initramfs-2.6.32-696.el6.x86_64.img
   ```

4. Build the `initramfs` file by issuing the following command:

   ```
   # dracut -f
   ```

5. Reboot the server, and then open the HII.

6. In the HII, disable iSCSI boot from BIOS, and then enable iSCSI HBA or boot for the adapter as follows:

   a. Select the adapter port, and then select **Device Level Configuration**.

   b. On the Device Level Configuration page, for the **Virtualization Mode**, select **NPAR**.

   c. Open the NIC Partitioning Configuration page and set the **iSCSI Offload Mode** to **Enabled**. (iSCSI HBA support is on partition 3 for a two--port adapter and on partition 2 for a four-port adapter.)

   d. Open the **NIC Configuration** menu and set the **Boot Protocol** to **UEFI iSCSI**.

   e. Open the iSCSI Configuration page and configure iSCSI settings.

7. Save the configuration and reboot the server.

The OS can now boot through the offload interface.

Migrating to Offload iSCSI for RHEL 7.2/7.3

To migrate from a software iSCSI installation to an offload iSCSI for RHEL 7.2/7.3:

1. Update Open-iSCSI tools and iscsiuio by issuing the following commands:

```
#rpm -ivh qlgc-open-iscsi-2.0_873.111.rhel7u3-3.x86_64.rpm --force
#rpm -ivh iscsiuio-2.11.5.3-2.rhel7u3.x86_64.rpm --force
```

NOTE

To retain the original contents of the inbox RPM during installation, you must use the `-ivh` option (instead of the `-Uvh` option), followed by the `--force` option.

2. Reload all the daemon services by issuing the following command:

```
# systemctl daemon-reload
```

3. Restart iscsid and iscsiuio services by issuing the following commands:

```
# systemctl restart iscsiuio
# systemctl restart iscsid
```

4. Edit the `/usr/libexec/iscsi-mark-root-node` file and locate the following statement:

```
if [ "$transport" = bnx2i ]; then
start_iscsiuio=1
```

   Add `|| [ "$transport" = qedi ]` to the IF expression as follows:

```
if [ "$transport" = bnx2i ] || [ "$transport" = qedi ]; then
start_iscsiuio=1
```

5. Edit the `/usr/libexec/iscsi-mark-root-nodes` file as follows:

   a. Locate the following statement:

```
if [ "$transport" = bnx2i ]; then
```

   b. Change the statement to:

```
if [ "$transport" = bnx2i ] || [ "$transport" = qedi ]; then
```

6. Edit the `/etc/default/grub` file as follows:

   a. Locate the following statement:

```
GRUB_CMDLINE_LINUX="iscsi_firmware ip=ibft"
```

   b. Change this statement to:

      ```
      GRUB_CMDLINE_LINUX="rd.iscsi.firmware"
      ```

7. Create a new `grub.cfg` file by issuing the following command:

   ```
   # grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
   ```

8. Build the `initramfs` file by issuing the following command:

   ```
   # dracut -f
   ```

9. Reboot the server, and then open the HII.
10. In the HII, disable iSCSI boot from BIOS, and then enable iSCSI HBA or boot for the adapter as follows:
    a. Select the adapter port, and then select **Device Level Configuration**.
    b. On the Device Level Configuration page, for the **Virtualization Mode**, select **NPAR**.
    c. Open the NIC Partitioning Configuration page and set the **iSCSI Offload Mode** to **Enabled**. (iSCSI HBA support is on partition 3 for a two--port adapter and on partition 2 for a four-port adapter.)
    d. Open the **NIC Configuration** menu and set the **Boot Protocol** to **UEFI iSCSI**.
    e. Open the iSCSI Configuration page and configure iSCSI settings.
11. Save the configuration and reboot the server.

The OS can now boot through the offload interface.

Migrating to Offload iSCSI for SLES 11 SP4

To migrate from a software iSCSI installation to an offload iSCSI for SLES 11 SP4:

1. Update Open-iSCSI tools and iscsiuio to the latest available versions by issuing the following commands:

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.sles11sp4-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.3-2.sles11sp4.x86_64.rpm --force
```

NOTE

To retain the original contents of the inbox RPM during installation, you must use the `-ivh` option (instead of the `-Uvh` option), followed by the `--force` option.

2. Edit the `/etc/elilo.conf` file, make the following changes, and then save the file:

   ❑ Remove the `ip=ibft` parameter (if present)
   ❑ Add `iscsi_firmware`

3. Edit the `/etc/sysconfig/kernel` file as follows:

   a. Locate the line that begins with `INITRD_MODULES`. This line will look similar to the following, but may contain different parameters:

   ```
   INITRD_MODULES="ata_piix ata_generic"
   ```

   or

   ```
   INITRD_MODULES="ahci"
   ```

   b. Edit the line by adding `qedi` to the end of the existing line (inside the quotation marks). For example:

   ```
   INITRD_MODULES="ata_piix ata_generic qedi"
   ```

   or

   ```
   INITRD_MODULES="ahci qedi"
   ```

   c. Save the file.

4. Edit the `/etc/modprobe.d/unsupported-modules` file, change the value for `allow_unsupported_modules` to **`1`**, and then save the file:

   ```
   allow_unsupported_modules 1
   ```

5. Locate and delete the following files:

   ❑ `/etc/init.d/boot.d/K01boot.open-iscsi`
   ❑ `/etc/init.d/boot.open-iscsi`

6. Create a backup of initrd, and then build a new initrd by issuing the following commands.

   ```
   # cd /boot/
   # mkinitrd
   ```

7. Reboot the server, and then open the HII.

8. In the HII, disable iSCSI boot from BIOS, and then enable iSCSI HBA or boot for the adapter as follows:

   a. Select the adapter port, and then select **Device Level Configuration**.

   b. On the Device Level Configuration page, for the **Virtualization Mode**, select **NPAR**.

   c. Open the NIC Partitioning Configuration page and set the **iSCSI Offload Mode** to **Enabled**. (iSCSI HBA support is on partition 3 for a two--port adapter and on partition 2 for a four-port adapter.)

   d. Open the **NIC Configuration** menu and set the **Boot Protocol** to **UEFI iSCSI**.

   e. Open the iSCSI Configuration page and configure iSCSI settings.

9. Save the configuration and reboot the server.

The OS can now boot through the offload interface.

Migrating to Offload iSCSI for SLES 12 SP1/SP2

To migrate from a software iSCSI installation to an offload iSCSI for SLES 12 SP1/SP2:

1. Boot into the iSCSI non-offload/L2 boot-from-SAN operating system. Issue the following commands to install the Open-iSCSI and iscsiuio RPMs:

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.sles12p2-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.3-2.sles12sp2.x86_64.rpm --force
```

NOTE

To retain the original contents of the inbox RPM during installation, you must use the -ivh option (instead of the -Uvh option), followed by the --force option.

2. Reload all the daemon services by issuing the following command:

```
# systemctl daemon-reload
```

3. Enable iscsid and iscsiuio services if they are not already enabled by issuing the following commands:

```
# systemctl enable iscsid
# systemctl enable iscsiuio
```

4. Issue the following command:

```
cat /proc/cmdline
```

5. Check if the OS has preserved any boot options, such as ip=ibft or rd.iscsi.ibft.

   ❑ If there are preserved boot options, continue with Step 6.

   ❑ If there are no preserved boot options, skip to Step 6 c.

6. Edit the `/etc/default/grub` file and modify the `GRUB_CMDLINE_LINUX` value:
   a. Remove `rd.iscsi.ibft` (if present).
   b. Remove any `ip=<value>` boot options (if present).
   c. Add `rd.iscsi.firmware`. For older distros, add `iscsi_firmware`.
7. Create a backup of the original `grub.cfg` file. The file is in the following locations:
   ❑ Legacy boot: `/boot/grub2/grub.cfg`
   ❑ UEFI boot: `/boot/efi/EFI/sles/grub.cfg` for SLES
8. Create a new `grub.cfg` file by issuing the following command:

```
# grub2-mkconfig -o <new file name>
```

9. Compare the old `grub.cfg` file with the new `grub.cfg` file to verify your changes.
10. Replace the original `grub.cfg` file with the new `grub.cfg` file.
11. Build the `initramfs` file by issuing the following command:

```
# dracut -f
```

12. Reboot the server, and then open the HII.
13. In the HII, disable iSCSI boot from BIOS, and then enable iSCSI HBA or boot for the adapter as follows:
    a. Select the adapter port, and then select **Device Level Configuration**.
    b. On the Device Level Configuration page, for the **Virtualization Mode**, select **NPAR**.
    c. Open the NIC Partitioning Configuration page and set the **iSCSI Offload Mode** to **Enabled**. (iSCSI HBA support is on partition 3 for a two--port adapter and on partition 2 for a four-port adapter.)
    d. Open the **NIC Configuration** menu and set the **Boot Protocol** to **UEFI iSCSI**.
    e. Open the iSCSI Configuration page and configure iSCSI settings.
14. Save the configuration and reboot the server.

The OS can now boot through the offload interface.

Linux Multipath Considerations

iSCSI boot from SAN installations on Linux operating systems are currently supported only in a single-path configuration. To enable multipath configurations, perform the installation in a single path, using either an L2 or L4 path. After the server boots into the installed operating system, perform the required configurations for enabling multipath I/O (MPIO).

See the appropriate procedure in this section to migrate from L2 to L4 and configure MPIO for your OS:

- Migrating and Configuring MPIO to Offloaded Interface for RHEL 6.9/6.10
- Migrating and Configuring MPIO to Offloaded Interface for RHEL 7.2/7.3
- Migrating and Configuring MPIO to Offloaded Interface for SLES 11 SP4
- Migrating and Configuring MPIO to Offloaded Interface for SLES 12 SP1/SP2

Migrating and Configuring MPIO to Offloaded Interface for RHEL 6.9/6.10

To migrate from L2 to L4 and configure MPIO to boot the OS over an offloaded interface for RHEL 6.9/6.10:

1. Configure the iSCSI boot settings for L2 BFS on both ports of the adapter. The boot will log in through only one port, but will create an iSCSI Boot Firmware Table (iBFT) interface for both ports.
2. While booting to the CD, ensure that you specify the following kernel parameters:

   ```
   ip=ibft
   linux dd
   ```

3. Provide the DUD and complete the installation.
4. Boot to the OS with L2.
5. Update Open-iSCSI tools and iscsiuio by issuing the following commands:

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.rhel6u9-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.5-6.rhel6u9.x86_64.rpm --force
```

6. Edit the `/boot/efi/EFI/redhat/grub.conf` file, make the following, changes, and then save the file:
   a. Remove `ifname=eth5:14:02:ec:ce:dc:6d`.
   b. Remove `ip=ibft`.
   c. Add `selinux=0`.

7. Build the `initramfs` file by issuing the following command:

   ```
   # dracut -f
   ```

8. Reboot and change the adapter boot settings to use L4 or iSCSI (HW) for both ports and to boot through L4.

9. After booting into the OS, set up the multipath daemon `multipathd.conf`:

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

10. Start the multipathd service:

    ```
    # service multipathd start
    ```

11. Rebuild `initramfs` with multipath support.

    ```
    # dracut --force --add multipath --include /etc/multipath
    ```

12. Reboot the server and boot into OS with multipath.

NOTE

For any additional changes in the `/etc/multipath.conf` file to take effect, you must rebuild the initrd image and reboot the server.

Migrating and Configuring MPIO to Offloaded Interface for RHEL 7.2/7.3

To migrate from L2 to L4 and configure MPIO to boot the OS over an offloaded interface for RHEL 7.2/7.3:

1. Configure the iSCSI boot settings for L2 BFS on both ports of the adapter. The boot will log in through only one port, but will create an iSCSI Boot Firmware Table (iBFT) interface for both ports.

2. While booting to the CD, ensure that you specify the following kernel parameters:

   ```
   rd.iscsi.ibft
   dd
   ```

3. Provide the DUD and complete the installation.

4. Boot to the OS with L2.

5. Update Open-iSCSI tools and iscsiuio by issuing the following commands:

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.rhel7u3-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.5-6.rhel7u3.x86_64.rpm --force
```

6. Reload all the daemon services by issuing the following command:

   ```
   # systemctl daemon-reload
   ```

7. Restart iscsid and iscsiuio services by issuing the following commands:

   ```
   # systemctl restart iscsiuio
   # systemctl restart iscsid
   ```

8. Edit the `/usr/libexec/iscsi-mark-root-nodes` file and locate the following statement:

   ```
   if [ "$transport" = bnx2i ]; then
   ```

   Change the statement to:

   ```
   if [ "$transport" = bnx2i ] || [ "$transport" = qedi ]; then
   ```

9. Edit the `/etc/default/grub` file, and then locate the following statement:

   ```
   GRUB_CMDLINE_LINUX="iscsi_firmware ip=ibft"
   ```

   Change this statement to:

   ```
   GRUB_CMDLINE_LINUX="rd.iscsi.firmware"
   ```

10. Create a new `grub.cfg` file by issuing the following command:

    ```
    # grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
    ```

11. Build the `initramfs` file by issuing the following command:

    ```
    # dracut -f
    ```

12. Reboot and change the adapter boot settings to use L4 for both port and boot through L4.

13. After booting into OS, set up the multipath daemon `multipathd.conf`:

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

14. Start multipathd and verify that multipath.service is enabled.

    ```
    # systemctl start multipathd
    # systemctl status multipathd
    ```

15. Rebuild `initramfs` with multipath support.

    ```
    # dracut --force --add multipath --include /etc/multipath
    ```

16. Reboot the server and boot into OS with multipath.

NOTE

For any additional changes in the `/etc/multipath.conf` file to take effect, you must rebuild the initrd image and reboot the server.

Migrating and Configuring MPIO to Offloaded Interface for SLES 11 SP4

To migrate from L2 to L4 and configure MPIO to boot the OS over an offloaded interface for SLES 11 SP4:

1. Follow all the steps necessary to migrate non-offload (L2) interface to an offload (L4) interface, over a single path. See Migrating to Offload iSCSI for SLES 11 SP4.
2. After you boot into the OS using the L4 interface, prepare multipathing as follows:
   a. Reboot the server and open the HII by going to **System Setup/Utilities**.
   b. In the HII, select **System Configuration**, and select the second adapter port to be used for multipath.
   c. On the Main Configuration Page under **Port Level Configuration**, set **Boot Mode** to **iSCSI (HW)** and enable **iSCSI Offload**.
   d. On the Main Configuration Page under **iSCSI Configuration**, perform the necessary iSCSI configurations.
   e. Reboot the server and boot into OS.
3. Set MPIO services to remain persistent on re-boot as follows:

```
# chkconfig multipathd on
# chkconfig boot.multipath on
# chkconfig boot.udev on
```

4. Start multipath services as follows:

```
# /etc/init.d/boot.multipath start
# /etc/init.d/multipathd start
```

5. Run `multipath -v2 -d` to display multipath configuration with a dry run.
6. Locate the `multipath.conf` file under `/etc/multipath.conf`.

NOTE

If the file is not present, copy `multipath.conf` from the `/usr/share/doc/packages/multipath-tools` folder:

```
cp /usr/share/doc/packages/multipath-tools/multipath.
conf.defaults /etc/multipath.conf
```

7. Edit the `multipath.conf` to enable the default section.

8. Rebuild initrd image to include MPIO support:

   ```
   # mkinitrd -f multipath
   ```

9. Reboot the server and boot the OS with multipath support.

NOTE

For any additional changes in the `/etc/multipath.conf` file to take effect, you must rebuild the initrd image and reboot the server.

Migrating and Configuring MPIO to Offloaded Interface for SLES 12 SP1/SP2

To migrate from L2 to L4 and configure MPIO to boot the OS over an offloaded interface for SLES 12 SP1/SP2:

1. Configure the iSCSI boot settings for L2 BFS on both ports of the adapter. The boot will log in through only one port, but will create an iSCSI Boot Firmware Table (iBFT) interface for both ports.
2. While booting to the CD, ensure that you specify the following kernel parameters:

   ```
   dud=1
   rd.iscsi.ibft
   ```

3. Provide the DUD and complete the installation.
4. Boot to the OS with L2.
5. Update the Open-iSCSI tools by issuing the following commands:

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.sles12sp1-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.5-6.sles12sp1.x86_64.rpm --force
```

6. Edit the `/etc/default/grub` file by changing the `rd.iscsi.ibft` parameter to `rd.iscsi.firmware`, and then save the file:
7. Issue the following command:

   ```
   # grub2-mkconfig -o /boot/efi/EFI/suse/grub.cfg
   ```

8. To load the multipath module, issue the following command:

   ```
   # modprobe dm_multipath
   ```

9. To enable the multipath daemon, issue the following commands:

   ```
   # systemctl start multipathd.service
   # systemctl enable multipathd.service
   # systemctl start multipathd.socket
   ```

10. To run the multipath utility, issue the following commands:

    `# multipath` (may not show the multipath devices because it is booted with a single path on L2)

    ```
    # multipath -ll
    ```

11. To inject the multipath module in initrd, issue the following command:

    ```
    # dracut --force --add multipath --include /etc/multipath
    ```

12. Reboot the server and enter system settings by pressing the F9 key during the POST menu.

13. Change the UEFI configuration to use L4 iSCSI boot:

    a. Open System Configuration, select the adapter port, and then select **Port Level Configuration**.

    b. On the Port Level Configuration page, set the **Boot Mode** to **iSCSI (HW)** and set **iSCSI Offload** to **Enabled**.

    c. Save the settings, and then exit the System Configuration Menu.

14. Reboot the system. The OS should now boot through the offload interface.

NOTE

For any additional changes in the `/etc/multipath.conf` file to take effect, you must rebuild the initrd image and reboot the server.

10 FCoE Configuration

This chapter provides the following Fibre Channel over Ethernet (FCoE) configuration information:

- FCoE Boot from SAN
- "Injecting (Slipstreaming) Adapter Drivers into Windows Image Files" on page 184
- "Configuring Linux FCoE Offload" on page 185
- "VMware FCoE Boot from SAN" on page 194

NOTE

FCoE offload is supported on all 41*xxx* Series Adapters. Some FCoE features may not be fully enabled in the current release. For details, refer to Appendix D Feature Constraints.

FCoE Boot from SAN

This section describes the installation and boot procedures for the Windows, Linux, and ESXi operating systems, including:

- Preparing System BIOS for FCoE Build and Boot
- Windows FCoE Boot from SAN

NOTE

FCoE boot from SAN is supported on ESXi 5.5 and later. Not all adapter versions support FCoE and FCoE boot from SAN.

Preparing System BIOS for FCoE Build and Boot

To prepare the system BIOS, modify the system boot order and specify the BIOS boot protocol, if required.

Specifying the BIOS Boot Protocol

FCoE boot from SAN is supported in UEFI mode only. Set the platform in boot mode (protocol) using the system BIOS configuration to UEFI.

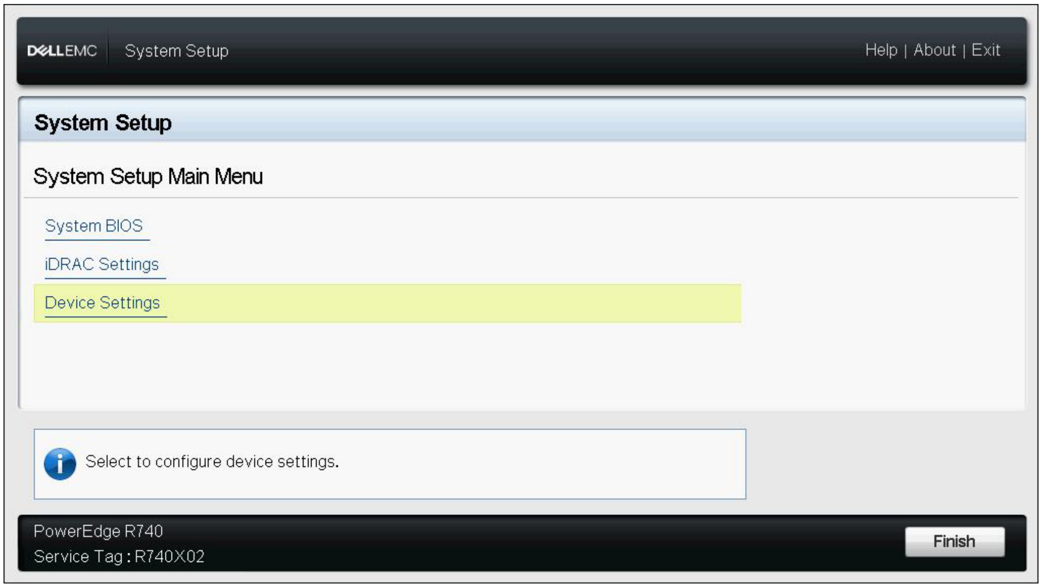
NOTE

FCoE BFS is not supported in legacy BIOS mode.

Configuring Adapter UEFI Boot Mode

To configure the boot mode to FCOE:

1. Restart the system.
2. Press the OEM hot key to enter System Setup (Figure 10-1). This is also known as UEFI HII.



Select to configure device settings.

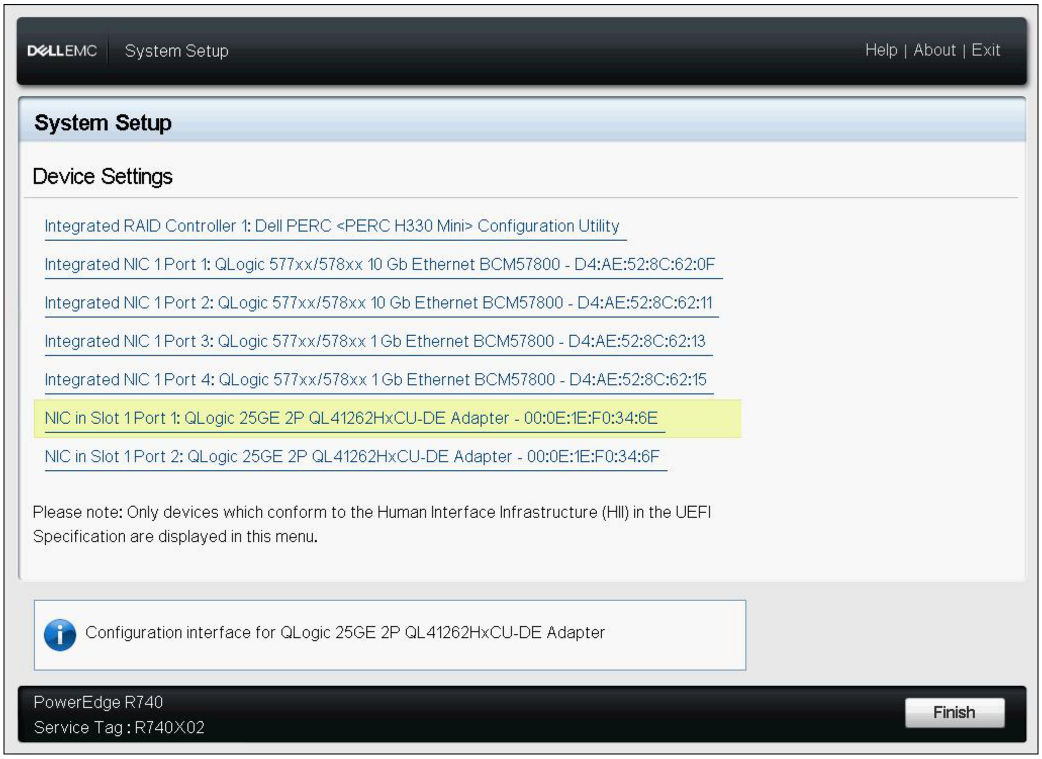
PowerEdge R740
Service Tag : R740X02

Figure 10-1. System Setup: Selecting Device Settings

NOTE

SAN boot is supported in the UEFI environment only. Make sure the system boot option is UEFI, and not legacy.

3. On the Device Settings page, select the QLogic device (Figure 10-2).



System Setup

Device Settings

- Integrated RAID Controller 1: Dell PERC <PERC H330 Mini> Configuration Utility
- Integrated NIC 1 Port 1: QLogic 577xx/578xx 10 Gb Ethernet BCM57800 - D4:AE:52:8C:62:0F
- Integrated NIC 1 Port 2: QLogic 577xx/578xx 10 Gb Ethernet BCM57800 - D4:AE:52:8C:62:11
- Integrated NIC 1 Port 3: QLogic 577xx/578xx 1 Gb Ethernet BCM57800 - D4:AE:52:8C:62:13
- Integrated NIC 1 Port 4: QLogic 577xx/578xx 1 Gb Ethernet BCM57800 - D4:AE:52:8C:62:15
- NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E
- NIC in Slot 1 Port 2: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6F

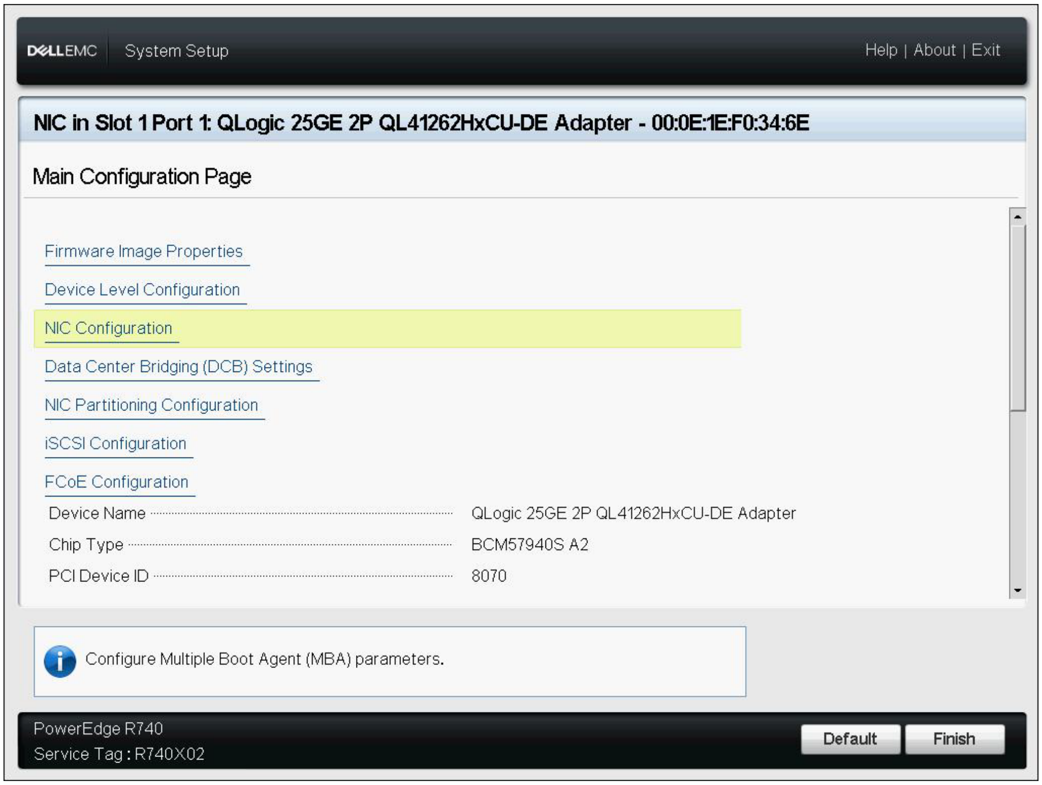
Please note: Only devices which conform to the Human Interface Infrastructure (HII) in the UEFI Specification are displayed in this menu.

Configuration interface for QLogic 25GE 2P QL41262HxCU-DE Adapter

PowerEdge R740
Service Tag : R740X02

Figure 10-2. System Setup: Device Settings, Port Selection

4. On the Main Configuration Page, select **NIC Configuration** (Figure 10-3), and then press ENTER.



DELLEMC System Setup
Help | About | Exit

NIC in Slot 1 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:F0:34:6E

Main Configuration Page

Firmware Image Properties
Device Level Configuration
NIC Configuration
Data Center Bridging (DCB) Settings
NIC Partitioning Configuration
iSCSI Configuration
FCoE Configuration

Device Name QLogic 25GE 2P QL41262HxCU-DE Adapter
Chip Type BCM57940S A2
PCI Device ID 8070

Configure Multiple Boot Agent (MBA) parameters.

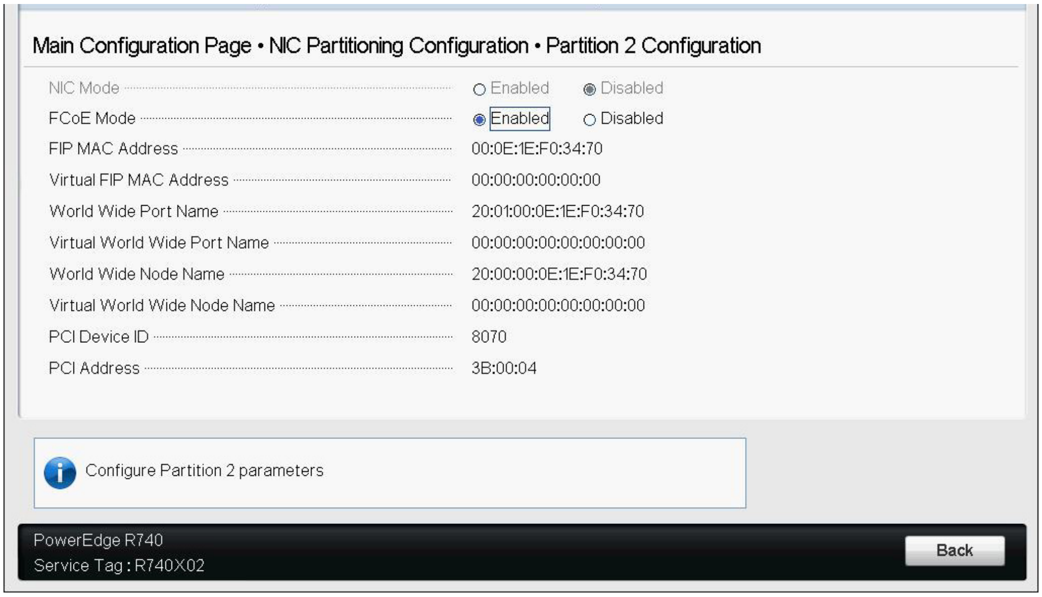
PowerEdge R740
Service Tag : R740X02

Figure 10-3. System Setup: NIC Configuration

5. On the NIC Configuration page, select **Boot Mode**, press ENTER, and then select **FCoE** as a preferred boot mode.

NOTE

FCoE is not listed as a boot option if the **FCoE Mode** feature is disabled at the port level. If the **Boot Mode** preferred is **FCoE**, make sure the **FCoE Mode** feature is enabled as shown in Figure 10-4. Not all adapter versions support FCoE.



Main Configuration Page • NIC Partitioning Configuration • Partition 2 Configuration

NIC Mode	○ Enabled	● Disabled
FCoE Mode	● Enabled	○ Disabled
FIP MAC Address	00:0E:1E:F0:34:70	
Virtual FIP MAC Address	00:00:00:00:00:00	
World Wide Port Name	20:01:00:0E:1E:F0:34:70	
Virtual World Wide Port Name	00:00:00:00:00:00:00:00	
World Wide Node Name	20:00:00:0E:1E:F0:34:70	
Virtual World Wide Node Name	00:00:00:00:00:00:00:00	
PCI Device ID	8070	
PCI Address	3B:00:04	

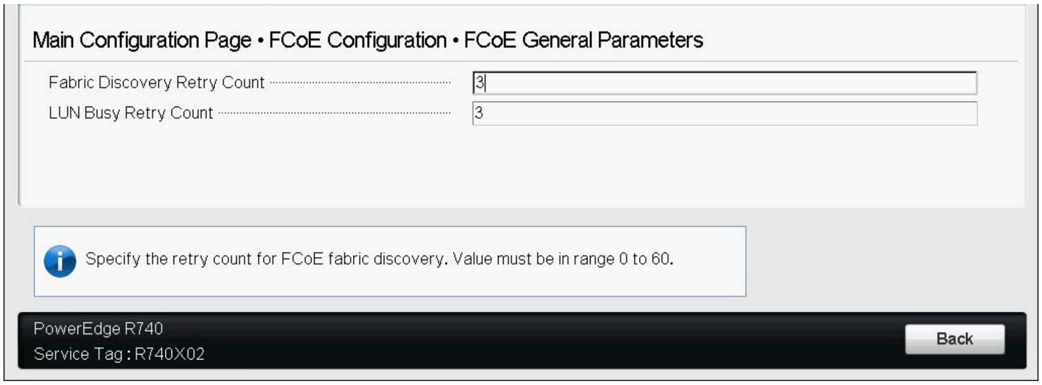
Configure Partition 2 parameters

PowerEdge R740
Service Tag : R740X02

Figure 10-4. System Setup: FCoE Mode Enabled

To configure the FCoE boot parameters:

1. On the Device HII Main Configuration Page, select **FCoE Configuration**, and then press ENTER.
2. On the FCoE Configuration Page, select **FCoE General Parameters**, and then press ENTER.
3. On the FCoE General Parameters page (Figure 10-5), press the UP ARROW and DOWN ARROW keys to select a parameter, and then press ENTER to select and input the following values:
   - **Fabric Discovery Retry Count**: Default value or as required
   - **LUN Busy Retry Count**: Default value or as required



Main Configuration Page • FCoE Configuration • FCoE General Parameters

Fabric Discovery Retry Count: 3

LUN Busy Retry Count: 3

Specify the retry count for FCoE fabric discovery. Value must be in range 0 to 60.

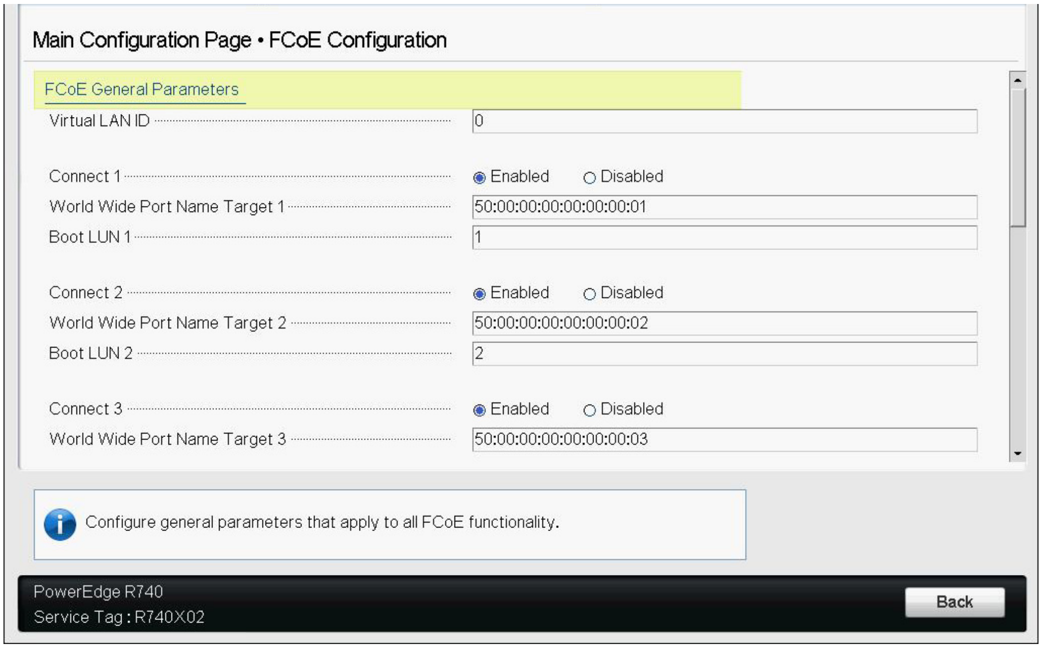
PowerEdge R740
Service Tag : R740X02

Figure 10-5. System Setup: FCoE General Parameters

4. Return to the FCoE Configuration page.
5. Press ESC, and then select **FCoE Target Parameters**.
6. Press ENTER.
7. In the **FCoE General Parameters Menu**, enable **Connect** to the preferred FCoE target.
8. Type values for the following parameters (Figure 10-6) for the iSCSI target, and then press ENTER:

   ❑ **World Wide Port Name Target *n***
   ❑ **Boot LUN *n***

   Where the value of *n* is between 1 and 8, enabling you to configure 8 FCoE targets.



Main Configuration Page • FCoE Configuration

FCoE General Parameters

Virtual LAN ID: 0

Connect 1: ● Enabled ○ Disabled

World Wide Port Name Target 1: 50:00:00:00:00:00:00:01

Boot LUN 1: 1

Connect 2: ● Enabled ○ Disabled

World Wide Port Name Target 2: 50:00:00:00:00:00:00:02

Boot LUN 2: 2

Connect 3: ● Enabled ○ Disabled

World Wide Port Name Target 3: 50:00:00:00:00:00:00:03

Configure general parameters that apply to all FCoE functionality.

Figure 10-6. System Setup: FCoE General Parameters

Windows FCoE Boot from SAN

FCoE boot from SAN information for Windows includes:

- Windows Server 2012 R2 and 2016 FCoE Boot Installation
- Configuring FCoE
- FCoE Crash Dump

Windows Server 2012 R2 and 2016 FCoE Boot Installation

For Windows Server 2012R2/2016 boot from SAN installation, Cavium requires the use of a “slipstream” DVD, or ISO image, with the latest Cavium QLogic drivers injected. See “Injecting (Slipstreaming) Adapter Drivers into Windows Image Files” on page 184.

The following procedure prepares the image for installation and booting in FCoE mode.

To set up Windows Server 2012R2/2016 FCoE boot:

1. Remove any local hard drives on the system to be booted (remote system).
2. Prepare the Windows OS installation media by following the slipstreaming steps in “Injecting (Slipstreaming) Adapter Drivers into Windows Image Files” on page 184.
3. Load the latest Cavium QLogic FCoE boot images into the adapter NVRAM.

4. Configure the FCoE target to allow a connection from the remote device. Ensure that the target has sufficient disk space to hold the new OS installation.
5. Configure the UEFI HII to set the FCoE boot type on the required adapter port, correct initiator, and target parameters for FCoE boot.
6. Save the settings and reboot the system. The remote system should connect to the FCoE target, and then boot from the DVD-ROM device.
7. Boot from DVD and begin installation.
8. Follow the on-screen instructions.

   On the window that shows the list of disks available for the installation, the FCoE target disk should be visible. This target is a disk connected through the FCoE boot protocol, located in the remote FCoE target.
9. To proceed with Windows Server 2012R2/2016 installation, select **Next**, and then follow the on-screen instructions. The server will undergo a reboot multiple times as part of the installation process.
10. After the server boots to the OS, you should run the driver installer to complete the Cavium QLogic drivers and application installation.

Configuring FCoE

By default, DCB is enabled on Cavium QLogic 41*xxx* FCoE- and DCB-compatible C-NICs. Cavium QLogic 41*xxx* FCoE requires a DCB-enabled interface. For Windows operating systems, use QConvergeConsole GUI or a command line utility to configure the DCB parameters.

FCoE Crash Dump

Crash dump functionality is currently supported for FCoE boot for the FastLinQ 41*xxx* Series Adapters.

No additional configuration is required for FCoE crash-dump generation when in FCoE boot mode.

Injecting (Slipstreaming) Adapter Drivers into Windows Image Files

To inject adapter drivers into the Windows image files:

1. Obtain the latest driver package for the applicable Windows Server version (2012, 2012 R2, or 2016).
2. Extract the driver package to a working directory:
   a. Open a command line session and navigate to the folder that contains the driver package.
   b. To start the driver installer, issue the following command:

   ```
   setup.exe /a
   ```

   c. In the `Network location` field, enter the path of the folder to which to extract the driver package. For example, type **`c:\temp`**.
   d. Follow the driver installer instructions to install the drivers in the specified folder. In this example, the Cavium QLogic driver files are installed here:

   ```
   c:\temp\Program File 64\QLogic Corporation\QDrivers
   ```

3. Download the Windows Assessment and Deployment Kit (ADK) version 10 from Microsoft:

   https://developer.microsoft.com/en-us/windows/hardware/windows-assessment-deployment-kit

4. Open a command line session (with administrator privilege), and navigate through the release CD to the `Tools\Slipstream` folder.
5. Locate the `slipstream.bat` script file, and then issue the following command:

   ```
   slipstream.bat <path>
   ```

   Where `<path>` is the drive and subfolder that you specified in Step 2. For example:

   ```
   slipstream.bat "c:\temp\Program Files 64\QLogic
   Corporation\QDrivers"
   ```

NOTE

Note the following regarding the operating system installation media:

- Operating system installation media is expected to be a local drive. Network paths for operating system installation media are not supported.
- The `slipstream.bat` script injects the driver components in all the SKUs that are supported by the operating system installation media.

6. Burn a DVD containing the resulting driver ISO image file located in the working directory.
7. Install the Windows Server operating system using the new DVD.

Configuring Linux FCoE Offload

The Cavium FastLinQ 41*xxx* Series Adapter FCoE software consists of a single kernel module called qedf.ko (qedf). The qedf module is dependent on additional parts of the Linux kernel for specific functionality:

- **`qed.ko`** is the Linux eCore kernel module used for common Cavium FastLinQ 41*xxx* hardware initialization routines.
- **`libfcoe.ko`** is the Linux FCoE kernel library needed to conduct FCoE forwarder (FCF) solicitation and FCoE initialization protocol (FIP) fabric login (FLOGI).
- **`libfc.ko`** is the Linux FC kernel library needed for several functions, including:
  - ❑ Name server login and registration
  - ❑ rport session management
- **`scsi_transport_fc.ko`** is the Linux FC SCSI transport library used for remote port and SCSI target management.

These modules must be loaded before qedf can be functional, otherwise errors such as “unresolved symbol” can result. If the qedf module is installed in the distribution update path, the requisite modules are automatically loaded by modprobe. Cavium FastLinQ 41*xxx* Series Adapters support FCoE offload.

This section provides the following information about FCoE offload in Linux:

- Differences Between qedf and bnx2fc
- Configuring qedf.ko
- Verifying FCoE Devices in Linux

- Prerequisites for FCoE Boot from SAN
- Configuring FCoE Boot from SAN

Differences Between qedf and bnx2fc

Significant differences exist between qedf—the driver for the Cavium FastLinQ 41*xxx* 10/25GbE Controller (FCoE)—and the previous QLogic FCoE offload driver, bnx2fc. Differences include:

- qedf directly binds to a PCI function exposed by the CNA.
- qedf does not need the open-fcoe user space tools (fipvlan, fcoemon, fcoeadm) to initiate discovery.
- qedf issues FIP VLAN requests directly and does not need the fipvlan utility.
- qedf does not need an FCoE interface created by fipvlan for fcoemon.
- qedf does not sit on top of the net_device.
- qedf is not dependent on network drivers (such as bnx2x and cnic).
- qedf will automatically initiate FCoE discovery on link up (because it is not dependent on fipvlan or fcoemon for FCoE interface creation).

NOTE

FCoE interfaces no longer sit on top of the network interface. The qedf driver automatically creates FCoE interfaces that are separate from the network interface. Thus, FCoE interfaces do not show up in the FCoE interface dialog box in the installer. Instead, the disks show up automatically as SCSI disks, similar to the way Fibre Channel drivers work.

Configuring qedf.ko

No explicit configuration is required for qedf.ko. The driver automatically binds to the exposed FCoE functions of the CNA and begins discovery. This functionality is similar to the functionality and operation of the Cavium QLogic FC driver, qla2xx, as opposed to the older bnx2fc driver.

NOTE

For more information on FastLinQ driver installation, see Chapter 3 Driver Installation.

The load qedf.ko kernel module performs the following:

```
# modprobe qed
# modprobe libfcoe
# modprobe qedf
```

Verifying FCoE Devices in Linux

Follow these steps to verify that the FCoE devices were detected correctly after installing and loading the qedf kernel module.

To verify FCoE devices in Linux:

1. Check lsmod to verify that the qedf and associated kernel modules were loaded:

```
# lsmod | grep qedf
69632   1   qedf libfc
143360  2   qedf,libfcoe scsi_transport_fc
65536   2   qedf,libfc qed
806912  1   qedf scsi_mod
262144  14  sg,hpsa,qedf,scsi_dh_alua,scsi_dh_rdac,dm_multipath,
scsi_transport_fc,scsi_transport_sas,libfc,scsi_transport_iscsi,scsi_dh_emc,
libata,sd_mod,sr_mod
```

2. Check dmesg to verify that the FCoE devices were detected properly. In this example, the two detected FCoE CNA devices are SCSI host numbers 4 and 5.

```
# dmesg | grep qedf
[  235.321185] [0000:00:00.0]: [qedf_init:3728]: QLogic FCoE Offload Driver
v8.18.8.0.
....
[  235.322253] [0000:21:00.2]: [__qedf_probe:3142]:4: QLogic FastLinQ FCoE
Module qedf 8.18.8.0, FW 8.18.10.0
[  235.606443] scsi host4: qedf
....
[  235.624337] [0000:21:00.3]: [__qedf_probe:3142]:5: QLogic FastLinQ FCoE
Module qedf 8.18.8.0, FW 8.18.10.0
[  235.886681] scsi host5: qedf
....
[  243.991851] [0000:21:00.3]: [qedf_link_update:489]:5: LINK UP (40 GB/s).
```

3. Check for discovered FCoE devices using the `lsscsi` or `lsblk -S` commands. An example of each command follows.

```
# lsscsi
```

```
[0:2:0:0]     disk    DELL     PERC H700         2.10  /dev/sda
[2:0:0:0]     cd/dvd  TEAC     DVD-ROM DV-28SW   R.2A  /dev/sr0
[151:0:0:0]   disk    HP       P2000G3 FC/iSCSI  T252  /dev/sdb
[151:0:0:1]   disk    HP       P2000G3 FC/iSCSI  T252  /dev/sdc
[151:0:0:2]   disk    HP       P2000G3 FC/iSCSI  T252  /dev/sdd
[151:0:0:3]   disk    HP       P2000G3 FC/iSCSI  T252  /dev/sde
[151:0:0:4]   disk    HP       P2000G3 FC/iSCSI  T252  /dev/sdf

# lsblk -S
NAME HCTL       TYPE  VENDOR    MODEL            REV  TRAN
sdb  5:0:0:0    disk  SANBlaze  VLUN P2T1L0      V7.3 fc
sdc  5:0:0:1    disk  SANBlaze  VLUN P2T1L1      V7.3 fc
sdd  5:0:0:2    disk  SANBlaze  VLUN P2T1L2      V7.3 fc
sde  5:0:0:3    disk  SANBlaze  VLUN P2T1L3      V7.3 fc
sdf  5:0:0:4    disk  SANBlaze  VLUN P2T1L4      V7.3 fc
sdg  5:0:0:5    disk  SANBlaze  VLUN P2T1L5      V7.3 fc
sdh  5:0:0:6    disk  SANBlaze  VLUN P2T1L6      V7.3 fc
sdi  5:0:0:7    disk  SANBlaze  VLUN P2T1L7      V7.3 fc
sdj  5:0:0:8    disk  SANBlaze  VLUN P2T1L8      V7.3 fc
sdk  5:0:0:9    disk  SANBlaze  VLUN P2T1L9      V7.3 fc
```

Configuration information for the host is located in `/sys/class/fc_host/hostX`, where `X` is the number of the SCSI host. In the preceding example, `X` is 4. The `hostX` file contains attributes for the FCoE function, such as worldwide port name and fabric ID.

Prerequisites for FCoE Boot from SAN

The following are required for Linux FCoE boot from SAN to function correctly with the Cavium FastLinQ 41*xxx* 10/25GbE Controller.

General

You no longer need to use the FCoE disk tabs in the Red Hat and SUSE installers because the FCoE interfaces are not exposed from the network interface and are automatically activated by the qedf driver.

SLES 11

- The driver update disk is required because the necessary FCoE drivers are not inbox.
- The installer parameter `dud=1` is required to ensure that the installer will ask for the driver update disk.

- Do not use the installer parameter `withfcoe=1` because the software FCoE will conflict with the hardware offload if network interfaces from qede are exposed.

SLES 12

- The driver update disk is required because the necessary drivers are not inbox in SP 2 and earlier.
- Driver update disk is recommended for SP 3 and later.
- The installer parameter `dud=1` is required to ensure that the installer will ask for the driver update disk.
- Do not use the installer parameter `withfcoe=1` because the software FCoE will conflict with the hardware offload if network interfaces from qede are exposed.

RHEL 6

- The driver update disk is required because the necessary FCoE drivers are not inbox.
- Use the installer parameter `dd` to ensure that the installer will ask for the driver update disk.

RHEL 7

- The driver update disk is required because the necessary FCoE drivers are not inbox on 7.3 and earlier.
- For RHEL 7.4 and later, you must blacklist the inbox drivers to ensure that the out-of-box drivers from the driver update disk will load correctly. For details, see "Configuring FCoE Boot from SAN for RHEL 7.4 and Later" on page 191.

Configuring FCoE Boot from SAN

This section provides FCoE boot from SAN procedures for the following Linux distributions:

- Configuring FCoE Boot from SAN for RHEL 6.10
- Configuring FCoE Boot from SAN for RHEL 7.4 and Later
- Configuring FCoE Boot from SAN for SLES 12 SP3 and Later
- Using an FCoE Boot Device as a kdump Target

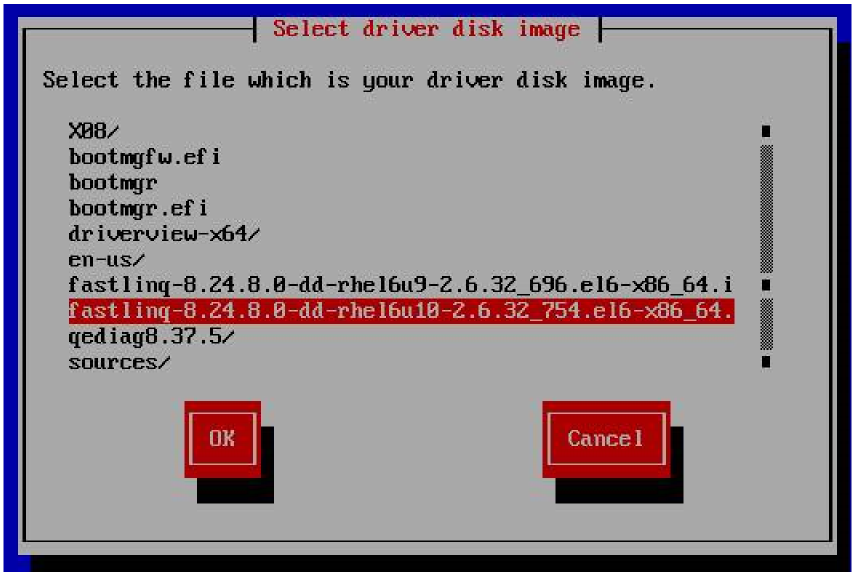
Configuring FCoE Boot from SAN for RHEL 6.10

To install RHEL 6.10:

1. Boot from the RHEL 6.10 installation media with the FCoE target already connected in UEFI.
2. Press any key to install an out-of-box driver.
3. Select the `Install system with basic video driver` option, and then press the E key.
4. Select the `kernel` line, and then press the E key to edit the line.
5. Modify the kernel line by typing **`linux dd`** at the end, as shown in the following:

   ```
   grub edit> kernel /images/pxeboot/vmlinuz nomodeset askmethod
   linux dd
   ```

6. Press ENTER to initiate the driver update process.
7. Press the B key to continue with the installation.
8. At the **Select the file which is your driver disk image** prompt, select the out-of-box driver (FastLinQ driver update disk), and then click **OK**. Figure 10-7 shows an example.



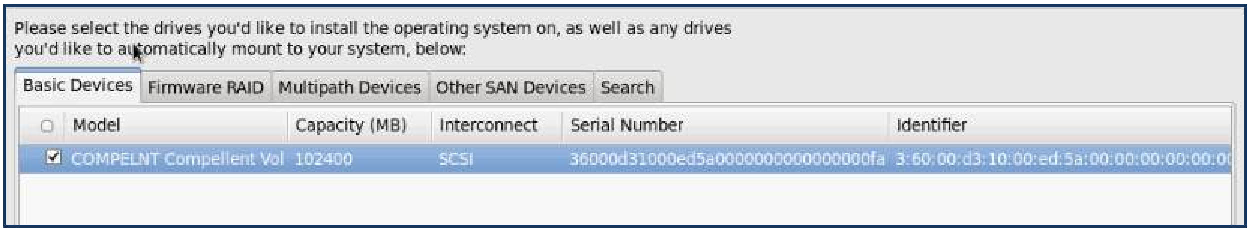
```
Select driver disk image

Select the file which is your driver disk image.

  X08/
  bootmgfw.efi
  bootmgr
  bootmgr.efi
  driverview-x64/
  en-us/
  fastlinq-8.24.8.0-dd-rhel6u9-2.6.32_696.el6-x86_64.i
  fastlinq-8.24.8.0-dd-rhel6u10-2.6.32_754.el6-x86_64.
  qediag8.37.5/
  sources/
```

Figure 10-7. Selecting the Driver Disk Image

9. Load the FastLinQ driver update disk and then click **Next** to continue with the installation.(You can skip the media test.)

10. At the **What type of devices will your installation involve?** prompt, select **Specialized Storage Devices**.

11. At the **Please select the drives...** prompt, on the Basic Devices page, select the FCoE LUN. Figure 10-8 shows an example.



Please select the drives you'd like to install the operating system on, as well as any drives you'd like to automatically mount to your system, below:

○	Model	Capacity (MB)	Interconnect	Serial Number	Identifier
☑	COMPELNT Compellent Vol	102400	SCSI	36000d31000ed5a0000000000000000fa	3:60:00:d3:10:00:ed:5a:00:00:00:00:00:0

Figure 10-8. Selecting the Drives

12. Complete the installation and boot to the OS.

13. Turn off the lldpad and fcoe services that are used for software FCoE. (If they are active, they can interfere with the normal operation of the hardware offload FCoE.) Issue the following commands:

```
# service fcoe stop
# service lldpad stop
# chkconfig fcoe off
# chkconfig lldpad off
```

Configuring FCoE Boot from SAN for RHEL 7.4 and Later

To install RHEL 7.4 and later:

1. Boot from the RHEL 7.*x* installation media with the FCoE target already connected in UEFI.

```
Install Red Hat Enterprise Linux 7.x
Test this media & install Red Hat Enterprise 7.x
Troubleshooting -->

Use the UP and DOWN keys to change the selection
Press 'e' to edit the selected item or 'c' for a command
prompt
```

2. To install an out-of-box driver, press the E key. Otherwise, proceed to Step 7.

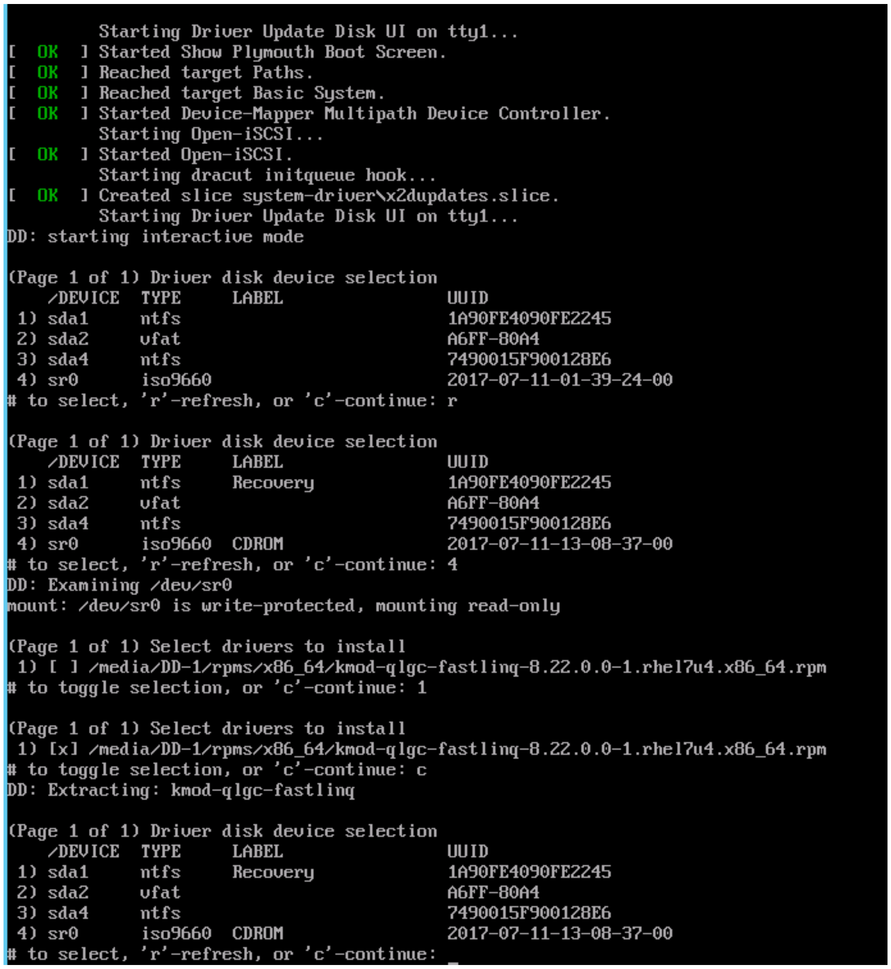
3. Select the `kernel` line, and then press the E key to edit the line.

4. Issue the following command, and then press ENTER.

   `linux dd modprobe.blacklist=qed modprobe.blacklist=qede modprobe.blacklist=qedr modprobe.blacklist=qedi modprobe.blacklist=qedf`

   You can use the `inst.dd` option instead of `linux dd`.

5. The installation process prompts you to install the out-of-box driver as shown in the Figure 10-9 example.



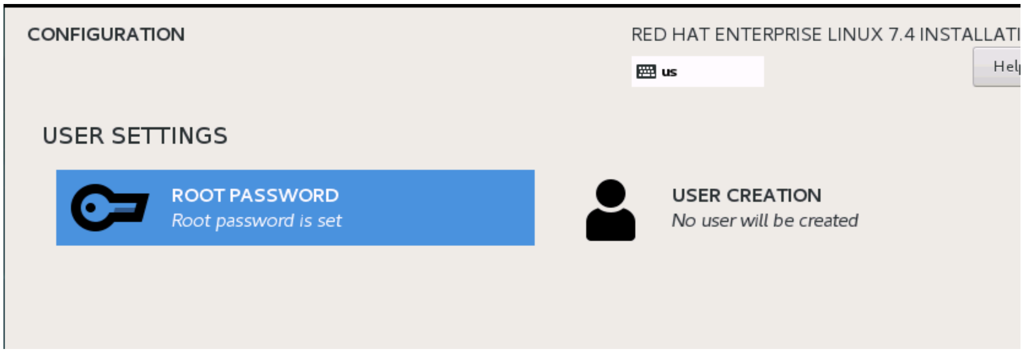
```
         Starting Driver Update Disk UI on tty1...
[  OK  ] Started Show Plymouth Boot Screen.
[  OK  ] Reached target Paths.
[  OK  ] Reached target Basic System.
[  OK  ] Started Device-Mapper Multipath Device Controller.
         Starting Open-iSCSI...
[  OK  ] Started Open-iSCSI.
         Starting dracut initqueue hook...
[  OK  ] Created slice system-driver\x2dupdates.slice.
         Starting Driver Update Disk UI on tty1...
DD: starting interactive mode

(Page 1 of 1) Driver disk device selection
    /DEVICE  TYPE     LABEL                UUID
 1) sda1     ntfs                          1A90FE4090FE2245
 2) sda2     vfat                          A6FF-80A4
 3) sda4     ntfs                          7490015F900128E6
 4) sr0      iso9660                       2017-07-11-01-39-24-00
# to select, 'r'-refresh, or 'c'-continue: r

(Page 1 of 1) Driver disk device selection
    /DEVICE  TYPE     LABEL                UUID
 1) sda1     ntfs     Recovery             1A90FE4090FE2245
 2) sda2     vfat                          A6FF-80A4
 3) sda4     ntfs                          7490015F900128E6
 4) sr0      iso9660  CDROM                2017-07-11-13-08-37-00
# to select, 'r'-refresh, or 'c'-continue: 4
DD: Examining /dev/sr0
mount: /dev/sr0 is write-protected, mounting read-only

(Page 1 of 1) Select drivers to install
 1) [ ] /media/DD-1/rpms/x86_64/kmod-qlgc-fastlinq-8.22.0.0-1.rhel7u4.x86_64.rpm
# to toggle selection, or 'c'-continue: 1

(Page 1 of 1) Select drivers to install
 1) [x] /media/DD-1/rpms/x86_64/kmod-qlgc-fastlinq-8.22.0.0-1.rhel7u4.x86_64.rpm
# to toggle selection, or 'c'-continue: c
DD: Extracting: kmod-qlgc-fastlinq

(Page 1 of 1) Driver disk device selection
    /DEVICE  TYPE     LABEL                UUID
 1) sda1     ntfs     Recovery             1A90FE4090FE2245
 2) sda2     vfat                          A6FF-80A4
 3) sda4     ntfs                          7490015F900128E6
 4) sr0      iso9660  CDROM                2017-07-11-13-08-37-00
# to select, 'r'-refresh, or 'c'-continue: _
```

Figure 10-9. Prompt for Out-of-Box Installation

6. If required for your setup, load the FastLinQ driver update disk when prompted for additional driver disks. Otherwise, press the C key if you have no other driver update disks to install.

7. Continue with the installation. You can skip the media test. Click **Next** to continue with the installation.

8. In the Configuration window (Figure 10-10), select the language to use during the installation process, and then click **Continue**.



CONFIGURATION

RED HAT ENTERPRISE LINUX 7.4 INSTALLATI

USER SETTINGS

ROOT PASSWORD
Root password is set

USER CREATION
No user will be created

Figure 10-10. Red Hat Enterprise Linux 7.4 Configuration

9. In the Installation Summary window, click **Installation Destination**. The disk label is *sda*, indicating a single-path installation. If you configured multipath, the disk has a device mapper label.
10. In the **Specialized & Network Disks** section, select the FCoE LUN.
11. Type the root user's password, and then click **Next** to complete the installation.
12. During the first boot, add the following kernel command line to fall into shell.

    ```
    rd.driver.pre=qed,qede,qedr,qedf,qedi
    ```

13. After a successful system boot, edit the `/etc/modprobe.d/anaconda-blacklist.conf` file to remove the blacklist entry for the selected driver.
14. Rebuild the ramdisk by issuing the **`dracut -f`** command, and then reboot.
15. Turn off the lldpad and fcoe services that are used for software FCoE. (If they are active, they can interfere with the typical operation of the hardware offload FCoE.) Issue the appropriate commands based on your operating system:

    ❑ For SLES 11:

    ```
    # service boot.fcoe stop
    # service boot.lldpad stop
    # chkconfig boot.fcoe off
    # chkconfig boot.lldpad off
    ```

    ❑ For RHEL 7 and SLES 12:

    ```
    # systemctl stop fcoe
    # systemctl stop lldpad
    ```

```
# systemctl disable fcoe
# systemctl disable lldpad
```

Configuring FCoE Boot from SAN for SLES 12 SP3 and Later

No additional steps, other than injecting DUD for out-of-box driver, are necessary to perform boot from SAN installations when using SLES 12 SP3.

Using an FCoE Boot Device as a kdump Target

When using a device exposed using the qedf driver as a kdump target for crash dumps, Cavium recommends that you specify the kdump `crashkernel` memory parameter at the kernel command line to be a minimum of 512MB. Otherwise, the kernel crash dump may not succeed.

For details on how to set the kdump `crashkernel` size, refer to your Linux distribution documentation.

VMware FCoE Boot from SAN

For VMware ESXi 6.5/6.7 boot from SAN installation, Cavium requires that you use a customized ESXi ISO image that is built with the latest Cavium QLogic Converged Network Adapter bundle injected. This section covers the following VMware FCoE boot from SAN procedures.

- Injecting (Slipstreaming) Adapter Drivers into Image Files
- Installing the Customized ISO

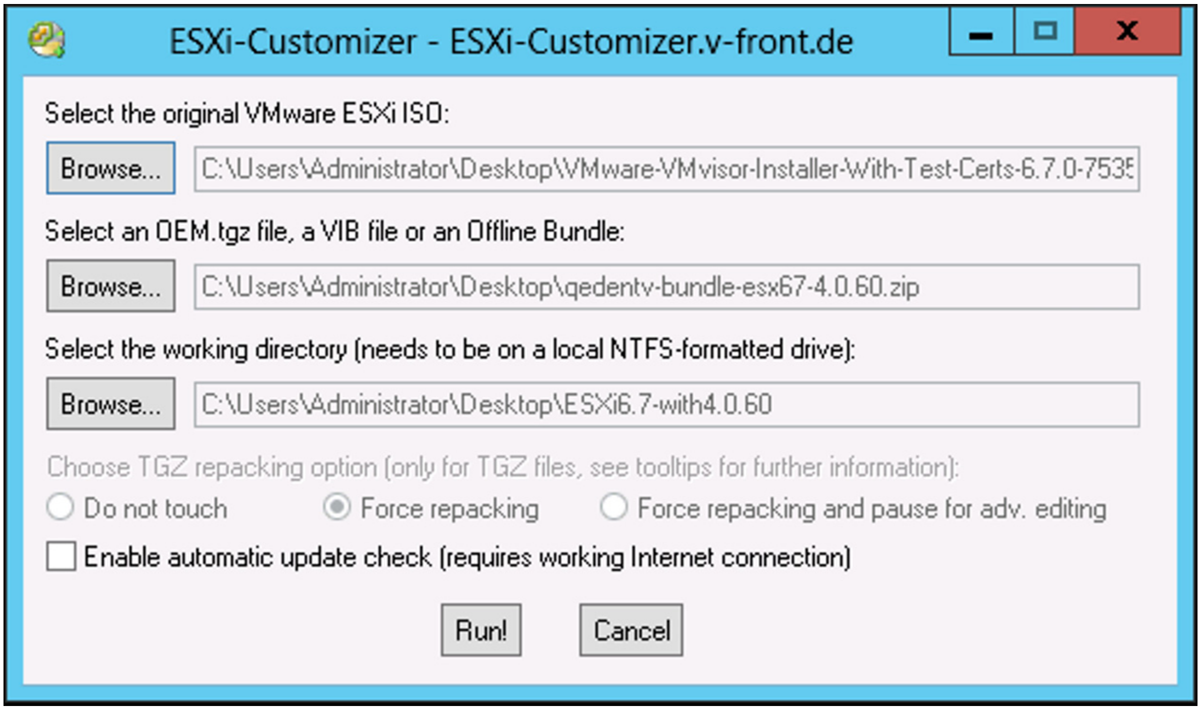
Injecting (Slipstreaming) Adapter Drivers into Image Files

This procedure uses ESXi-Customizer tool v2.7.2 as an example, but you can use any ESXi customizer.

To inject adapter drivers into an ESXi image file:

1. Download ESXi-Customizer v2.7.2 or later.
2. Go to the `ESXi customizer` directory.
3. Issue the **`ESXi-Customizer.cmd`** command.
4. In the ESXi-Customizer dialog box, complete the following.
    a. Select the original VMware ESXi ISO file.
    b. Select either the QLogic FCoE driver VIB file or the QLogic offline qedentv bundle ZIP file.
    c. In the working directory, select the folder in which the customized ISO needs to be created.
    d. Click **Run**.

Figure 10-11 shows an example.



ESXi-Customizer - ESXi-Customizer.v-front.de

Select the original VMware ESXi ISO:

C:\Users\Administrator\Desktop\VMware-VMvisor-Installer-With-Test-Certs-6.7.0-7535

Select an OEM.tgz file, a VIB file or an Offline Bundle:

C:\Users\Administrator\Desktop\qedentv-bundle-esx67-4.0.60.zip

Select the working directory (needs to be on a local NTFS-formatted drive):

C:\Users\Administrator\Desktop\ESXi6.7-with4.0.60

Choose TGZ repacking option (only for TGZ files, see tooltips for further information):

○ Do not touch ● Force repacking ○ Force repacking and pause for adv. editing

□ Enable automatic update check (requires working Internet connection)

Figure 10-11. ESXi-Customizer Dialog Box

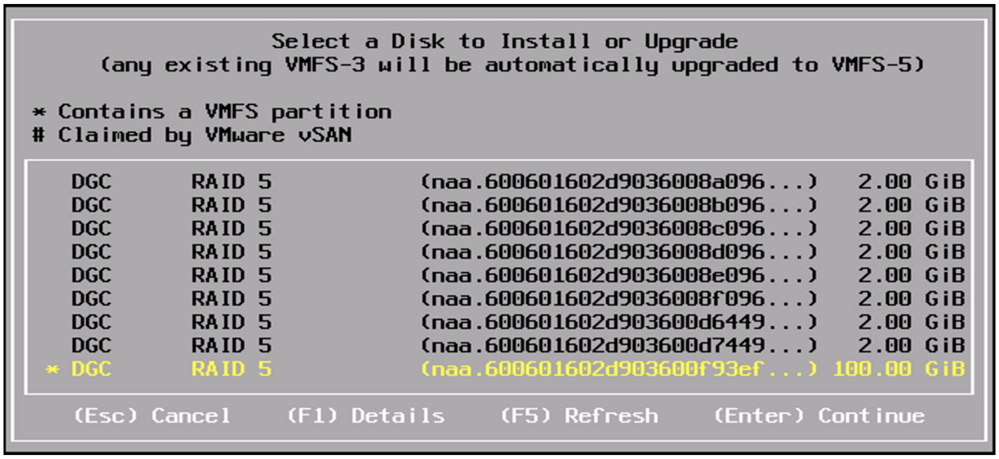
5. Burn a DVD that contains the customized ISO build located in the working directory specified in Step 4c.
6. Use the new DVD to install the ESXi OS.

Installing the Customized ISO

1. Load the latest Cavium QLogic FCOE boot images into the adapter NVRAM.
2. Configure the FCOE target to allow a valid connection with the remote machine. Ensure that the target has sufficient free disk space to hold the new OS installation.
3. Configure the UEFI HII to set the FCOE boot type on the required adapter port, the correct initiator, and the target parameters for FCOE boot.
4. Save the settings and reboot the system.

   The initiator should connect to an FCOE target and then boot the system from the DVD-ROM device.
5. Boot from the DVD and begin installation.
6. Follow the on-screen instructions.

On the window that shows the list of disks available for the installation, the FCOE target disk should be visible because the injected Converged Network Adapter bundle is inside the customized ESXi ISO. Figure 10-12 shows an example.

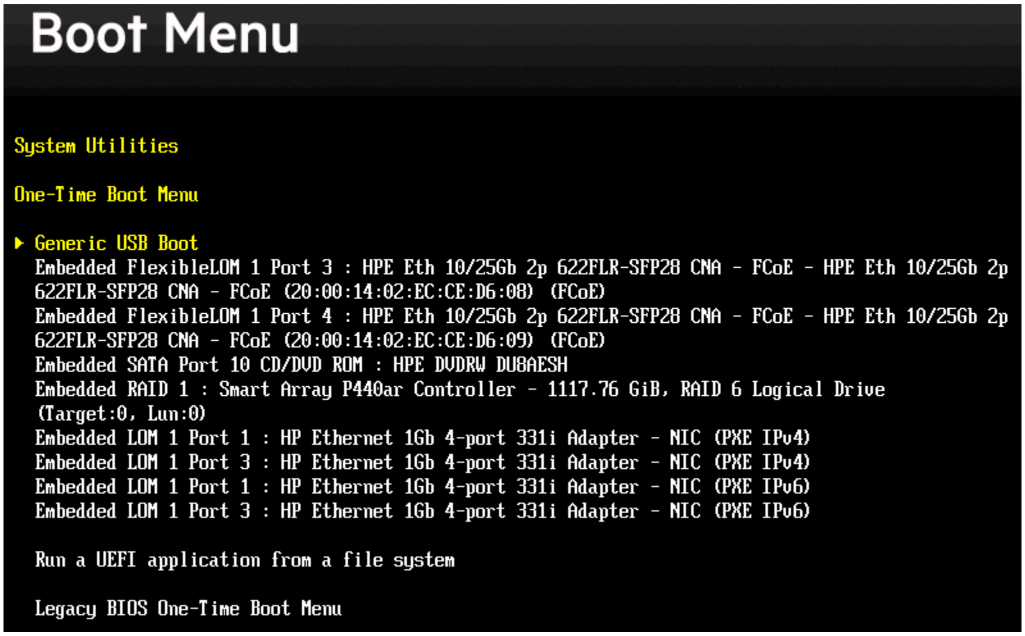


```
               Select a Disk to Install or Upgrade
      (any existing VMFS-3 will be automatically upgraded to VMFS-5)

 * Contains a VMFS partition
 # Claimed by VMware vSAN

   DGC      RAID 5          (naa.600601602d9036008a096...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d9036008b096...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d9036008c096...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d9036008d096...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d9036008e096...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d9036008f096...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d903600d6449...)   2.00 GiB
   DGC      RAID 5          (naa.600601602d903600d7449...)   2.00 GiB
 * DGC      RAID 5          (naa.600601602d903600f93ef...) 100.00 GiB

    (Esc) Cancel    (F1) Details    (F5) Refresh    (Enter) Continue
```

Figure 10-12. Select a VMware Disk to Install

7. Select the LUN on which ESXi can install, and then press ENTER.
8. On the next window, click **Next**, and then follow the on-screen instructions.
9. When installation completes, reboot the server and eject the DVD.
10. During the server boot, press the F9 key to access the **One-Time Boot Menu**, and then select **Boot media to QLogic adapter port**.
11. Under **Generic USB Boot**, select the newly installed ESXi to load through boot from SAN.

In the example shown in Figure 10-13, the first two ports indicate Cavium QLogic adapters.



```
Boot Menu

System Utilities

One-Time Boot Menu

▶ Generic USB Boot
  Embedded FlexibleLOM 1 Port 3 : HPE Eth 10/25Gb 2p 622FLR-SFP28 CNA - FCoE - HPE Eth 10/25Gb 2p
  622FLR-SFP28 CNA - FCoE (20:00:14:02:EC:CE:D6:08) (FCoE)
  Embedded FlexibleLOM 1 Port 4 : HPE Eth 10/25Gb 2p 622FLR-SFP28 CNA - FCoE - HPE Eth 10/25Gb 2p
  622FLR-SFP28 CNA - FCoE (20:00:14:02:EC:CE:D6:09) (FCoE)
  Embedded SATA Port 10 CD/DVD ROM : HPE DVDRW DU8AESH
  Embedded RAID 1 : Smart Array P440ar Controller - 1117.76 GiB, RAID 6 Logical Drive
  (Target:0, Lun:0)
  Embedded LOM 1 Port 1 : HP Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv4)
  Embedded LOM 1 Port 3 : HP Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv4)
  Embedded LOM 1 Port 1 : HP Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv6)
  Embedded LOM 1 Port 3 : HP Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv6)

  Run a UEFI application from a file system

  Legacy BIOS One-Time Boot Menu
```

Figure 10-13. VMware Generic USB Boot Options

11 SR-IOV Configuration

Single root input/output virtualization (SR-IOV) is a specification by the PCI SIG that enables a single PCI Express (PCIe) device to appear as multiple, separate physical PCIe devices. SR-IOV permits isolation of PCIe resources for performance, interoperability, and manageability.

NOTE

Some SR-IOV features may not be fully enabled in the current release.

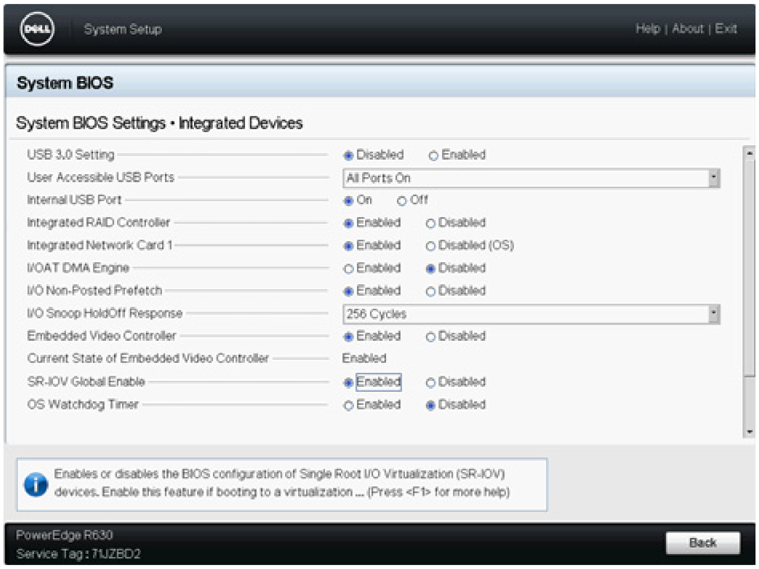
This chapter provides instructions for:

- Configuring SR-IOV on Windows
- "Configuring SR-IOV on Linux" on page 205
- "Configuring SR-IOV on VMware" on page 211

Configuring SR-IOV on Windows

To configure SR-IOV on Windows:

1. Access the server BIOS System Setup, and then click **System BIOS Settings**.
2. On the System BIOS Settings page, click **Integrated Devices**.
3. On the Integrated Devices page (Figure 11-1):
   a. Set the **SR-IOV Global Enable** option to **Enabled**.
   b. Click **Back**.



System BIOS

System BIOS Settings • Integrated Devices

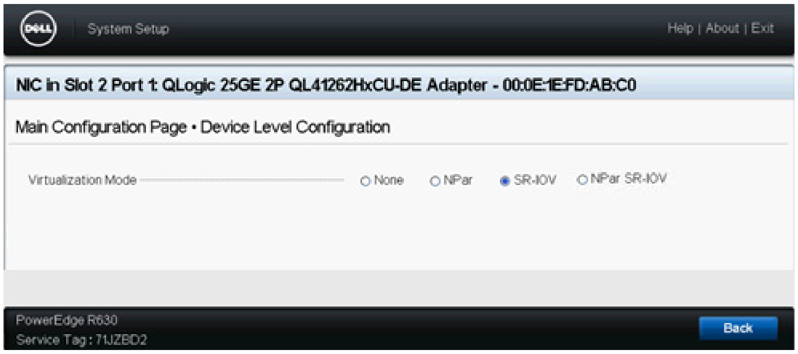
USB 3.0 Setting: ● Disabled ○ Enabled
User Accessible USB Ports: All Ports On
Internal USB Port: ● On ○ Off
Integrated RAID Controller: ● Enabled ○ Disabled
Integrated Network Card 1: ● Enabled ○ Disabled (OS)
I/OAT DMA Engine: ○ Enabled ● Disabled
I/O Non-Posted Prefetch: ● Enabled ○ Disabled
I/O Snoop HoldOff Response: 256 Cycles
Embedded Video Controller: ● Enabled ○ Disabled
Current State of Embedded Video Controller: Enabled
SR-IOV Global Enable: ● Enabled ○ Disabled
OS Watchdog Timer: ○ Enabled ● Disabled

Enables or disables the BIOS configuration of Single Root I/O Virtualization (SR-IOV) devices. Enable this feature if booting to a virtualization ... (Press <F1> for more help)

PowerEdge R630
Service Tag : 71JZBD2

Figure 11-1. System Setup for SR-IOV: Integrated Devices

4. On the Main Configuration Page for the selected adapter, click **Device Level Configuration**.
5. On the Main Configuration Page - Device Level Configuration (Figure 11-2):
   a. Set the **Virtualization Mode** to **SR-IOV**, or **NPAR+SR-IOV** if you are using NPAR mode.
   a.
   b. Click **Back**.



NIC in Slot 2 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:FD:AB:C0

Main Configuration Page • Device Level Configuration

Virtualization Mode: ○ None ○ NPar ● SR-IOV ○ NPar SR-IOV

PowerEdge R630
Service Tag : 71JZBD2

Figure 11-2. System Setup for SR-IOV: Device Level Configuration

6. On the Main Configuration Page, click **Finish**.
7. In the Warning - Saving Changes message box, click **Yes** to save the configuration.

8. In the Success - Saving Changes message box, click **OK**.

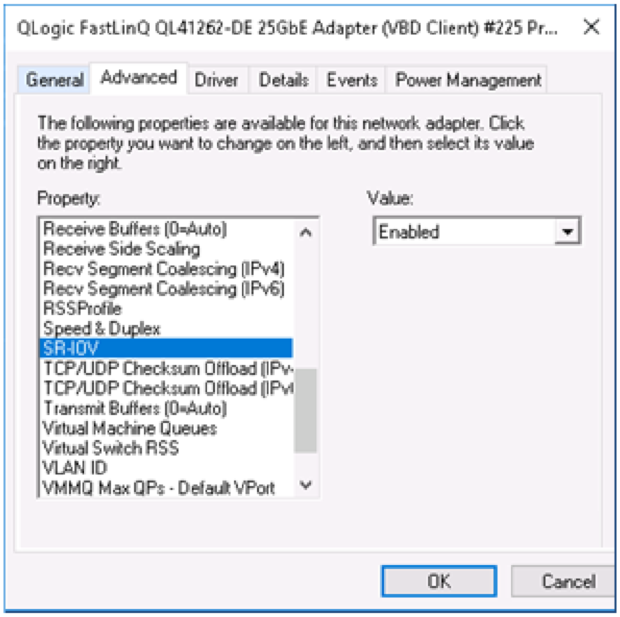
9. To enable SR-IOV on the miniport adapter:

   a. Access Device Manager.

   b. Open the miniport adapter properties, and then click the **Advanced** tab.

   c. On the Advanced properties page (Figure 11-3) under **Property**, select **SR-IOV**, and then set the value to **Enabled**.

   d. Click **OK**.


QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

Receive Buffers (0=Auto)
Receive Side Scaling
Recv Segment Coalescing (IPv4)
Recv Segment Coalescing (IPv6)
RSSProfile
Speed & Duplex
SR-IOV
TCP/UDP Checksum Offload (IPv
TCP/UDP Checksum Offload (IPv
Transmit Buffers (0=Auto)
Virtual Machine Queues
Virtual Switch RSS
VLAN ID
VMMQ Max QPs - Default VPort

Value: Enabled

Figure 11-3. Adapter Properties, Advanced: Enabling SR-IOV

10. To create a Virtual Machine Switch with SR-IOV (Figure 11-4 on page 201):

    a. Launch the Hyper-V Manager.

    b. Select **Virtual Switch Manager**.

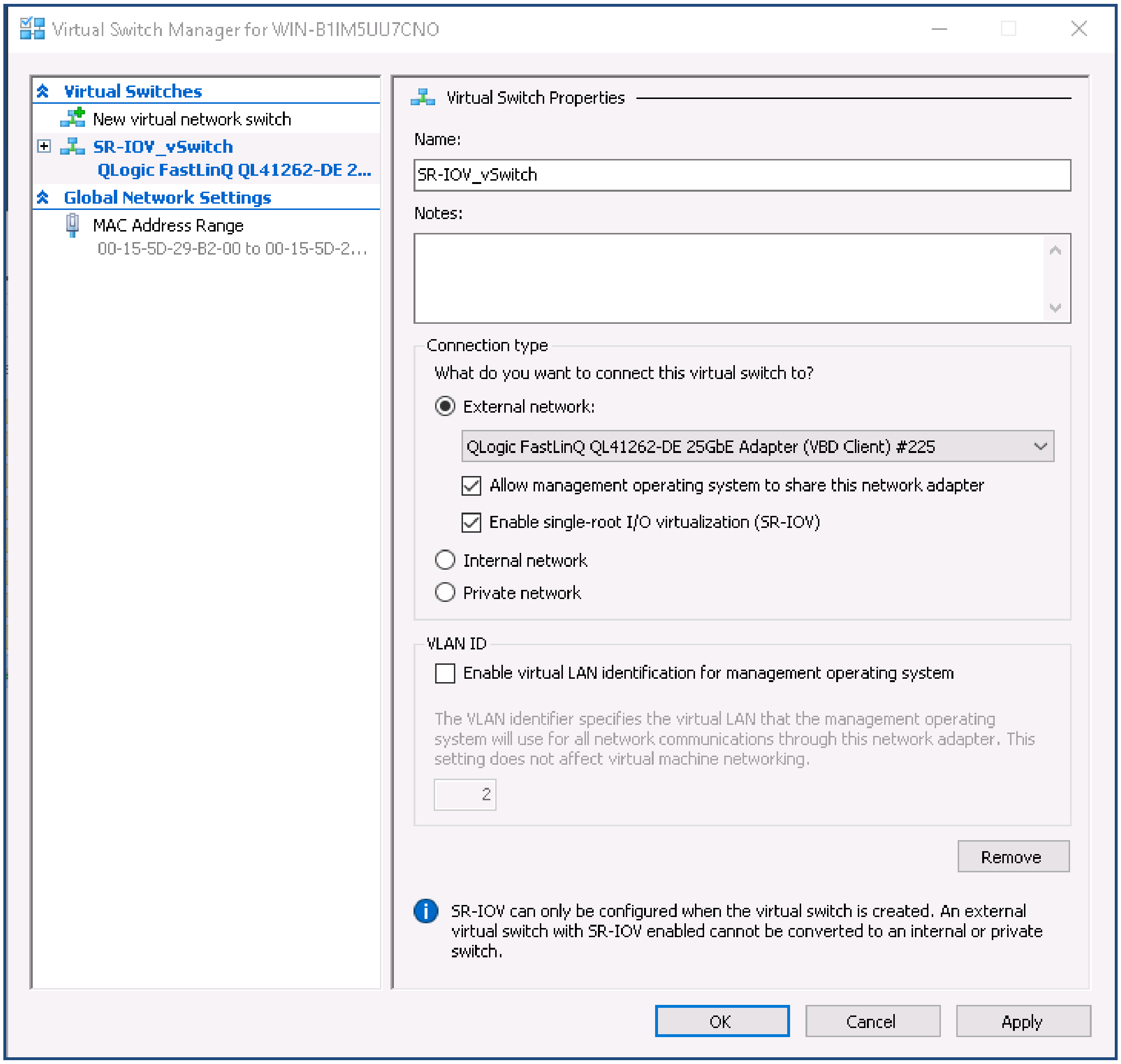
    c. In the **Name** box, type a name for the virtual switch.

    d. Under **Connection type**, select **External network**.

    e. Select the **Enable single-root I/O virtualization (SR-IOV)** check box, and then click **Apply**.

NOTE

Be sure to enable SR-IOV when you create the vSwitch. This option is unavailable after the vSwitch is created.



Virtual Switch Manager for WIN-B1IM5UU7CNO

- Virtual Switches
  - New virtual network switch
  - SR-IOV_vSwitch
    QLogic FastLinQ QL41262-DE 2...
- Global Network Settings
  - MAC Address Range
    00-15-5D-29-B2-00 to 00-15-5D-2...

Virtual Switch Properties

Name: SR-IOV_vSwitch

Notes:

Connection type

What do you want to connect this virtual switch to?

● External network:

QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225

☑ Allow management operating system to share this network adapter

☑ Enable single-root I/O virtualization (SR-IOV)

○ Internal network

○ Private network

VLAN ID

□ Enable virtual LAN identification for management operating system

The VLAN identifier specifies the virtual LAN that the management operating system will use for all network communications through this network adapter. This setting does not affect virtual machine networking.

2

SR-IOV can only be configured when the virtual switch is created. An external virtual switch with SR-IOV enabled cannot be converted to an internal or private switch.

Figure 11-4. Virtual Switch Manager: Enabling SR-IOV

f. The Apply Networking Changes message box advises you that **Pending changes may disrupt network connectivity**. To save your changes and continue, click **Yes**.

11. To get the virtual machine switch capability, issue the following Windows PowerShell command:

    ```
    PS C:\Users\Administrator> Get-VMSwitch -Name SR-IOV_vSwitch | fl
    ```

    Output of the `Get-VMSwitch` command includes the following SR-IOV capabilities:

    ```
    IovVirtualFunctionCount            : 96
    IovVirtualFunctionsInUse           : 1
    ```

12. To create a virtual machine (VM) and export the virtual function (VF) in the VM:

    a. Create a virtual machine.

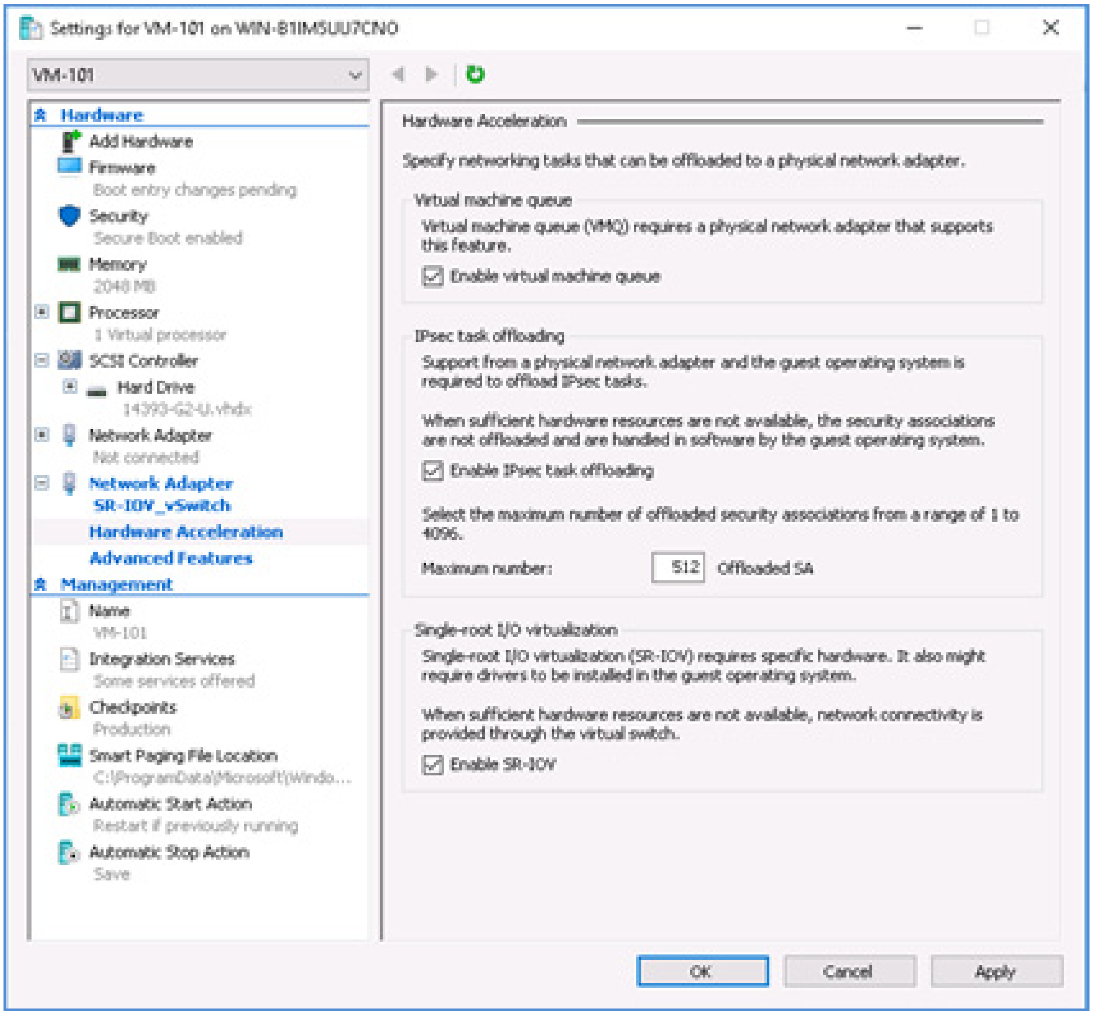
    b. Add the VMNetworkadapter to the virtual machine.

    c. Assign a virtual switch to the VMNetworkadapter.

    d. In the Settings for VM <VM_Name> dialog box (Figure 11-5), Hardware Acceleration page, under **Single-root I/O virtualization**, select the **Enable SR-IOV** check box, and then click **OK**.

    NOTE

    After the virtual adapter connection is created, the SR-IOV setting can be enabled or disabled at any time (even while traffic is running).



Settings for VM-101 on WIN-B1IM5UU7CNO

VM-101

- Hardware
  - Add Hardware
  - Firmware
    Boot entry changes pending
  - Security
    Secure Boot enabled
  - Memory
    2048 MB
  - Processor
    1 Virtual processor
  - SCSI Controller
    - Hard Drive
      14393-G2-U.vhdx
  - Network Adapter
    Not connected
  - Network Adapter
    SR-IOV_vSwitch
    - Hardware Acceleration
    - Advanced Features
- Management
  - Name
    VM-101
  - Integration Services
    Some services offered
  - Checkpoints
    Production
  - Smart Paging File Location
    C:\ProgramData\Microsoft\Windo...
  - Automatic Start Action
    Restart if previously running
  - Automatic Stop Action
    Save

Hardware Acceleration

Specify networking tasks that can be offloaded to a physical network adapter.

Virtual machine queue

Virtual machine queue (VMQ) requires a physical network adapter that supports this feature.

☑ Enable virtual machine queue

IPsec task offloading

Support from a physical network adapter and the guest operating system is required to offload IPsec tasks.

When sufficient hardware resources are not available, the security associations are not offloaded and are handled in software by the guest operating system.

☑ Enable IPsec task offloading

Select the maximum number of offloaded security associations from a range of 1 to 4096.

Maximum number: 512 Offloaded SA

Single-root I/O virtualization

Single-root I/O virtualization (SR-IOV) requires specific hardware. It also might require drivers to be installed in the guest operating system.

When sufficient hardware resources are not available, network connectivity is provided through the virtual switch.

☑ Enable SR-IOV

Figure 11-5. Settings for VM: Enabling SR-IOV

13. Install the Cavium QLogic drivers for the adapters detected in the VM. Use the latest drivers available from your vendor for your host OS (do not use inbox drivers).

NOTE

Be sure to use the same driver package on both the VM and the host system. For example, use the same qeVBD and qeND driver version on the Windows VM and in the Windows Hyper-V host.

After installing the drivers, the Cavium QLogic adapter is listed in the VM. Figure 11-6 shows an example.

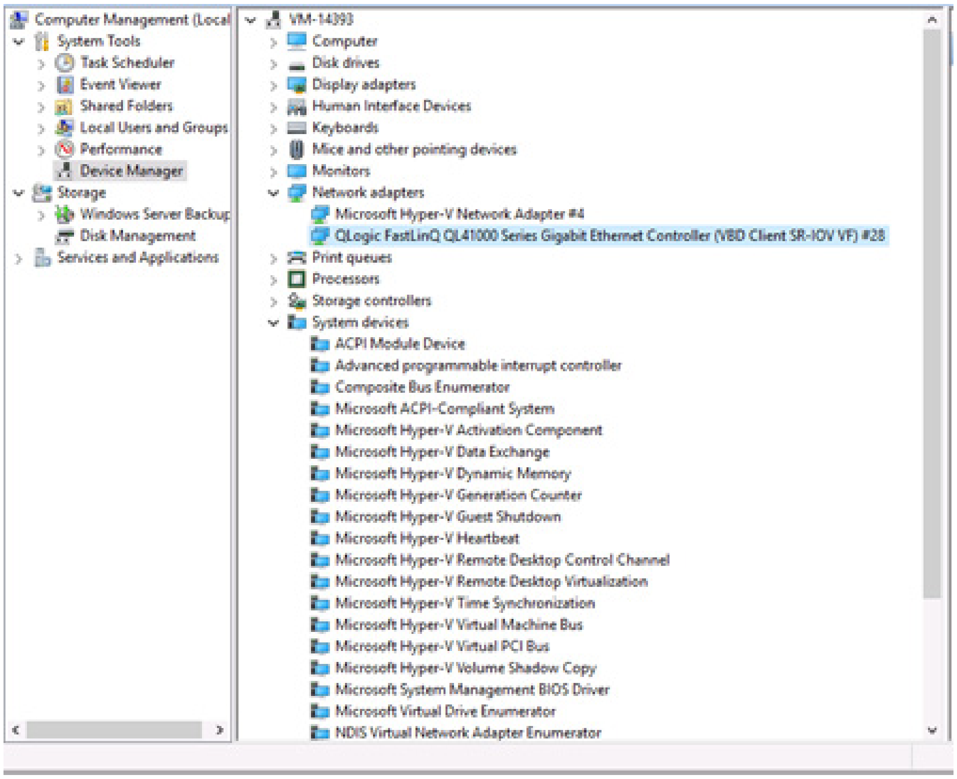
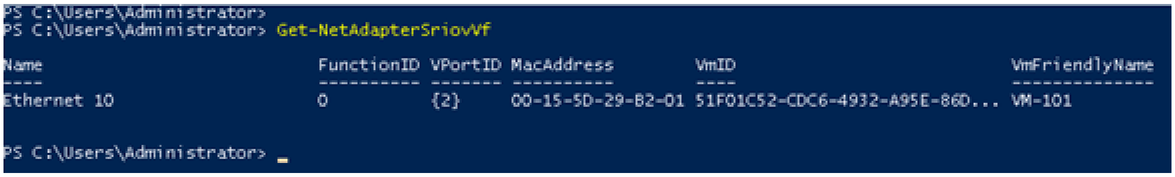

Computer Management (Local
System Tools
Task Scheduler
Event Viewer
Shared Folders
Local Users and Groups
Performance
Device Manager
Storage
Windows Server Backup
Disk Management
Services and Applications
VM-14393
Computer
Disk drives
Display adapters
Human Interface Devices
Keyboards
Mice and other pointing devices
Monitors
Network adapters
Microsoft Hyper-V Network Adapter #4
QLogic FastLinQ QL41000 Series Gigabit Ethernet Controller (VBD Client SR-IOV VF) #28
Print queues
Processors
Storage controllers
System devices
ACPI Module Device
Advanced programmable interrupt controller
Composite Bus Enumerator
Microsoft ACPI-Compliant System
Microsoft Hyper-V Activation Component
Microsoft Hyper-V Data Exchange
Microsoft Hyper-V Dynamic Memory
Microsoft Hyper-V Generation Counter
Microsoft Hyper-V Guest Shutdown
Microsoft Hyper-V Heartbeat
Microsoft Hyper-V Remote Desktop Control Channel
Microsoft Hyper-V Remote Desktop Virtualization
Microsoft Hyper-V Time Synchronization
Microsoft Hyper-V Virtual Machine Bus
Microsoft Hyper-V Virtual PCI Bus
Microsoft Hyper-V Volume Shadow Copy
Microsoft System Management BIOS Driver
Microsoft Virtual Drive Enumerator
NDIS Virtual Network Adapter Enumerator

Figure 11-6. Device Manager: VM with QLogic Adapter

14. To view the SR-IOV VF details, issue the following Windows PowerShell command:

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

Figure 11-7 shows example output.



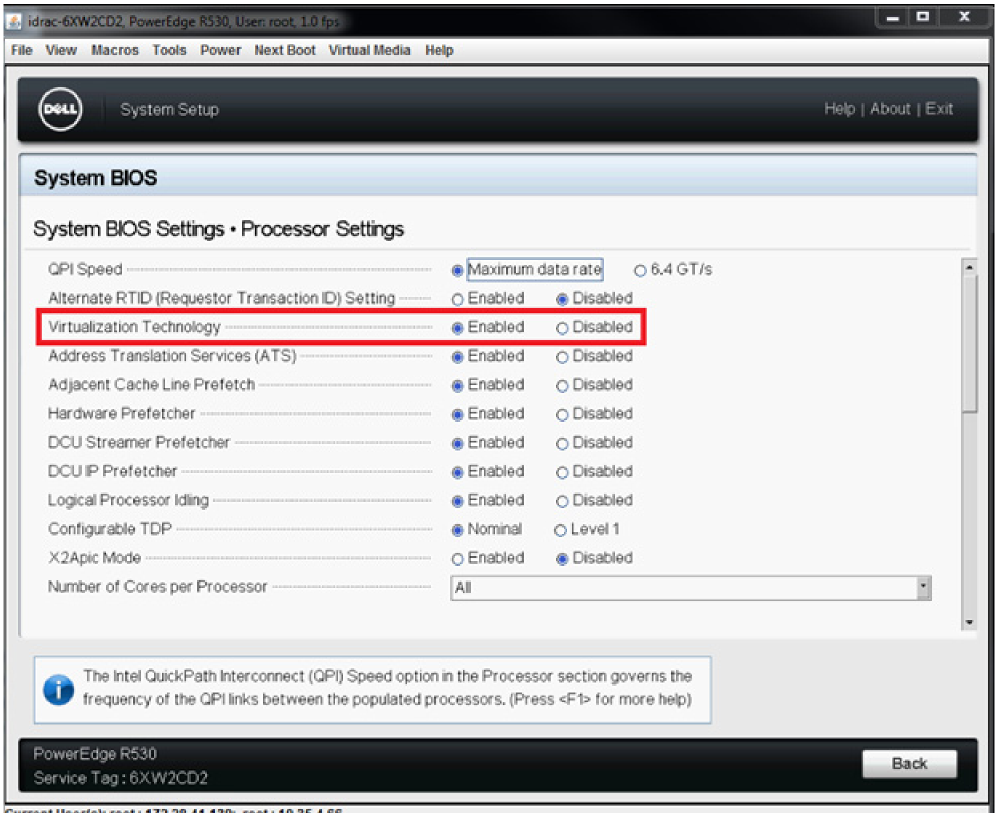
```
PS C:\Users\Administrator>
PS C:\Users\Administrator> Get-NetAdapterSriovVf

Name                                     FunctionID VPortID MacAddress        VmID                                 VmFriendlyName
----                                     ---------- ------- ----------        ----                                 --------------
Ethernet 10                              0          {2}     00-15-5D-29-B2-01 51F01C52-CDC6-4932-A95E-86D...       VM-101

PS C:\Users\Administrator> _
```

Figure 11-7. Windows PowerShell Command: Get-NetadapterSriovVf

Configuring SR-IOV on Linux

To configure SR-IOV on Linux:

1. Access the server BIOS System Setup, and then click **System BIOS Settings**.
2. On the System BIOS Settings page, click **Integrated Devices**.
3. On the System Integrated Devices page (see Figure 11-1 on page 199):
   a. Set the **SR-IOV Global Enable** option to **Enabled**.
   b. Click **Back**.
4. On the System BIOS Settings page, click **Processor Settings**.
5. On the Processor Settings (Figure 11-8) page:
   a. Set the **Virtualization Technology** option to **Enabled**.
   b. Click **Back**.


idrac-6XW2CD2, PowerEdge R530, User: root, 1.0 fps
File View Macros Tools Power Next Boot Virtual Media Help
System Setup
Help | About | Exit
Back

System BIOS

System BIOS Settings • Processor Settings

QPI Speed	● Maximum data rate	○ 6.4 GT/s
Alternate RTID (Requestor Transaction ID) Setting	○ Enabled	● Disabled
Virtualization Technology	● Enabled	○ Disabled
Address Translation Services (ATS)	● Enabled	○ Disabled
Adjacent Cache Line Prefetch	● Enabled	○ Disabled
Hardware Prefetcher	● Enabled	○ Disabled
DCU Streamer Prefetcher	● Enabled	○ Disabled
DCU IP Prefetcher	● Enabled	○ Disabled
Logical Processor Idling	● Enabled	○ Disabled
Configurable TDP	● Nominal	○ Level 1
X2Apic Mode	○ Enabled	● Disabled
Number of Cores per Processor	All	

The Intel QuickPath Interconnect (QPI) Speed option in the Processor section governs the frequency of the QPI links between the populated processors. (Press <F1> for more help)

PowerEdge R530
Service Tag: 6XW2CD2

Figure 11-8. System Setup: Processor Settings for SR-IOV

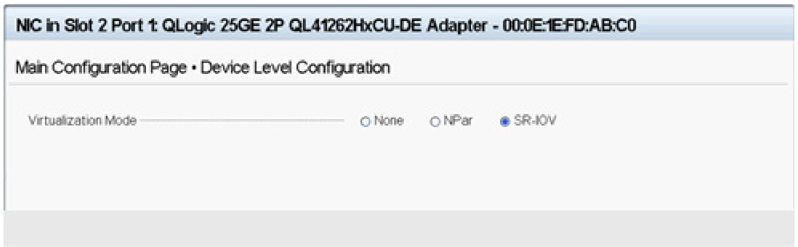
6. On the System Setup page, select **Device Settings**.

7. On the Device Settings page, select **Port 1** for the Cavium QLogic adapter.

8. On the Device Level Configuration page (Figure 11-9):

   a. Set the **Virtualization Mode** to **SR-IOV**.

   b. Click **Back**.



NIC in Slot 2 Port 1: QLogic 25GE 2P QL41262HxCU-DE Adapter - 00:0E:1E:FD:AB:C0

Main Configuration Page • Device Level Configuration

Virtualization Mode ○ None ○ NPar ● SR-IOV

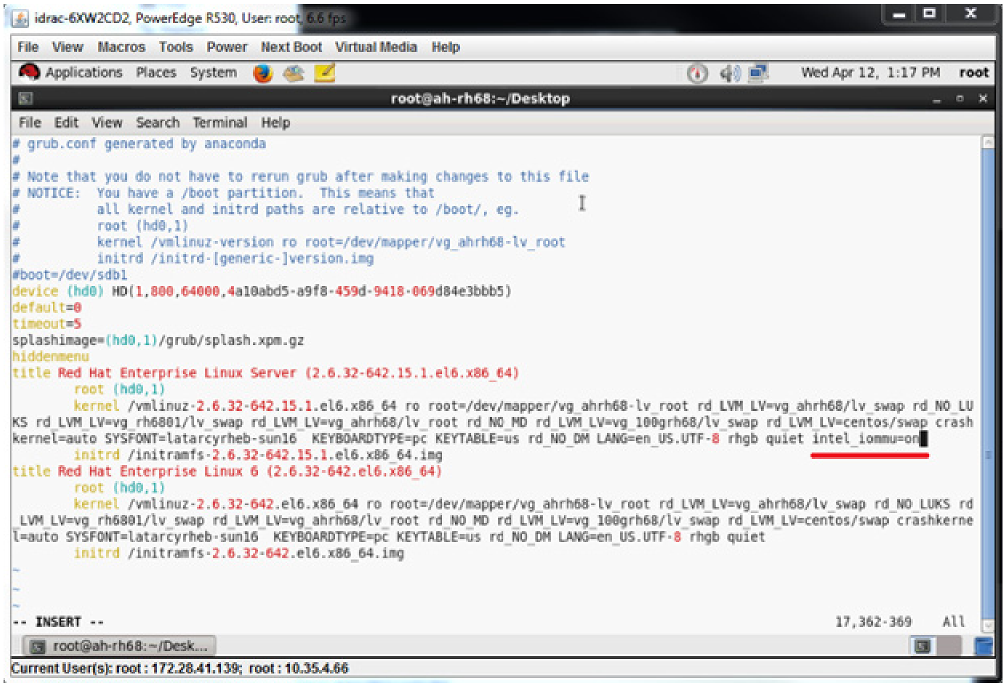
Figure 11-9. System Setup for SR-IOV: Integrated Devices

9. On the Main Configuration Page, click **Finish**, save your settings, and then reboot the system.

10. To enable and verify virtualization:

    a. Open the `grub.conf` file and configure the `iommu` parameter as shown in Figure 11-10. (For details, see “Enabling IOMMU for SR-IOV in UEFI-based Linux OS Installations” on page 210.)

       - For Intel-based systems, add `intel_iommu=on`.
       - For AMD-based systems, add `amd_iommu=on`.


idrac-6XW2CD2, PowerEdge R530, User: root, 6.6 fps
File View Macros Tools Power Next Boot Virtual Media Help
Applications Places System
Wed Apr 12, 1:17 PM root
root@ah-rh68:~/Desktop
File Edit View Search Terminal Help
root@ah-rh68:~/Desk...
Current User(s): root : 172.28.41.139; root : 10.35.4.66

```
# grub.conf generated by anaconda
#
# Note that you do not have to rerun grub after making changes to this file
# NOTICE:  You have a /boot partition.  This means that
#          all kernel and initrd paths are relative to /boot/, eg.
#          root (hd0,1)
#          kernel /vmlinuz-version ro root=/dev/mapper/vg_ahrh68-lv_root
#          initrd /initrd-[generic-]version.img
#boot=/dev/sdb1
device (hd0) HD(1,800,64000,4a10abd5-a9f8-459d-9418-069d84e3bbb5)
default=0
timeout=5
splashimage=(hd0,1)/grub/splash.xpm.gz
hiddenmenu
title Red Hat Enterprise Linux Server (2.6.32-642.15.1.el6.x86_64)
        root (hd0,1)
        kernel /vmlinuz-2.6.32-642.15.1.el6.x86_64 ro root=/dev/mapper/vg_ahrh68-lv_root rd_LVM_LV=vg_ahrh68/lv_swap rd_NO_LU
KS rd_LVM_LV=vg_rh6801/lv_swap rd_LVM_LV=vg_ahrh68/lv_root rd_NO_MD rd_LVM_LV=vg_100grh68/lv_swap rd_LVM_LV=centos/swap crash
kernel=auto SYSFONT=latarcyrheb-sun16  KEYBOARDTYPE=pc KEYTABLE=us rd_NO_DM LANG=en_US.UTF-8 rhgb quiet intel_iommu=on
        initrd /initramfs-2.6.32-642.15.1.el6.x86_64.img
title Red Hat Enterprise Linux 6 (2.6.32-642.el6.x86_64)
        root (hd0,1)
        kernel /vmlinuz-2.6.32-642.el6.x86_64 ro root=/dev/mapper/vg_ahrh68-lv_root rd_LVM_LV=vg_ahrh68/lv_swap rd_NO_LUKS rd
_LVM_LV=vg_rh6801/lv_swap rd_LVM_LV=vg_ahrh68/lv_root rd_NO_MD rd_LVM_LV=vg_100grh68/lv_swap rd_LVM_LV=centos/swap crashkerne
l=auto SYSFONT=latarcyrheb-sun16  KEYBOARDTYPE=pc KEYTABLE=us rd_NO_DM LANG=en_US.UTF-8 rhgb quiet
        initrd /initramfs-2.6.32-642.el6.x86_64.img
~
~
~
-- INSERT --                                                                                 17,362-369    All
```

Figure 11-10. Editing the grub.conf File for SR-IOV

b. Save the `grub.conf` file and then reboot the system.

c. To verify that the changes are in effect, issue the following command:

```
dmesg | grep  -I  iommu
```

A successful input–output memory management unit (IOMMU) command output should show, for example:

```
Intel-IOMMU: enabled
```

d. To view VF details (number of VFs and total VFs), issue the following command:

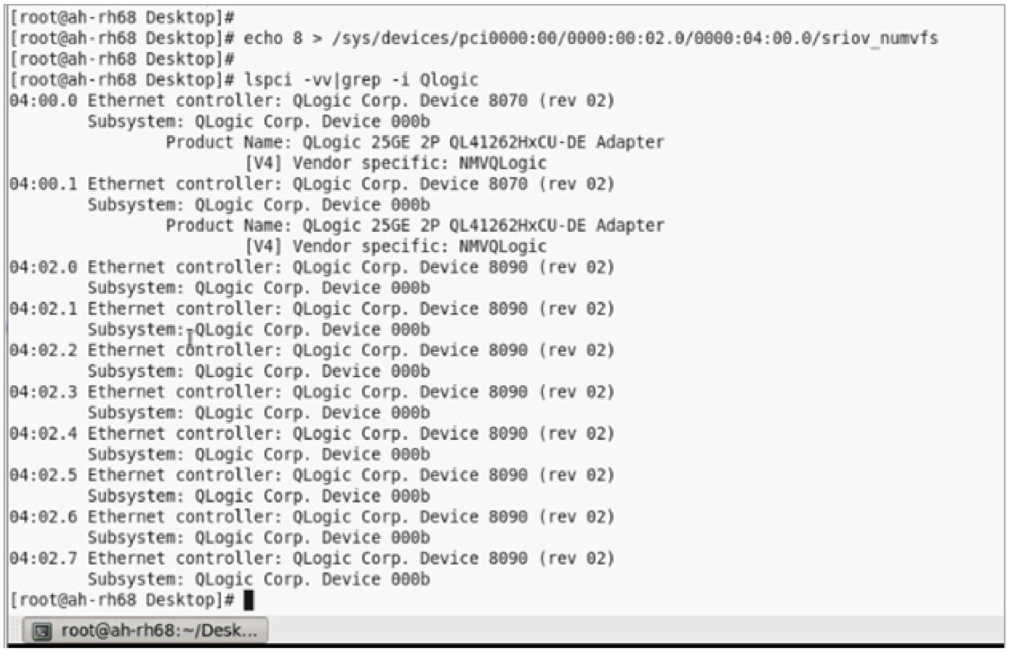
```
find /sys/|grep -I sriov
```

11. For a specific port, enable a quantity of VFs.

a. Issue the following command to enable, for example, 8 VFs on PCI instance 04:00.0 (bus 4, device 0, function 0):

```
[root@ah-rh68 ~]# echo 8 >
/sys/devices/pci0000:00/0000:00:02.0/0000:04:00.0/
sriov_numvfs
```

b. Review the command output (Figure 11-11) to confirm that actual VFs were created on bus 4, device 2 (from the 0000:00:02.0 parameter), functions 0 through 7. Note that the actual device ID is different on the PFs (8070 in this example) versus the VFs (8090 in this example).



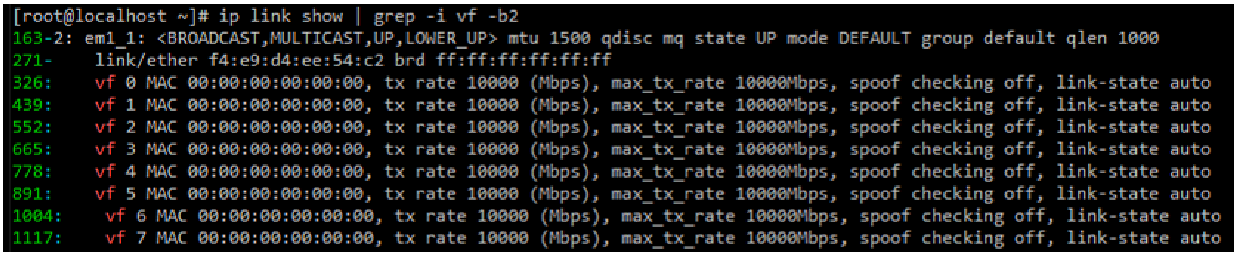
```
[root@ah-rh68 Desktop]#
[root@ah-rh68 Desktop]# echo 8 > /sys/devices/pci0000:00/0000:00:02.0/0000:04:00.0/sriov_numvfs
[root@ah-rh68 Desktop]#
[root@ah-rh68 Desktop]# lspci -vv|grep -i Qlogic
04:00.0 Ethernet controller: QLogic Corp. Device 8070 (rev 02)
        Subsystem: QLogic Corp. Device 000b
                Product Name: QLogic 25GE 2P QL41262HxCU-DE Adapter
                        [V4] Vendor specific: NMVQLogic
04:00.1 Ethernet controller: QLogic Corp. Device 8070 (rev 02)
        Subsystem: QLogic Corp. Device 000b
                Product Name: QLogic 25GE 2P QL41262HxCU-DE Adapter
                        [V4] Vendor specific: NMVQLogic
04:02.0 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.1 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.2 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.3 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.4 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.5 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.6 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
04:02.7 Ethernet controller: QLogic Corp. Device 8090 (rev 02)
        Subsystem: QLogic Corp. Device 000b
[root@ah-rh68 Desktop]# █
```

Figure 11-11. Command Output for sriov_numvfs

12. To view a list of all PF and VF interfaces, issue the following command:

```
# ip link show | grep -i vf -b2
```

Figure 11-12 shows example output.



```
[root@localhost ~]# ip link show | grep -i vf -b2
163-2: em1_1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP mode DEFAULT group default qlen 1000
271-    link/ether f4:e9:d4:ee:54:c2 brd ff:ff:ff:ff:ff:ff
326:    vf 0 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
439:    vf 1 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
552:    vf 2 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
665:    vf 3 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
778:    vf 4 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
891:    vf 5 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
1004:    vf 6 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
1117:    vf 7 MAC 00:00:00:00:00:00, tx rate 10000 (Mbps), max_tx_rate 10000Mbps, spoof checking off, link-state auto
```

Figure 11-12. Command Output for ip link show Command

13. Assign and verify MAC addresses:

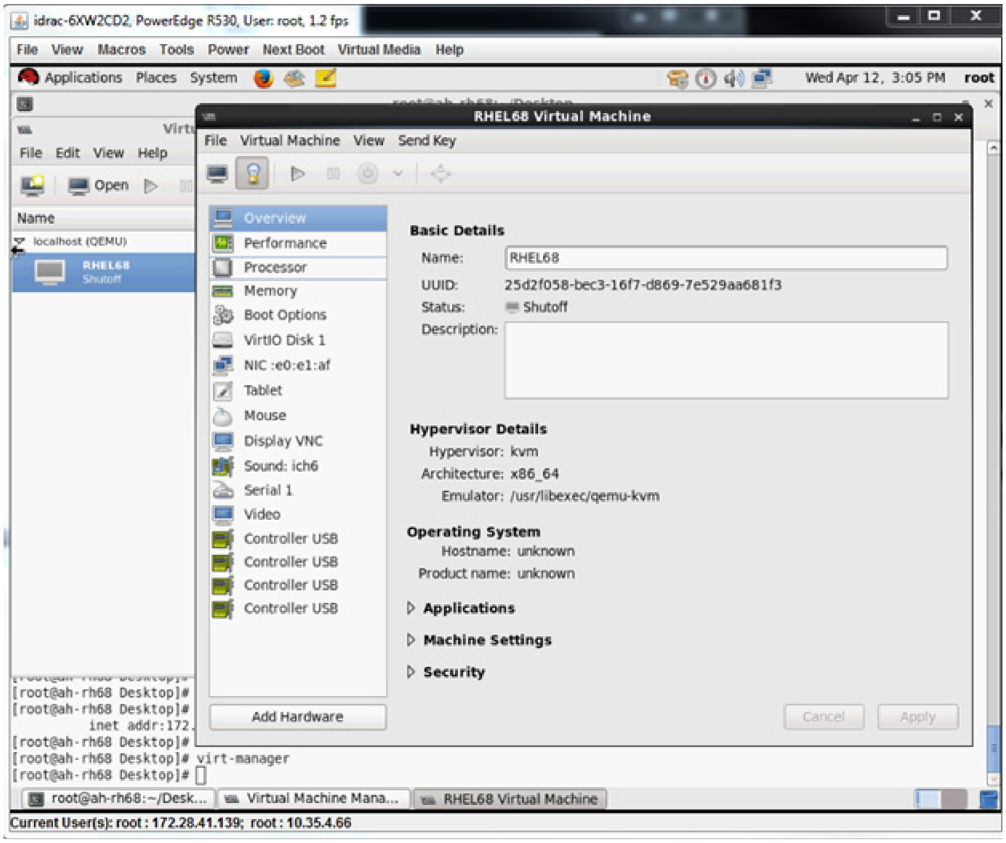
a. To assign a MAC address to the VF, issue the following command:

```
ip link set <pf device> vf <vf index> mac <mac address>
```

b. Ensure that the VF interface is up and running with the assigned MAC address.

14. Power off the VM and attach the VF. (Some OSs support hot-plugging of VFs to the VM.)

a. In the Virtual Machine dialog box (Figure 11-13), click **Add Hardware**.



Basic Details

Name: RHEL68

UUID: 25d2f058-bec3-16f7-d869-7e529aa681f3

Status: Shutoff

Description:

Hypervisor Details

Hypervisor: kvm

Architecture: x86_64

Emulator: /usr/libexec/qemu-kvm

Operating System

Hostname: unknown

Product name: unknown

```
[root@ah-rh68 Desktop]#
[root@ah-rh68 Desktop]#
          inet addr:172.
[root@ah-rh68 Desktop]#
[root@ah-rh68 Desktop]# virt-manager
[root@ah-rh68 Desktop]#
```

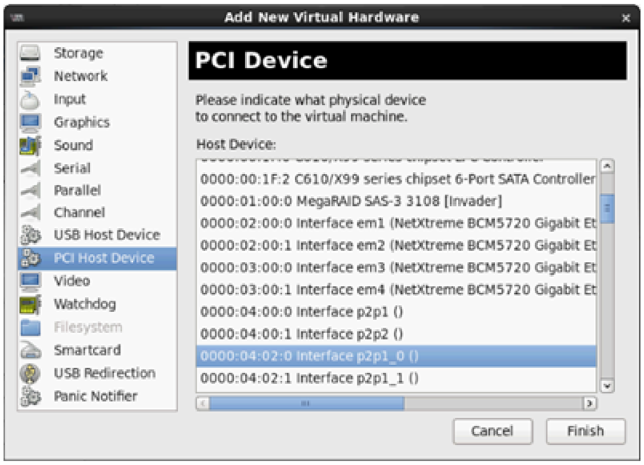
Current User(s): root : 172.28.41.139; root : 10.35.4.66

Figure 11-13. RHEL68 Virtual Machine

b. In the left pane of the Add New Virtual Hardware dialog box (Figure 11-14), click **PCI Host Device**.

c. In the right pane, select a host device.

d. Click **Finish**.



PCI Device

Please indicate what physical device
to connect to the virtual machine.

Host Device:

```
0000:00:1F:2 C610/X99 series chipset 6-Port SATA Controller
0000:01:00:0 MegaRAID SAS-3 3108 [Invader]
0000:02:00:0 Interface em1 (NetXtreme BCM5720 Gigabit Et
0000:02:00:1 Interface em2 (NetXtreme BCM5720 Gigabit Et
0000:03:00:0 Interface em3 (NetXtreme BCM5720 Gigabit Et
0000:03:00:1 Interface em4 (NetXtreme BCM5720 Gigabit Et
0000:04:00:0 Interface p2p1 ()
0000:04:00:1 Interface p2p2 ()
0000:04:02:0 Interface p2p1_0 ()
0000:04:02:1 Interface p2p1_1 ()
```

Figure 11-14. Add New Virtual Hardware

15. Power on the VM, and then issue the following command:

    ```
    check lspci -vv|grep -I ether
    ```

16. Install the drivers for the adapters detected in the VM. Use the latest drivers available from your vendor for your host OS (do not use inbox drivers). The same driver version must be installed on the host and the VM.

17. As needed, add more VFs in the VM.

Enabling IOMMU for SR-IOV in UEFI-based Linux OS Installations

Follow the appropriate procedure for your Linux OS.

NOTE

For AMD systems, replace `intel_iommu=on` with `amd_iommu=on`.

To enable IOMMU for SR-IOV on Red Hat 6.*x*:

- In the `/boot/efi/EFI/redhat/grub.conf` file, locate the kernel line, and then add the `intel_iommu=on` boot parameter.

To enable IOMMU for SR-IOV on Red Hat 7.*x*:

1. In the `/etc/default/grub` file, locate `GRUB_CMDLINE_LINUX`, and then add the `intel_iommu=on` boot parameter.

2. To update the grub configuration file, issue the following command:

   ```
   GRUB2-MKCONFIG -O /BOOT/EFI/EFI/REDHAT/GRUB.CFG
   ```

To enable IOMMU for SR-IOV on SUSE 12.*x*:

1. In the `/etc/default/grub` file, locate `GRUB_CMDLINE_LINUX_DEFAULT`, and then add the `intel_iommu=on` boot parameter.
2. To update the grub configuration file, issue the following command:

```
grub2-mkconfig -o /boot/grub2/grub.cfg
```

Configuring SR-IOV on VMware

To configure SR-IOV on VMware:

1. Access the server BIOS System Setup, and then click **System BIOS Settings**.
2. On the System BIOS Settings page, click **Integrated Devices**.
3. On the Integrated Devices page (see Figure 11-1 on page 199):
   a. Set the **SR-IOV Global Enable** option to **Enabled**.
   b. Click **Back**.
4. In the System Setup window, click **Device Settings**.
5. On the Device Settings page, select a port for the 25G 41*xxx* Series Adapter.
6. On the Device Level Configuration (see Figure 11-2 on page 199):
   a. Set the **Virtualization Mode** to **SR-IOV**.
   b. Click **Back**.
7. On the Main Configuration Page, click **Finish**.
8. Save the configuration settings and reboot the system.
9. To enable the needed quantity of VFs per port (in this example, 16 on each port of a dual-port adapter), issue the following command:

```
"esxcfg-module -s "max_vfs=16,16" qedentv"
```

NOTE

Each Ethernet function of the 41*xxx* Series Adapter must have its own entry.

10. Reboot the host.
11. To verify that the changes are complete at the module level, issue the following command:

```
"esxcfg-module -g qedentv"
```

```
[root@localhost:~] esxcfg-module -g qedentv
qedentv enabled = 1 options = 'max_vfs=16,16'
```

12. To verify if actual VFs were created, issue the `lspci` command as follows:

```
[root@localhost:~] lspci | grep -i QLogic | grep -i 'ethernet\|network' | more
0000:05:00.0 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx 10/25
GbE Ethernet Adapter [vmnic6]
0000:05:00.1 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx 10/25
GbE Ethernet Adapter [vmnic7]
0000:05:02.0 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_0]
0000:05:02.1 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_1]
0000:05:02.2 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_2]
0000:05:02.3 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_3]
.
.
.
0000:05:03.7 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_15]
0000:05:0e.0 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_0]
0000:05:0e.1 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_1]
0000:05:0e.2 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_2]
0000:05:0e.3 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_3]
.
.
.
0000:05:0f.6 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_14]
0000:05:0f.7 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_15]
```

13. Attach VFs to the VM as follows:

    a. Power off the VM and attach the VF. (Some OSs support hot-plugging of VFs to the VM.)

    b. Add a host to a VMware vCenter Server Virtual Appliance (vCSA).

   c. Click **Edit Settings** of the VM.

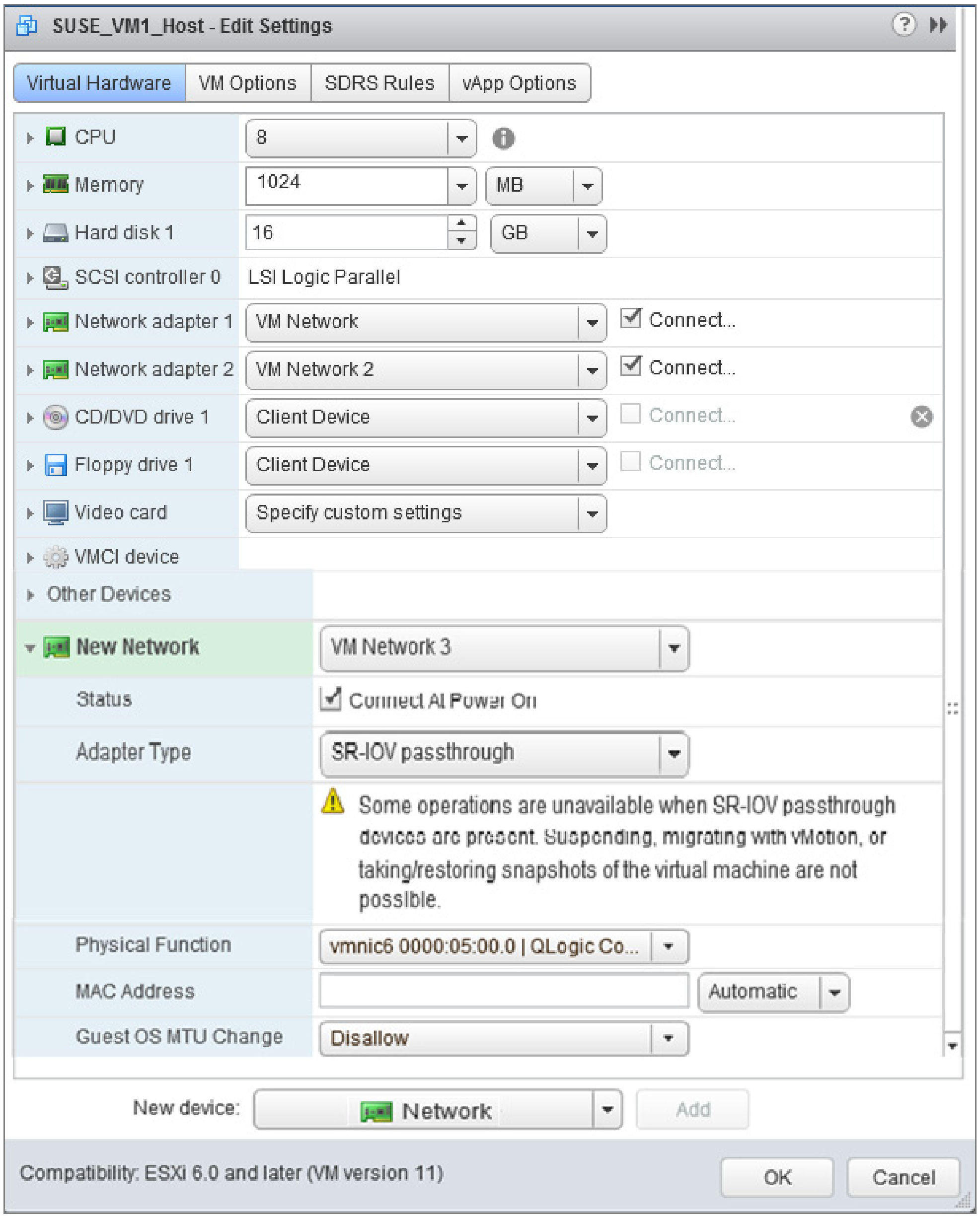
14. Complete the Edit Settings dialog box (Figure 11-15) as follows:

   a. In the **New Device** box, select **Network**, and then click **Add**.

   b. For **Adapter Type**, select **SR-IOV Passthrough**.

   c. For **Physical Function**, select the Cavium QLogic VF.

   d. To save your configuration changes and close this dialog box, click **OK**.



SUSE_VM1_Host - Edit Settings

CPU	8		
Memory	1024	MB	
Hard disk 1	16	GB	
SCSI controller 0	LSI Logic Parallel		
Network adapter 1	VM Network		☑ Connect...
Network adapter 2	VM Network 2		☑ Connect...
CD/DVD drive 1	Client Device		□ Connect...
Floppy drive 1	Client Device		□ Connect...
Video card	Specify custom settings		
VMCI device			
Other Devices			
New Network	VM Network 3		
Status	☑ Connect At Power On		
Adapter Type	SR-IOV passthrough		
	⚠ Some operations are unavailable when SR-IOV passthrough devices are present. Suspending, migrating with vMotion, or taking/restoring snapshots of the virtual machine are not possible.		
Physical Function	vmnic6 0000:05:00.0 | QLogic Co...		
MAC Address		Automatic	
Guest OS MTU Change	Disallow		

Compatibility: ESXi 6.0 and later (VM version 11)

Figure 11-15. VMware Host Edit Settings

15. To validate the VFs per port, issue the `esxcli` command as follows:

```
[root@localhost:~] esxcli network sriovnic vf list -n vmnic6
VF ID  Active  PCI Address  Owner World ID
-----  ------  -----------  --------------
    0    true  005:02.0     60591
    1    true  005:02.1     60591
```

```
 2   false  005:02.2      -
 3   false  005:02.3      -
 4   false  005:02.4      -
 5   false  005:02.5      -
 6   false  005:02.6      -
 7   false  005:02.7      -
 8   false  005:03.0      -
 9   false  005:03.1      -
10   false  005:03.2      -
11   false  005:03.3      -
12   false  005:03.4      -
13   false  005:03.5      -
14   false  005:03.6      -
15   false  005:03.7      -
```

16. Install the Cavium QLogic drivers for the adapters detected in the VM. Use the latest drivers available from your vendor for your host OS (do not use inbox drivers). The same driver version must be installed on the host and the VM.
17. Power on the VM, and then issue the **`ifconfig -a`** command to verify that the added network interface is listed.
18. As needed, add more VFs in the VM.

12 NVMe-oF Configuration with RDMA

Non-Volatile Memory Express over Fabrics (NVMe-oF) enables the use of alternate transports to PCIe to extend the distance over which an NVMe host device and an NVMe storage drive or subsystem can connect. NVMe-oF defines a common architecture that supports a range of storage networking fabrics for the NVMe block storage protocol over a storage networking fabric. This architecture includes enabling a front-side interface into storage systems, scaling out to large quantities of NVMe devices, and extending the distance within a data center over which NVMe devices and NVMe subsystems can be accessed.

The NVMe-oF configuration procedures and options described in this chapter apply to Ethernet-based RDMA protocols, including RoCE and iWARP. The development of NVMe-oF with RDMA is defined by a technical sub-group of the NVMe organization.

This chapter demonstrates how to configure NVMe-oF for a simple network. The example network comprises the following:

- Two servers: an initiator and a target. The target server is equipped with a PCIe SSD drive.
- Operating system: RHEL 7.4 or SLES 12 SP3 on both servers.
- Two adapters: One 41*xxx* Series Adapter installed in each server. Each port can be independently configured to use RoCE, RoCEv2, or iWARP as the RDMA protocol over which NVMe-oF runs.
- For RoCE and RoCEv2, an optional switch configured for data center bridging (DCB), relevant quality of service (QoS) policy, and VLANs to carry the NVMe-oF's RoCE/RoCEv2 DCB traffic class priority. The switch is not needed when NVMe-oF is using iWARP.

Figure 12-1 illustrates an example network.

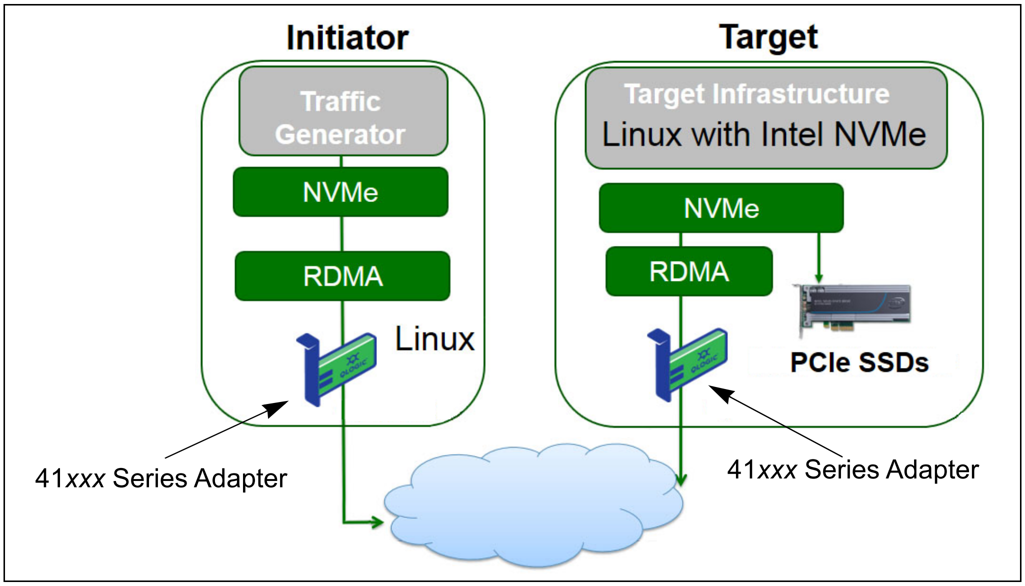


Figure 12-1. NVMe-oF Network

The NVMe-oF configuration process covers the following procedures:

- Installing Device Drivers on Both Servers
- Configuring the Target Server
- Configuring the Initiator Server
- Preconditioning the Target Server
- Testing the NVMe-oF Devices
- Optimizing Performance

Installing Device Drivers on Both Servers

After installing your operating system (RHEL 7.4 or SLES 12 SP3), install device drivers on both servers. To upgrade the kernel to the latest Linux upstream kernel, go to:

https://www.kernel.org/pub/linux/kernel/v4.x/

1. Install and load the latest FastLinQ drivers (qed, qede, libqedr/qedr) following all installation instructions in the README.
2. (Optional) If you upgraded the OS kernel, you must reinstall and load the latest driver as follows:
   a. Install the latest FastLinQ firmware following all installation instructions in the README.
   b. Install the OS RDMA support applications and libraries by issuing the following commands:

      ```
      # yum groupinstall "Infiniband Support"
      # yum install tcl-devel libibverbs-devel libnl-devel
      glib2-devel libudev-devel lsscsi perftest
      # yum install gcc make git ctags ncurses ncurses-devel
      openssl* openssl-devel elfutils-libelf-devel*
      ```

   c. To ensure that NVMe OFED support is in the selected OS kernel, issue the following command:

      ```
      make menuconfig
      ```

   d. Under **Device Drivers**, ensure that the following are enabled (set to **M**):

      ```
      NVM Express block devices
      NVM Express over Fabrics RDMA host driver
      NVMe Target support
      NVMe over Fabrics RDMA target support
      ```

   e. (Optional) If the **Device Drivers** options are not already present, rebuild the kernel by issuing the following commands:

      ```
      # make
      # make modules
      # make modules_install
      # make install
      ```

   f. If changes were made to the kernel, reboot to that new OS kernel. For instructions on how to set the default boot kernel, go to:

      https://wiki.centos.org/HowTos/Grub2

3. Enable and start the RDMA service as follows:

```
# systemctl enable rdma.service
# systemctl start rdma.service
```

   Disregard the `RDMA Service Failed` error. All OFED modules required by qedr are already loaded.

Configuring the Target Server

Configure the target server after the reboot process. After the server is operating, you cannot change the configuration without rebooting. If you are using a startup script to configure the target server, consider pausing the script (using the `wait` command or something similar) as needed to ensure that each command finishes before executing the next command.

To configure the target service:

1. Load target modules. Issue the following commands after each server reboot:

```
# modprobe qedr
# modprobe nvmet; modprobe nvmet-rdma
# lsmod | grep nvme  (confirm that the modules are loaded)
```

2. Create the target subsystem NVMe Qualified Name (NQN) with the name indicated by `<nvme-subsystem-name>`. Use the NVMe-oF specifications; for example, `nqn.<YEAR>-<Month>.org.<your-company>`.

```
# mkdir /sys/kernel/config/nvmet/subsystems/<nvme-subsystem-name>
# cd /sys/kernel/config/nvmet/subsystems/<nvme-subsystem-name>
```

3. Create multiple unique NQNs for additional NVMe devices as needed.

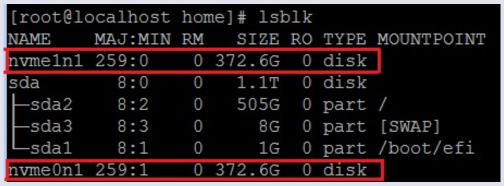
4. Set the target parameters, as listed in Table 12-1.

Table 12-1. Target Parameters

Command	Description
`# echo 1 > attr_allow_any_host`	Allows any host to connect.
`# mkdir namespaces/1`	Creates a namespace.

Table 12-1. Target Parameters (Continued)

Command	Description

`# echo -n /dev/nvme0n1 >namespaces/1/device_path`	Sets the NVMe device path. The NVMe device path can differ between systems. Check the device path using the `lsblk` command. This system has two NVMe devices: `nvme0n1` and `nvme1n1`.
	
	`[root@localhost home]# lsblk`
	`NAME    MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT`
	`nvme1n1 259:0    0 372.6G  0 disk`
	`sda       8:0    0   1.1T  0 disk`
	`├─sda2    8:2    0   505G  0 part /`
	`├─sda3    8:3    0     8G  0 part [SWAP]`
	`└─sda1    8:1    0     1G  0 part /boot/efi`
	`nvme0n1 259:1    0 372.6G  0 disk`

`# echo 1 > namespaces/1/enable`	Enables the namespace.

`# mkdir /sys/kernel/config/nvmet/ports/1`
`# cd /sys/kernel/config/nvmet/ports/1`
	Creates NVMe port 1.

`# echo 1.1.1.1 > addr_traddr`	Sets the same IP address. For example, 1.1.1.1 is the IP address for the target port of the 41*xxx* Series Adapter.
`# echo rdma > addr_trtype`	Sets the transport type RDMA.
`# echo 4420 > addr_trsvcid`	Sets the RDMA port number. The socket port number for NVMe-oF is typically 4420. However, any port number can be used if it is used consistently throughout the configuration.
`# echo ipv4 > addr_adrfam`	Sets the IP address type.

5. Create a symbolic link (symlink) to the newly created NQN subsystem:

   ```
   # ln -s /sys/kernel/config/nvmet/subsystems/
   nvme-subsystem-name subsystems/nvme-subsystem-name
   ```

6. Confirm that the NVMe target is listening on the port as follows:

   ```
   # dmesg | grep nvmet_rdma
   [ 8769.470043] nvmet_rdma: enabling port 1 (1.1.1.1:4420)
   ```

Configuring the Initiator Server

You must configure the initiator server after the reboot process. After the server is operating, you cannot change the configuration without rebooting. If you are using a startup script to configure the initiator server, consider pausing the script (using the `wait` command or something similar) as needed to ensure that each command finishes before executing the next command.

To configure the initiator server:

1. Load the NVMe modules. Issue these commands after each server reboot:

```
# modprobe qedr
# modprobe nvme-rdma
```

2. Download, compile and install the `nvme-cli` initiator utility. Issue these commands at the first configuration—you do not need to issue these commands after each reboot.

```
# git clone https://github.com/linux-nvme/nvme-cli.git
# cd nvme-cli
# make && make install
```

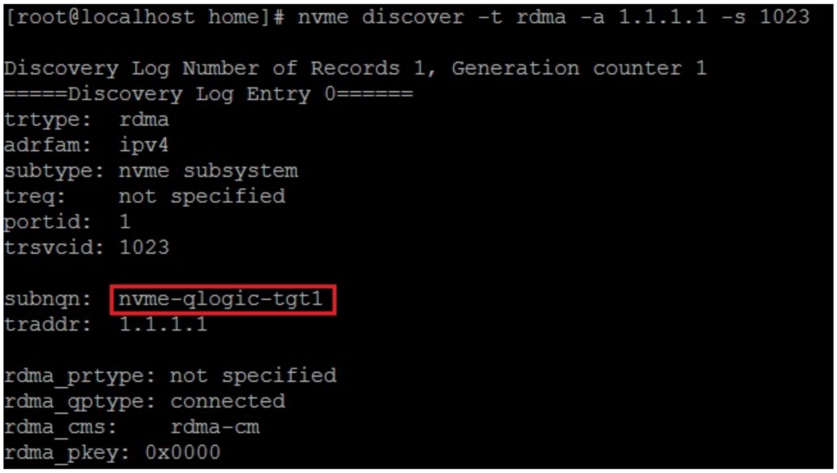
3. Verify the installation version as follows:

```
# nvme version
```

4. Discover the NVMe-oF target as follows:

```
# nvme discover -t rdma -a 1.1.1.1 -s 1023
```

Make note of the subsystem NQN (`subnqn`) of the discovered target (Figure 12-2) for use in Step 5.



```
[root@localhost home]# nvme discover -t rdma -a 1.1.1.1 -s 1023

Discovery Log Number of Records 1, Generation counter 1
=====Discovery Log Entry 0======
trtype:  rdma
adrfam:  ipv4
subtype: nvme subsystem
treq:    not specified
portid:  1
trsvcid: 1023

subnqn:  nvme-qlogic-tgt1
traddr:  1.1.1.1

rdma_prtype: not specified
rdma_qptype: connected
rdma_cms:    rdma-cm
rdma_pkey: 0x0000
```

Figure 12-2. Subsystem NQN

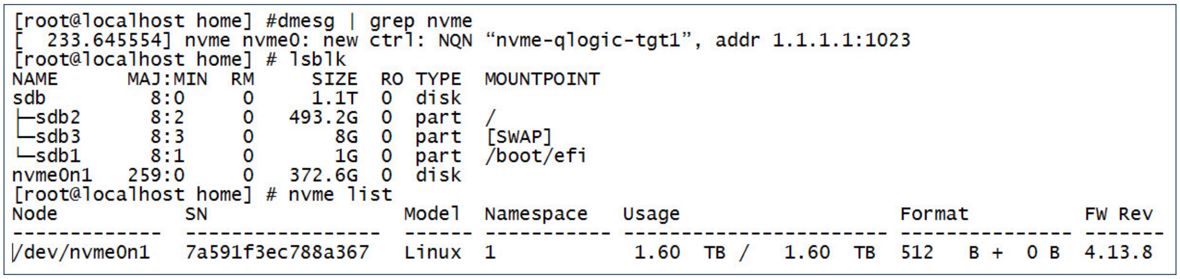
5. Connect to the discovered NVMe-oF target (`nvme-qlogic-tgt1`) using the NQN. Issue the following command after each server reboot. For example:

   ```
   # nvme connect -t rdma -n nvme-qlogic-tgt1 -a 1.1.1.1 -s 1023
   ```

6. Confirm the NVMe-oF target connection with the NVMe-oF device as follows:

   ```
   # dmesg | grep nvme
   # lsblk
   # list nvme
   ```

   Figure 12-3 shows an example.



```
[root@localhost home] #dmesg | grep nvme
[  233.645554] nvme nvme0: new ctrl: NQN “nvme-qlogic-tgt1”, addr 1.1.1.1:1023
[root@localhost home] # lsblk
NAME      MAJ:MIN  RM     SIZE  RO TYPE  MOUNTPOINT
sdb         8:0     0     1.1T  0  disk
├─sdb2      8:2     0   493.2G  0  part  /
└─sdb3      8:3     0       8G  0  part  [SWAP]
└─sdb1      8:1     0       1G  0  part  /boot/efi
nvme0n1   259:0     0   372.6G  0  disk
[root@localhost home] # nvme list
Node             SN                   Model  Namespace  Usage                      Format           FW Rev
---------------- -------------------- ------ ---------- -------------------------- ---------------- -------
/dev/nvme0n1     7a591f3ec788a367     Linux  1            1.60   TB /   1.60   TB  512   B +  0 B   4.13.8
```

Figure 12-3. Confirm NVMe-oF Connection

Preconditioning the Target Server

NVMe target servers that are tested out-of-the-box show a higher-than-expected performance. Before running a benchmark, the target server needs to be *prefilled* or *preconditioned*.

To precondition the target server:

1. Secure-erase the target server with vendor-specific tools (similar to formatting). This test example uses an Intel NVMe SSD device, which requires the Intel Data Center Tool that is available at the following link:

   https://downloadcenter.intel.com/download/23931/Intel-Solid-State-Drive-Data-Center-Tool

2. Precondition the target server (`nvme0n1`) with data, which guarantees that all available memory is filled. This example uses the “DD” disk utility:

   ```
   # dd if=/dev/zero bs=1024k of=/dev/nvme0n1
   ```

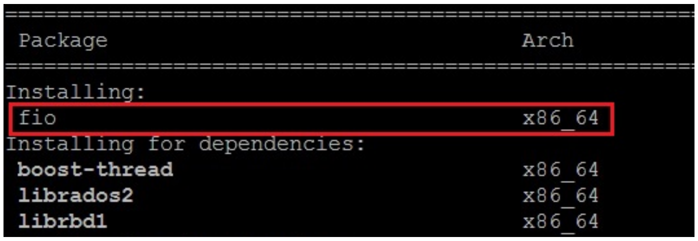
Testing the NVMe-oF Devices

Compare the latency of the local NVMe device on the target server with that of the NVMe-oF device on the initiator server to show the latency that NVMe adds to the system.

To test the NVMe-oF device:

1. Update the Repository (Repo) source and install the Flexible Input/Output (FIO) benchmark utility on both the target and initiator servers by issuing the following commands:

```
# yum install epel-release
# yum install fio
```



```
==============================================================
 Package                                          Arch
==============================================================
Installing:
 fio                                              x86_64
Installing for dependencies:
 boost-thread                                     x86_64
 librados2                                        x86_64
 librbd1                                          x86_64
```

Figure 12-4. FIO Utility Installation

2. Run the FIO utility to measure the latency of the initiator NVMe-oF device. Issue the following command:

```
# fio --filename=/dev/nvme0n1 --direct=1 --time_based
--rw=randread --refill_buffers --norandommap --randrepeat=0
--ioengine=libaio --bs=4k --iodepth=1 --numjobs=1
--runtime=60 --group_reporting --name=temp.out
```

   FIO reports two latency types: submission and completion. Submission latency (slat) measures application-to-kernel latency. Completion latency (clat), measures end-to-end kernel latency. The industry-accepted method is to read *clat percentiles* in the 99.00th range.

   In this example, the initiator device NVMe-oF latency is 30µsec.

3. Run FIO to measure the latency of the local NVMe device on the target server. Issue the following command:

```
# fio --filename=/dev/nvme0n1 --direct=1 --time_based
--rw=randread --refill_buffers --norandommap --randrepeat=0
--ioengine=libaio --bs=4k --iodepth=1 --numjobs=1
--runtime=60 --group_reporting --name=temp.out
```

In this example, the target NVMe device latency is 8µsec. The total latency that results from the use of NVMe-oF is the difference between the initiator device NVMe-oF latency (30µsec) and the target device NVMe-oF latency (8µsec), or 22µsec.

4. Run FIO to measure bandwidth of the local NVMe device on the target server. Issue the following command:

```
fio --verify=crc32 --do_verify=1 --bs=8k --numjobs=1
--iodepth=32 --loops=1 --ioengine=libaio --direct=1
--invalidate=1 --fsync_on_close=1 --randrepeat=1
--norandommap --time_based --runtime=60
--filename=/dev/nvme0n1 --name=Write-BW-to-NVMe-Device
--rw=randwrite
```

Where `--rw` can be `randread` for reads only, `randwrite` for writes only, or `randrw` for reads and writes.

Optimizing Performance

To optimize performance on both initiator and target servers:

1. Configure the following system BIOS settings:
   - ❑ Power Profiles = 'Max Performance' or equivalent
   - ❑ ALL C-States = Disabled
   - ❑ Hyperthreading = Disabled

2. Configure the Linux kernel parameters by editing the `grub` file (`/etc/default/grub`).

   a. Add parameters to end of line `GRUB_CMDLINE_LINUX`:

   ```
   GRUB_CMDLINE_LINUX="nosoftlockup intel_idle.max_cstate=0
   processor.max_cstate=1 mce=ignore_ce idle=poll"
   ```

   b. Save the `grub` file.

   c. Rebuild the `grub` file.

   - ■ To rebuild the `grub` file for a legacy BIOS boot, issue the following command:

     `# grub2-mkconfig -o /boot/grub2/grub.cfg` (Legacy BIOS boot)

   - ■ To rebuild the `grub` file for an EFI boot, issue the following command:

     `# grub2-mkconfig -o /boot/efi/EFI/<os>/grub.cfg` (EFI boot)

   d. Reboot the server to implement the changes.

3. Set the IRQ affinity for all 41*xxx* Series Adapters. The `multi_rss-affin.sh` file is a script file that is listed in “.IRQ Affinity (multi_rss-affin.sh)” on page 225.

   ```
   # systemctl stop irqbalance
   # ./multi_rss-affin.sh eth1
   ```

   NOTE

   A different version of this script, `qedr_affin.sh`, is in the 41*xxx* Linux Source Code Package in the `\add-ons\performance\roce` directory. For an explanation of the IRQ affinity settings, refer to the `multiple_irqs.txt` file in that directory.

4. Set the CPU frequency. The `cpufreq.sh` file is a script that is listed in “CPU Frequency (cpufreq.sh)” on page 226.

   ```
   # ./cpufreq.sh
   ```

The following sections list the scripts that are used in Steps 3 and 4.

.IRQ Affinity (multi_rss-affin.sh)

The following script sets the IRQ affinity.

```
#!/bin/bash
#RSS affinity setup script
#input: the device name (ethX)
#OFFSET=0   0/1   0/1/2   0/1/2/3
#FACTOR=1    2      3       4
OFFSET=0
FACTOR=1
LASTCPU='cat /proc/cpuinfo | grep processor | tail -n1 | cut -d":" -f2'
MAXCPUID='echo 2 $LASTCPU ^  p | dc'
OFFSET='echo 2 $OFFSET ^  p | dc'
FACTOR='echo 2 $FACTOR ^  p | dc'
CPUID=1

for eth in $*; do

NUM='grep $eth /proc/interrupts | wc -l'
NUM_FP=$((${NUM}))

INT='grep -m 1 $eth /proc/interrupts | cut -d ":" -f 1'

echo  "$eth: ${NUM} (${NUM_FP} fast path) starting irq ${INT}"
```

```
CPUID=$((CPUID*OFFSET))
for ((A=1; A<=${NUM_FP}; A=${A}+1)) ; do
INT='grep -m $A $eth /proc/interrupts | tail -1  | cut -d ":" -f 1'
SMP='echo $CPUID 16 o p | dc'
echo ${INT} smp affinity set to ${SMP}
echo  $((${SMP})) > /proc/irq/$((${INT}))/smp_affinity
CPUID=$((CPUID*FACTOR))
if [ ${CPUID} -gt ${MAXCPUID} ]; then
CPUID=1
CPUID=$((CPUID*OFFSET))
fi
done
done
```

CPU Frequency (cpufreq.sh)

The following script sets the CPU frequency.

```
#Usage "./nameofscript.sh"
grep -E '^model name|^cpu MHz' /proc/cpuinfo
cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
for CPUFREQ in /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor; do [ -f
$CPUFREQ ] || continue; echo -n performance > $CPUFREQ; done
cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
```

To configure the network or memory settings:

```
sysctl -w net.ipv4.tcp_mem="16777216 16777216 16777216"
sysctl -w net.ipv4.tcp_wmem="4096 65536 16777216"
sysctl -w net.ipv4.tcp_rmem="4096 87380 16777216"
sysctl -w net.core.wmem_max=16777216
sysctl -w net.core.rmem_max=16777216
sysctl -w net.core.wmem_default=16777216
sysctl -w net.core.rmem_default=16777216
sysctl -w net.core.optmem_max=16777216
sysctl -w net.ipv4.tcp_low_latency=1
sysctl -w net.ipv4.tcp_timestamps=0
sysctl -w net.ipv4.tcp_sack=1
sysctl -w net.ipv4.tcp_window_scaling=0
sysctl -w net.ipv4.tcp_adv_win_scale=1
```

NOTE

The following commands apply only to the initiator server.

```
# echo 0 > /sys/block/nvme0n1/queue/add_random
# echo 2 > /sys/block/nvme0n1/queue/nomerges
```

13 Windows Server 2016

This chapter provides the following information for Windows Server 2016:

- Configuring RoCE Interfaces with Hyper-V
- “RoCE over Switch Embedded Teaming” on page 233
- “Configuring QoS for RoCE” on page 235
- “Configuring VMMQ” on page 243
- “Configuring VXLAN” on page 248
- “Configuring Storage Spaces Direct” on page 250
- “Deploying and Managing a Nano Server” on page 256

Configuring RoCE Interfaces with Hyper-V

In Windows Server 2016, Hyper-V with Network Direct Kernel Provider Interface (NDKPI) Mode-2, host virtual network adapters (host virtual NICs) support RDMA.

NOTE

DCBX is required for RoCE over Hyper-V. To configure DCBX, either:

- Configure through the HII (see “Preparing the Adapter” on page 65).
- Configure using QoS (see “Configuring QoS for RoCE” on page 235).

RoCE configuration procedures in this section include:

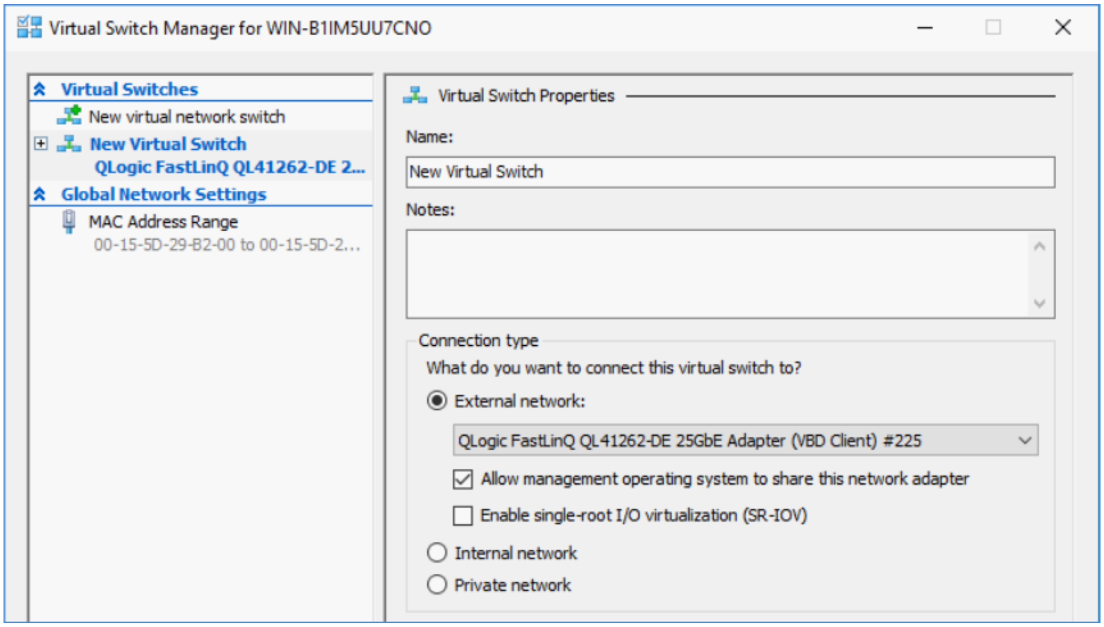
- Creating a Hyper-V Virtual Switch with an RDMA Virtual NIC
- Adding a VLAN ID to Host Virtual NIC
- Verifying If RoCE is Enabled
- Adding Host Virtual NICs (Virtual Ports)
- Mapping the SMB Drive and Running RoCE Traffic

Creating a Hyper-V Virtual Switch with an RDMA Virtual NIC

Follow the procedures in this section to create a Hyper-V virtual switch and then enable RDMA in the host VNIC.

To create a Hyper-V virtual switch with an RDMA virtual NIC:

1. Launch Hyper-V Manager.
2. Click **Virtual Switch Manager** (see Figure 13-1).



Virtual Switch Manager for WIN-B1IM5UU7CNO

Virtual Switches
- New virtual network switch
- New Virtual Switch
  QLogic FastLinQ QL41262-DE 2...

Global Network Settings
- MAC Address Range
  00-15-5D-29-B2-00 to 00-15-5D-2...

Virtual Switch Properties

Name: New Virtual Switch

Notes:

Connection type

What do you want to connect this virtual switch to?

● External network:

QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225

☑ Allow management operating system to share this network adapter

□ Enable single-root I/O virtualization (SR-IOV)

○ Internal network

○ Private network

Figure 13-1. Enabling RDMA in Host Virtual NIC

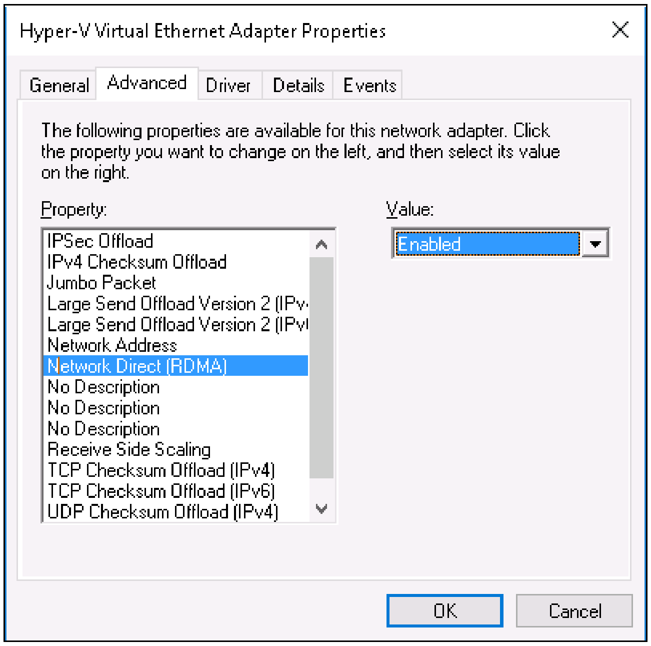
3. Create a virtual switch.
4. Select the **Allow management operating system to share this network adapter** check box.

In Windows Server 2016, a new parameter—Network Direct (RDMA)—is added in the Host virtual NIC.

To enable RDMA in a host virtual NIC:

1. Open the Hyper-V Virtual Ethernet Adapter Properties window.
2. Click the **Advanced** tab.
3. On the Advanced page (Figure 13-2):
   a. Under **Property**, select **Network Direct (RDMA)**.
   b. Under **Value**, select **Enabled**.

c. Click **OK**.


Hyper-V Virtual Ethernet Adapter Properties
General
Advanced
Driver
Details
Events
IPSec Offload
IPv4 Checksum Offload
Jumbo Packet
Large Send Offload Version 2 (IPv
Large Send Offload Version 2 (IPv
Network Address
Network Direct (RDMA)
No Description
No Description
No Description
Receive Side Scaling
TCP Checksum Offload (IPv4)
TCP Checksum Offload (IPv6)
UDP Checksum Offload (IPv4)
OK
Cancel

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property: Network Direct (RDMA)

Value: Enabled

Figure 13-2. Hyper-V Virtual Ethernet Adapter Properties

4. To enable RDMA, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Enable-NetAdapterRdma "vEthernet
(New Virtual Switch)"
PS C:\Users\Administrator>
```

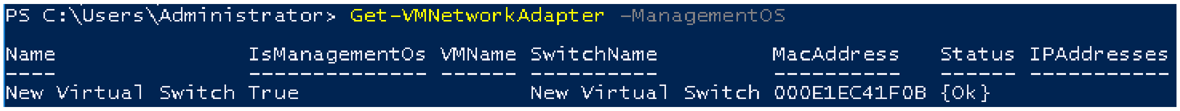
Adding a VLAN ID to Host Virtual NIC

To add a VLAN ID to a host virtual NIC:

1. To find the host virtual NIC name, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Get-VMNetworkAdapter -ManagementOS
```

Figure 13-3 shows the command output.



```
PS C:\Users\Administrator> Get-VMNetworkAdapter -ManagementOS

Name               IsManagementOs VMName SwitchName         MacAddress   Status IPAddresses
----               -------------- ------ ----------         ----------   ------ -----------
New Virtual Switch True                  New Virtual Switch 000E1EC41F0B {Ok}
```

Figure 13-3. Windows PowerShell Command: Get-VMNetworkAdapter

2. To set the VLAN ID to the host virtual NIC, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Set-VMNetworkAdaptervlan
-VMNetworkAdapterName "New Virtual Switch" -VlanId 5 -Access
-Management05
```

NOTE

Note the following about adding a VLAN ID to a host virtual NIC:

- A VLAN ID must be assigned to a host virtual NIC. The same VLAN ID must be assigned to all the interfaces, and on the switch.
- Make sure that the VLAN ID is not assigned to the physical interface when using a host virtual NIC for RoCE.
- If you are creating more than one host virtual NIC, you can assign a different VLAN to each host virtual NIC.

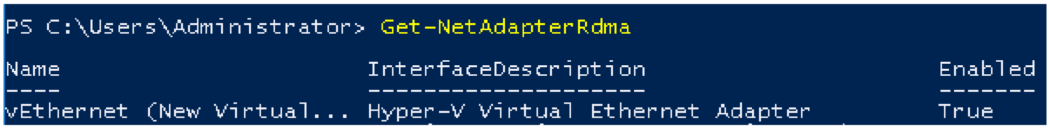
Verifying If RoCE is Enabled

To verify if the RoCE is enabled:

- Issue the following Windows PowerShell command:

  ```
  Get-NetAdapterRdma
  ```

  Command output lists the RDMA supported adapters as shown in Figure 13-4.



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

Name                      InterfaceDescription                     Enabled
----                      --------------------                     -------
vEthernet (New Virtual... Hyper-V Virtual Ethernet Adapter         True
```

Figure 13-4. Windows PowerShell Command: Get-NetAdapterRdma

Adding Host Virtual NICs (Virtual Ports)

To add host virtual NICs:

1. To add a host virtual NIC, issue the following command:

   ```
   Add-VMNetworkAdapter -SwitchName "New Virtual Switch" -Name
   SMB - ManagementOS
   ```

2. Enable RDMA on host virtual NICs as shown in “To enable RDMA in a host virtual NIC:” on page 229.

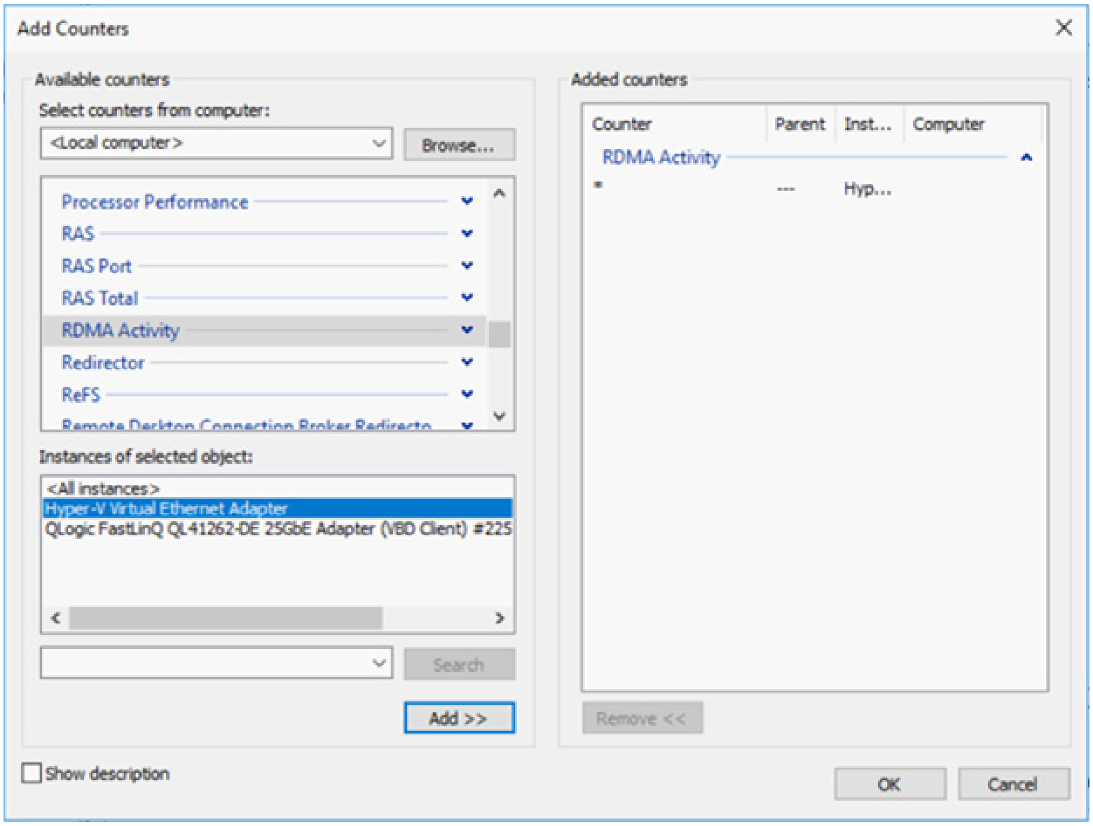
3. To assign a VLAN ID to the virtual port, issue the following command:

```
Set-VMNetworkAdapterVlan -VMNetworkAdapterName SMB -VlanId 5
-Access -ManagementOS
```

Mapping the SMB Drive and Running RoCE Traffic

To map the SMB drive and run the RoCE traffic:

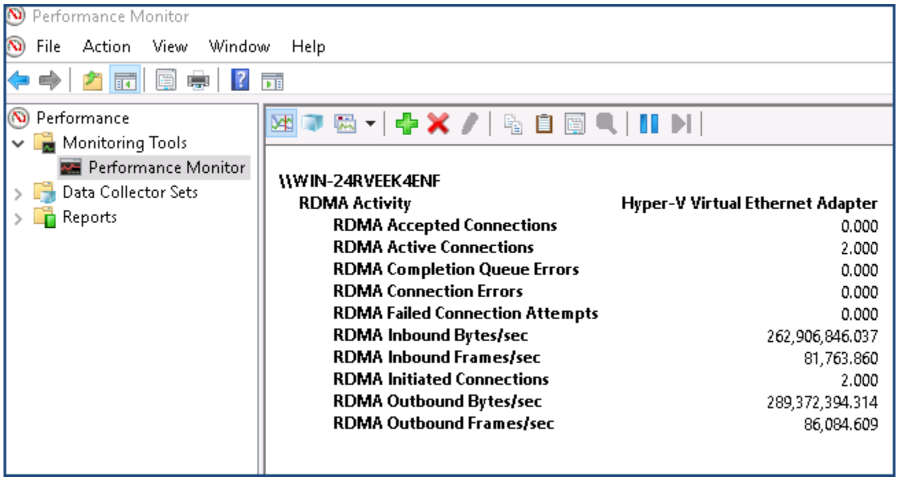
1. Launch the Performance Monitor (Perfmon).
2. Complete the Add Counters dialog box (Figure 13-5) as follows:
   a. Under **Available counters**, select **RDMA Activity.**
   b. Under **Instances of selected object**, select the adapter.
   c. Click **Add**.



Counter	Parent	Inst...	Computer
RDMA Activity			
*	---	Hyp...	

Figure 13-5. Add Counters Dialog Box

If the RoCE traffic is running, counters appear as shown in Figure 13-6.



```
\\WIN-24RVEEK4ENF
  RDMA Activity                            Hyper-V Virtual Ethernet Adapter
      RDMA Accepted Connections                                       0.000
      RDMA Active Connections                                         2.000
      RDMA Completion Queue Errors                                    0.000
      RDMA Connection Errors                                          0.000
      RDMA Failed Connection Attempts                                 0.000
      RDMA Inbound Bytes/sec                                262,906,846.037
      RDMA Inbound Frames/sec                                    81,763.860
      RDMA Initiated Connections                                      2.000
      RDMA Outbound Bytes/sec                               289,372,394.314
      RDMA Outbound Frames/sec                                   86,084.609
```

Figure 13-6. Performance Monitor Shows RoCE Traffic

RoCE over Switch Embedded Teaming

Switch Embedded Teaming (SET) is Microsoft's alternative NIC teaming solution available to use in environments that include Hyper-V and the Software Defined Networking (SDN) stack in Windows Server 2016 Technical Preview. SET integrates limited NIC Teaming functionality into the Hyper-V Virtual Switch.

Use SET to group between one and eight physical Ethernet network adapters into one or more software-based virtual network adapters. These adapters provide fast performance and fault tolerance if a network adapter failure occurs. To be placed on a team, SET member network adapters must all be installed in the same physical Hyper-V host.

RoCE over SET procedures included in this section:

- Creating a Hyper-V Virtual Switch with SET and RDMA Virtual NICs
- Enabling RDMA on SET
- Assigning a VLAN ID on SET
- Running RDMA Traffic on SET

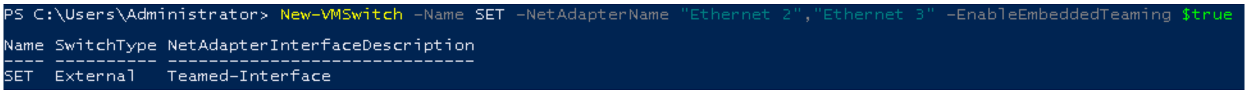
Creating a Hyper-V Virtual Switch with SET and RDMA Virtual NICs

To create a Hyper-V virtual switch with SET and RDMA virtual NICs:

- To create SET, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> New-VMSwitch -Name SET
-NetAdapterName "Ethernet 2","Ethernet 3"
-EnableEmbeddedTeaming $true
```

Figure 13-7 shows command output.



```
PS C:\Users\Administrator> New-VMSwitch -Name SET -NetAdapterName "Ethernet 2","Ethernet 3" -EnableEmbeddedTeaming $true

Name SwitchType NetAdapterInterfaceDescription
---- ---------- ------------------------------
SET  External   Teamed-Interface
```

Figure 13-7. Windows PowerShell Command: New-VMSwitch

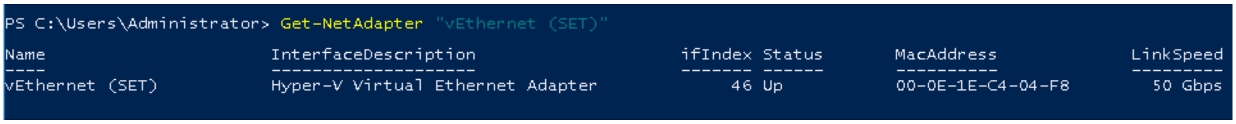
Enabling RDMA on SET

To enable RDMA on SET:

1. To view SET on the adapter, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Get-NetAdapter "vEthernet (SET)"
```

Figure 13-8 shows command output.



```
PS C:\Users\Administrator> Get-NetAdapter "vEthernet (SET)"

Name                       InterfaceDescription                    ifIndex Status       MacAddress             LinkSpeed
----                       --------------------                    ------- ------       ----------             ---------
vEthernet (SET)            Hyper-V Virtual Ethernet Adapter             46 Up           00-0E-1E-C4-04-F8        50 Gbps
```

Figure 13-8. Windows PowerShell Command: Get-NetAdapter

2. To enable RDMA on SET, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Enable-NetAdapterRdma "vEthernet
(SET)"
```

Assigning a VLAN ID on SET

To assign a VLAN ID on SET:

- To assign a VLAN ID on SET, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Set-VMNetworkAdapterVlan
-VMNetworkAdapterName "SET" -VlanId 5 -Access -ManagementOS
```

NOTE

Note the following when adding a VLAN ID to a host virtual NIC:

- Make sure that the VLAN ID is not assigned to the physical Interface when using a host virtual NIC for RoCE.
- If you are creating more than one host virtual NIC, a different VLAN can be assigned to each host virtual NIC.

Running RDMA Traffic on SET

For information about running RDMA traffic on SET, go to:

https://technet.microsoft.com/en-us/library/mt403349.aspx

Configuring QoS for RoCE

The two methods of configuring quality of service (QoS) include:

- Configuring QoS by Disabling DCBX on the Adapter
- Configuring QoS by Enabling DCBX on the Adapter

Configuring QoS by Disabling DCBX on the Adapter

All configuration must be completed on all of the systems in use before configuring QoS by disabling DCBX on the adapter. The priority-based flow control (PFC), enhanced transition services (ETS), and traffic classes configuration must be the same on the switch and server.

To configure QoS by disabling DCBX:

1. Disable DCBX on the adapter.
2. Using HII, set the **RoCE Priority** to `0`.
3. To install the DCB role in the host, issue the following Windows PowerShell command:

```
PS C:\Users\Administrators> Install-WindowsFeature
Data-Center-Bridging
```

4. To set the **DCBX Willing** mode to **False**, issue the following Windows PowerShell command:

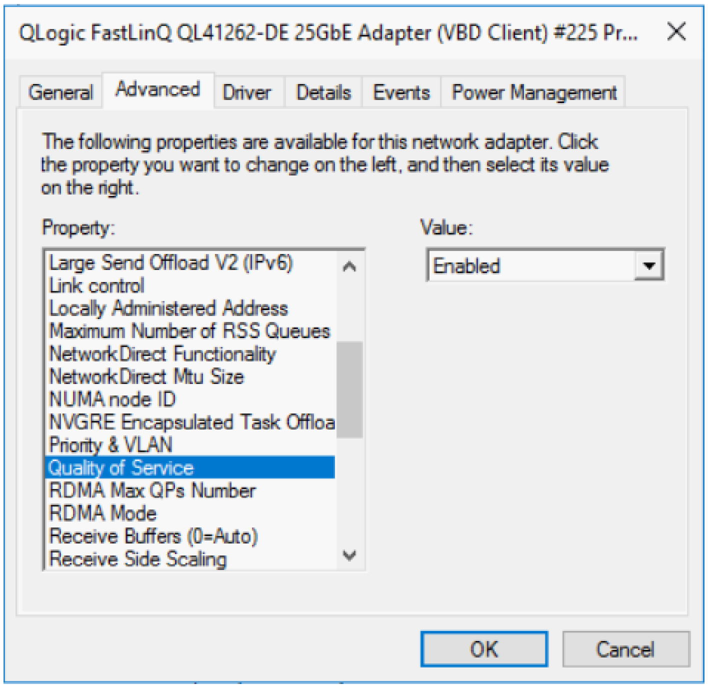
```
PS C:\Users\Administrators> set-NetQosDcbxSetting -Willing 0
```

5. Enable QoS in the miniport as follows:

   a. Open the miniport window, and then click the **Advanced** tab.

   b. On the adapter's Advanced Properties page (Figure 13-9) under **Property**, select **Quality of Service**, and then set the value to **Enabled**.

   c. Click **OK**.



QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...
General
Advanced
Driver
Details
Events
Power Management
OK
Cancel

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

Large Send Offload V2 (IPv6)
Link control
Locally Administered Address
Maximum Number of RSS Queues
NetworkDirect Functionality
NetworkDirect Mtu Size
NUMA node ID
NVGRE Encapsulated Task Offloa
Priority & VLAN
Quality of Service
RDMA Max QPs Number
RDMA Mode
Receive Buffers (0=Auto)
Receive Side Scaling

Value: Enabled

Figure 13-9. Advanced Properties: Enable QoS

6. Assign the VLAN ID to the interface as follows:

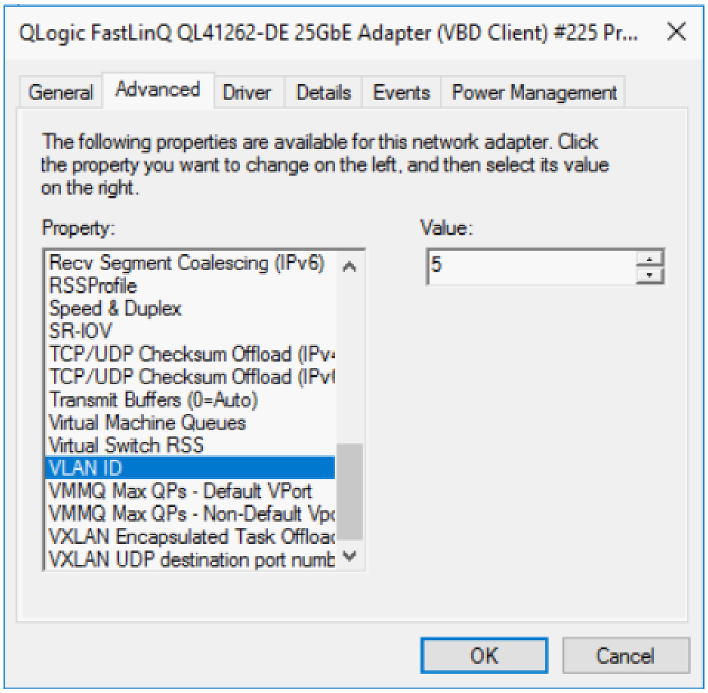
   a. Open the miniport window, and then click the **Advanced** tab.

   b. On the adapter's Advanced Properties page (Figure 13-10) under **Property**, select **VLAN ID**, and then set the value.

   c. Click **OK**.

NOTE

The preceding step is required for priority flow control (PFC).


QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...
General
Advanced
Driver
Details
Events
Power Management
OK
Cancel

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

Recv Segment Coalescing (IPv6)
RSSProfile
Speed & Duplex
SR-IOV
TCP/UDP Checksum Offload (IPv·
TCP/UDP Checksum Offload (IPv(
Transmit Buffers (0=Auto)
Virtual Machine Queues
Virtual Switch RSS
VLAN ID
VMMQ Max QPs - Default VPort
VMMQ Max QPs - Non-Default Vp(
VXLAN Encapsulated Task Offloa(
VXLAN UDP destination port numb

Value: 5

Figure 13-10. Advanced Properties: Setting VLAN ID

7. To enable PFC for RoCE on a specific priority, issue the following command:

   ```
   PS C:\Users\Administrators> Enable-NetQoSFlowControl
   -Priority 5
   ```

   NOTE

   If configuring RoCE over Hyper-V, do not assign a VLAN ID to the physical interface.

8. To disable priority flow control on any other priority, issue the following commands:

```
PS C:\Users\Administrator> Disable-NetQosFlowControl 0,1,2,3,4,6,7
PS C:\Users\Administrator> Get-NetQosFlowControl
Priority   Enabled    PolicySet        IfIndex IfAlias
--------   -------    ---------        ------- -------
0          False      Global
1          False      Global
2          False      Global
3          False      Global
4          False      Global
```

```
5          True       Global
6          False      Global
7          False      Global
```

9. To configure QoS and assign relevant priority to each type of traffic, issue the following commands (where Priority 5 is tagged for RoCE and Priority 0 is tagged for TCP):

```
PS C:\Users\Administrators> New-NetQosPolicy "SMB"
-NetDirectPortMatchCondition 445 -PriorityValue8021Action 5 -PolicyStore
ActiveStore

PS C:\Users\Administrators> New-NetQosPolicy "TCP" -IPProtocolMatchCondition
TCP -PriorityValue8021Action 0 -Policystore ActiveStore

PS C:\Users\Administrator> Get-NetQosPolicy -PolicyStore activestore

Name           : tcp
Owner          : PowerShell / WMI
NetworkProfile : All
Precedence     : 127
JobObject      :
IPProtocol     : TCP
PriorityValue  : 0

Name           : smb
Owner          : PowerShell / WMI
NetworkProfile : All
Precedence     : 127
JobObject      :
NetDirectPort  : 445
PriorityValue  : 5
```

10. To configure ETS for all traffic classes defined in the previous step, issue the following commands:

```
PS C:\Users\Administrators> New-NetQosTrafficClass -name "RDMA class"
-priority 5 -bandwidthPercentage 50 -Algorithm ETS

PS C:\Users\Administrators> New-NetQosTrafficClass -name "TCP class" -priority
0 -bandwidthPercentage 30 -Algorithm ETS

PS C:\Users\Administrator> Get-NetQosTrafficClass

Name        Algorithm Bandwidth(%) Priority   PolicySet       IfIndex IfAlias
----        --------- ------------ --------  ---------        ------- -------
```

```
[Default]   ETS       20           1-4,6-7     Global
RDMA class   ETS       50           5           Global
TCP class   ETS       30           0           Global
```

11. To see the network adapter QoS from the preceding configuration, issue the following Windows PowerShell command:

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

Name                       : SLOT 4 Port 1
Enabled                    : True
Capabilities               :                       Hardware     Current
                                                   --------     -------
                             MacSecBypass        : NotSupported NotSupported
                             DcbxSupport         : None         None
                             NumTCs(Max/ETS/PFC) : 4/4/4        4/4/4

OperationalTrafficClasses  : TC TSA    Bandwidth Priorities
                             -- ---    --------- ----------
                              0 ETS    20%       1-4,6-7
                              1 ETS    50%       5
                              2 ETS    30%       0

OperationalFlowControl     : Priority 5 Enabled
OperationalClassifications : Protocol  Port/Type Priority
                             --------  --------- --------
                             Default             0
                             NetDirect 445       5
```

12. Create a startup script to make the settings persistent across the system reboots.

13. Run RDMA traffic and verify as described in "RoCE Configuration" on page 63.

Configuring QoS by Enabling DCBX on the Adapter

All configuration must be completed on all of the systems in use. The PFC, ETS, and traffic classes configuration must be the same on the switch and server.

To configure QoS by enabling DCBX:

1. Enable DCBX (IEEE, CEE, or Dynamic).
2. Using HII, set the **RoCE Priority** to `0`.

NOTE

If the switch does not have a way of designating the RoCE traffic, you may need to set the **RoCE Priority** to the number used by the switch. Arista switches can do so, but some other switches cannot.

3. To install the DCB role in the host, issue the following Windows PowerShell command:

   ```
   PS C:\Users\Administrators> Install-WindowsFeature
   Data-Center-Bridging
   ```

NOTE

For this configuration, set the **DCBX Protocol** to **CEE**.

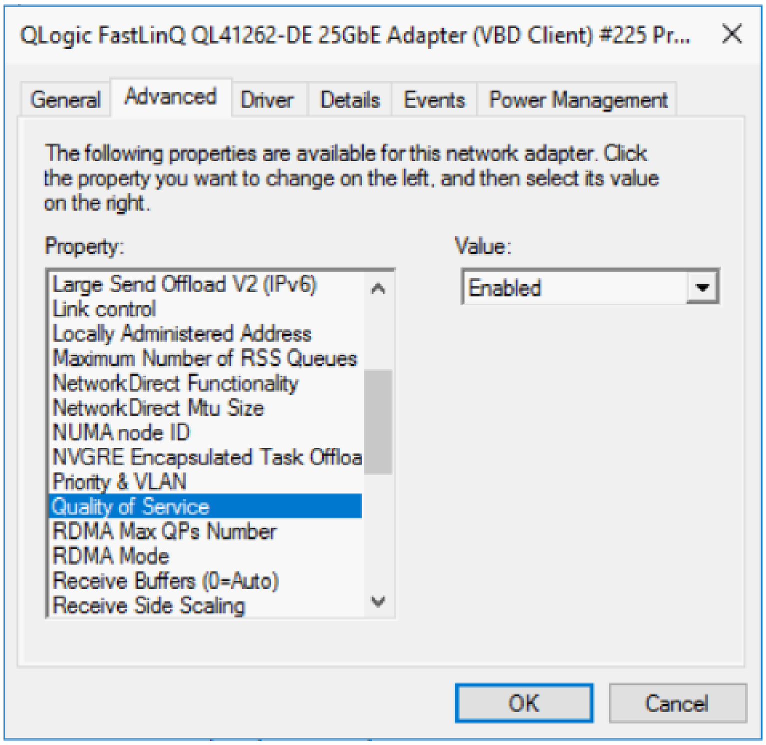
4. To set the **DCBX Willing** mode to **True**, issue the following command:

   ```
   PS C:\Users\Administrators> set-NetQosDcbxSetting -Willing 1
   ```

5. Enable QoS in the miniport as follows:

   a. On the adapter's Advanced Properties page (Figure 13-11) under **Property**, select **Quality of Service**, and then set the value to **Enabled**.

   b. Click **OK**.



QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

- Large Send Offload V2 (IPv6)
- Link control
- Locally Administered Address
- Maximum Number of RSS Queues
- NetworkDirect Functionality
- NetworkDirect Mtu Size
- NUMA node ID
- NVGRE Encapsulated Task Offloa
- Priority & VLAN
- Quality of Service
- RDMA Max QPs Number
- RDMA Mode
- Receive Buffers (0=Auto)
- Receive Side Scaling

Value: Enabled

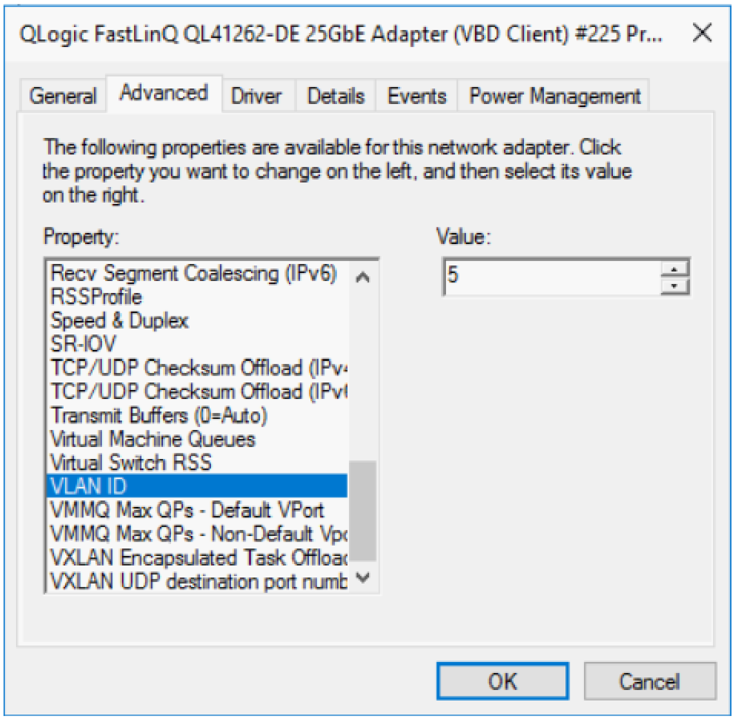
Figure 13-11. Advanced Properties: Enabling QoS

6. Assign the VLAN ID to the interface (required for PFC) as follows:

   a. Open the miniport window, and then click the **Advanced** tab.

   b. On the adapter's Advanced Properties page (Figure 13-12) under **Property**, select **VLAN ID**, and then set the value.

   c. Click **OK**.


QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...
General
Advanced
Driver
Details
Events
Power Management
Recv Segment Coalescing (IPv6)
RSSProfile
Speed & Duplex
SR-IOV
TCP/UDP Checksum Offload (IPv·
TCP/UDP Checksum Offload (IPv(
Transmit Buffers (0=Auto)
Virtual Machine Queues
Virtual Switch RSS
VLAN ID
VMMQ Max QPs - Default VPort
VMMQ Max QPs - Non-Default Vp(
VXLAN Encapsulated Task Offloa(
VXLAN UDP destination port numb
OK
Cancel

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property: VLAN ID

Value: 5

Figure 13-12. Advanced Properties: Setting VLAN ID

7. To configure the switch, issue the following Windows PowerShell command:

```
PS C:\Users\Administrators> Get-NetAdapterQoS

Name                           : Ethernet 5
Enabled                        : True
Capabilities                   :                       Hardware     Current
                                                       --------     -------
                                 MacSecBypass        : NotSupported NotSupported
                                 DcbxSupport         : CEE          CEE
                                 NumTCs(Max/ETS/PFC) : 4/4/4        4/4/4

OperationalTrafficClasses      : TC TSA    Bandwidth Priorities
                                 -- ---    --------- ----------
                                  0 ETS    5%        0-4,6-7
                                  1 ETS    95%       5

OperationalFlowControl         : Priority 5 Enabled
OperationalClassifications     : Protocol  Port/Type Priority
                                 --------  --------- --------
```

```
                             NetDirect 445       5

RemoteTrafficClasses       : TC TSA    Bandwidth Priorities
                             -- ---    --------- ----------
                              0 ETS    5%        0-4,6-7
                              1 ETS    95%       5

RemoteFlowControl          : Priority 5 Enabled
RemoteClassifications      : Protocol  Port/Type Priority
                             --------  --------- --------
                             NetDirect 445       5
```

NOTE

The preceding example is taken when the adapter port is connected to an Arista® 7060X switch. In this example, the switch PFC is enabled on Priority 5. RoCE App TLVs are defined. The two traffic classes are defined as TC0 and TC1, where TC1 is defined for RoCE. **DCBX Protocol** mode is set to **CEE**. For Arista switch configuration, refer to “Preparing the Ethernet Switch” on page 66“. When the adapter is in **Willing** mode, it accepts Remote Configuration and shows it as **Operational Parameters**.

Configuring VMMQ

Virtual machine multiqueue (VMMQ) configuration information includes:

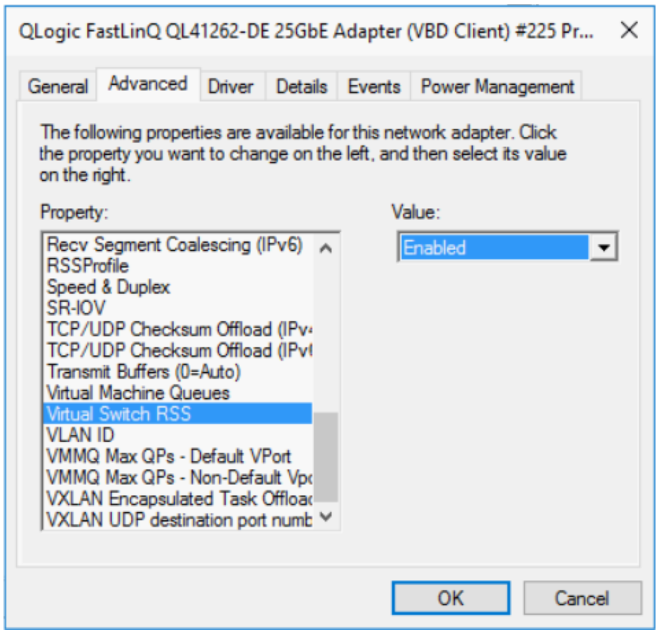
- Enabling VMMQ on the Adapter
- Creating a Virtual Machine Switch with or Without SR-IOV
- Enabling VMMQ on the Virtual Machine Switch
- Getting the Virtual Machine Switch Capability
- Creating a VM and Enabling VMMQ on VMNetworkadapters in the VM
- Default and Maximum VMMQ Virtual NIC
- Enabling and Disabling VMMQ on a Management NIC
- Monitoring Traffic Statistics

Enabling VMMQ on the Adapter

To enable VMMQ on the adapter:

1. Open the miniport window, and then click the **Advanced** tab.
2. On the Advanced Properties page (Figure 13-13) under **Property**, select **Virtual Switch RSS**, and then set the value to **Enabled**.

3. Click **OK**.


QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225 Pr...

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

- Recv Segment Coalescing (IPv6)
- RSSProfile
- Speed & Duplex
- SR-IOV
- TCP/UDP Checksum Offload (IPv4
- TCP/UDP Checksum Offload (IPv6
- Transmit Buffers (0=Auto)
- Virtual Machine Queues
- Virtual Switch RSS
- VLAN ID
- VMMQ Max QPs - Default VPort
- VMMQ Max QPs - Non-Default Vpo
- VXLAN Encapsulated Task Offload
- VXLAN UDP destination port numb

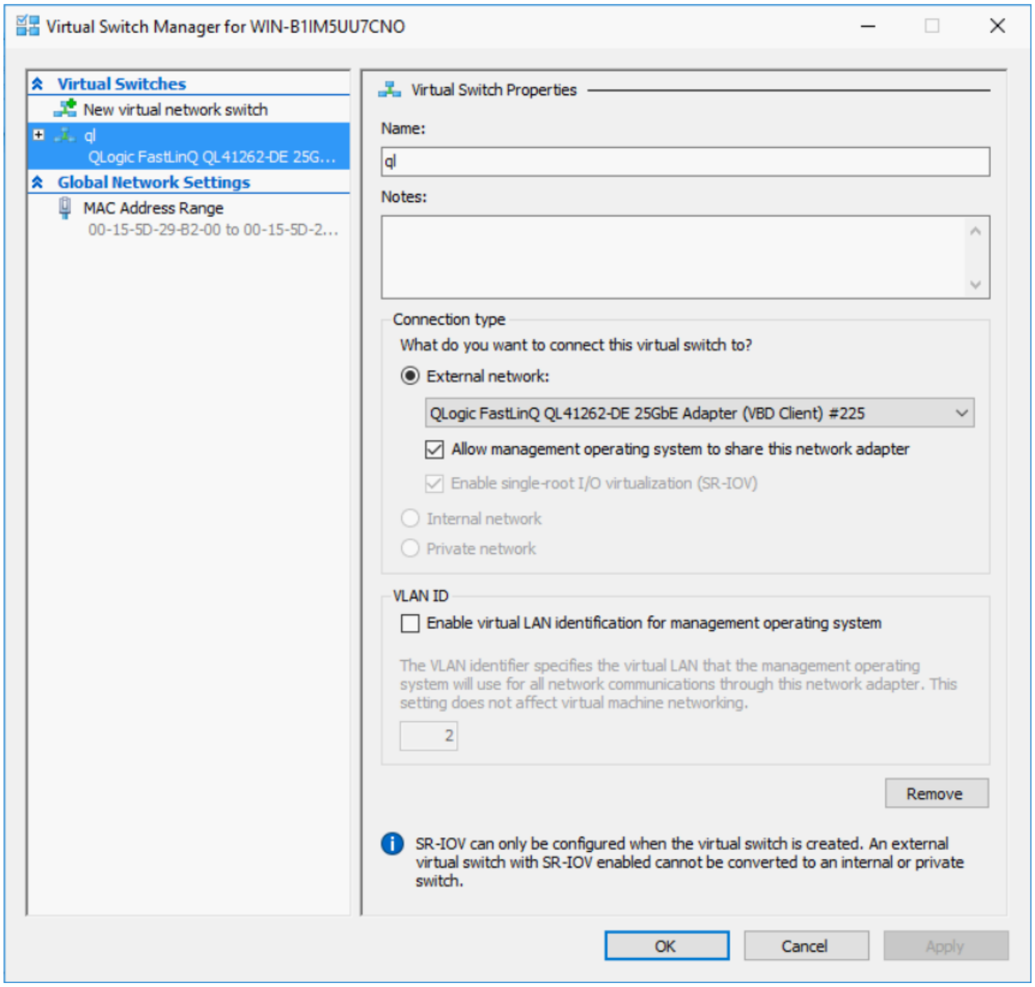
Value: Enabled

Figure 13-13. Advanced Properties: Enabling Virtual Switch RSS

Creating a Virtual Machine Switch with or Without SR-IOV

To create a virtual machine switch with or without SR-IOV:

1. Launch the Hyper-V Manager.
2. Select **Virtual Switch Manager** (see Figure 13-14).
3. In the **Name** box, type a name for the virtual switch.
4. Under **Connection type**:
   a. Click **External network**.
   b. Select the **Allow management operating system to share this network adapter** check box.



Virtual Switch Manager for WIN-B1IM5UU7CNO

Virtual Switches

New virtual network switch

ql
QLogic FastLinQ QL41262-DE 25G...

Global Network Settings

MAC Address Range
00-15-5D-29-B2-00 to 00-15-5D-2...

Virtual Switch Properties

Name: ql

Notes:

Connection type

What do you want to connect this virtual switch to?

● External network:

QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225

☑ Allow management operating system to share this network adapter

☑ Enable single-root I/O virtualization (SR-IOV)

○ Internal network

○ Private network

VLAN ID

□ Enable virtual LAN identification for management operating system

The VLAN identifier specifies the virtual LAN that the management operating system will use for all network communications through this network adapter. This setting does not affect virtual machine networking.

2

Remove

SR-IOV can only be configured when the virtual switch is created. An external virtual switch with SR-IOV enabled cannot be converted to an internal or private switch.

OK Cancel Apply

Figure 13-14. Virtual Switch Manager

5. Click **OK**.

Enabling VMMQ on the Virtual Machine Switch

To enable VMMQ on the virtual machine switch:

- Issue the following Windows PowerShell command:

```
PS C:\Users\Administrators> Set-VMSwitch -name q1
-defaultqueuevmmqenabled $true -defaultqueuevmmqqueuepairs 4
```

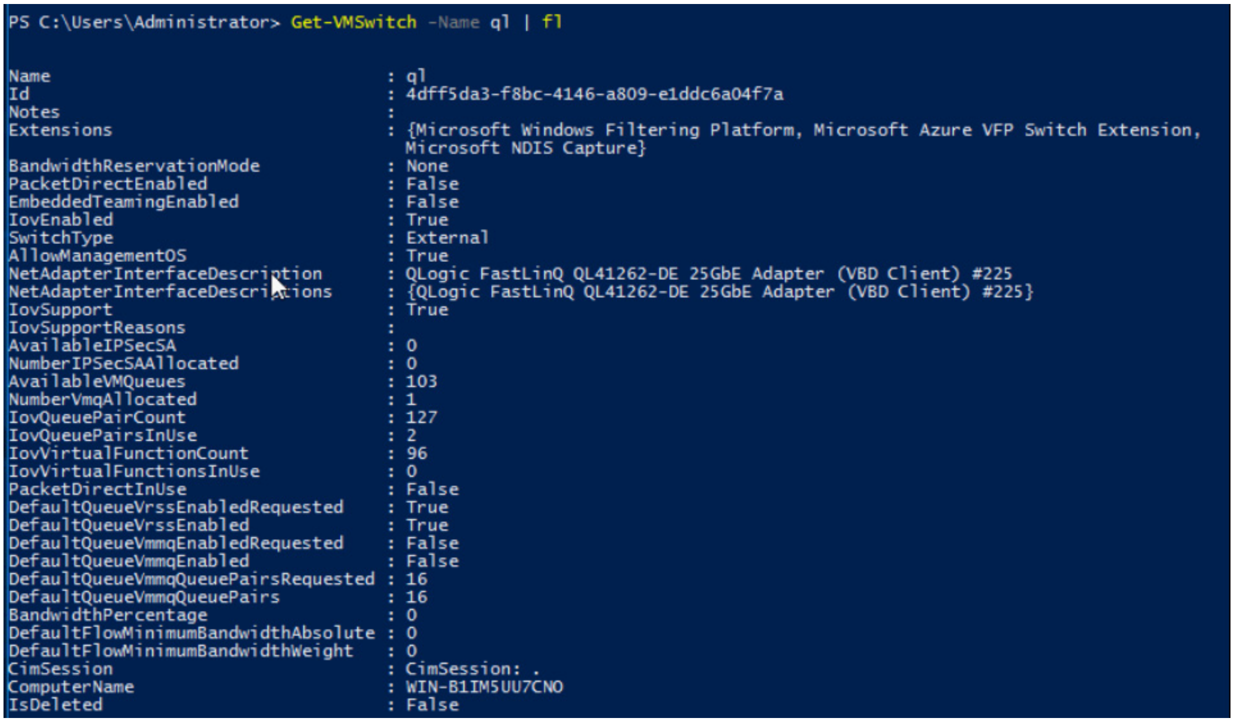
Getting the Virtual Machine Switch Capability

To get the virtual machine switch capability:

- Issue the following Windows PowerShell command:

  ```
  PS C:\Users\Administrator> Get-VMSwitch -Name ql | fl
  ```

  Figure 13-15 shows example output.



```
PS C:\Users\Administrator> Get-VMSwitch -Name ql | fl


Name                                : ql
Id                                  : 4dff5da3-f8bc-4146-a809-e1ddc6a04f7a
Notes                               :
Extensions                          : {Microsoft Windows Filtering Platform, Microsoft Azure VFP Switch Extension,
                                      Microsoft NDIS Capture}
BandwidthReservationMode            : None
PacketDirectEnabled                 : False
EmbeddedTeamingEnabled              : False
IovEnabled                          : True
SwitchType                          : External
AllowManagementOS                   : True
NetAdapterInterfaceDescription      : QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225
NetAdapterInterfaceDescriptions     : {QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225}
IovSupport                          : True
IovSupportReasons                   :
AvailableIPSecSA                    : 0
NumberIPSecSAAllocated              : 0
AvailableVMQueues                   : 103
NumberVmqAllocated                  : 1
IovQueuePairCount                   : 127
IovQueuePairsInUse                  : 2
IovVirtualFunctionCount             : 96
IovVirtualFunctionsInUse            : 0
PacketDirectInUse                   : False
DefaultQueueVrssEnabledRequested    : True
DefaultQueueVrssEnabled             : True
DefaultQueueVmmqEnabledRequested    : False
DefaultQueueVmmqEnabled             : False
DefaultQueueVmmqQueuePairsRequested : 16
DefaultQueueVmmqQueuePairs          : 16
BandwidthPercentage                 : 0
DefaultFlowMinimumBandwidthAbsolute : 0
DefaultFlowMinimumBandwidthWeight   : 0
CimSession                          : CimSession: .
ComputerName                        : WIN-B1IM5UU7CNO
IsDeleted                           : False
```

Figure 13-15. Windows PowerShell Command: Get-VMSwitch

Creating a VM and Enabling VMMQ on VMNetworkadapters in the VM

To create a virtual machine (VM) and enable VMMQ on VMNetworksadapters in the VM:

1. Create a VM.
2. Add the VMNetworkadapter to the VM.
3. Assign a virtual switch to the VMNetworkadapter.

4. To enable VMMQ on the VM, issue the following Windows PowerShell command:

```
PS C:\Users\Administrators> set-vmnetworkadapter -vmname vm1
-VMNetworkAdapterName "network adapter" -vmmqenabled $true
-vmmqqueuepairs 4
```

NOTE

For an SR-IOV capable virtual switch: If the VM switch and hardware acceleration is SR-IOV-enabled, you must create 10 VMs with 8 virtual NICs each to utilize VMMQ. This requirement is because the SR-IOV has precedence over VMMQ.

Example output of 64 virtual functions and 16 VMMQs is shown here:

```
PS C:\Users\Administrator> get-netadaptervport

Name         ID     MacAddress                VID   ProcMask   FID       State      ITR        QPairs
----         --     ----------                ---   --------   ---       -----      ---        ------
Ethernet 3   0      00-15-5D-36-0A-FB               0:0        PF        Activated  Unknown    4
Ethernet 3   1      00-0E-1E-C4-C0-A4               0:8        PF        Activated  Adaptive   4
Ethernet 3   2                                      0:0        0         Activated  Unknown    1
Ethernet 3   3                                      0:0        1         Activated  Unknown    1
Ethernet 3   4                                      0:0        2         Activated  Unknown    1
Ethernet 3   5                                      0:0        3         Activated  Unknown    1
Ethernet 3   6                                      0:0        4         Activated  Unknown    1
Ethernet 3   7                                      0:0        5         Activated  Unknown    1
Ethernet 3   8                                      0:0        6         Activated  Unknown    1
Ethernet 3   9                                      0:0        7         Activated  Unknown    1
Ethernet 3   10                                     0:0        8         Activated  Unknown    1
Ethernet 3   11                                     0:0        9         Activated  Unknown    1
.
.
.
Ethernet 3   64                                     0:0        62        Activated  Unknown    1
Ethernet 3   65                                     0:0        63        Activated  Unknown    1
Ethernet 3   66     00-15-5D-36-0A-04               0:16       PF        Activated  Adaptive   4
Ethernet 3   67     00-15-5D-36-0A-05               1:0        PF        Activated  Adaptive   4
Ethernet 3   68     00-15-5D-36-0A-06               0:0        PF        Activated  Adaptive   4
Name         ID     MacAddress                VID   ProcMask   FID       State      ITR        QPairs
----         --     ----------                ---   --------   ---       -----      ---        ------
Ethernet 3   69     00-15-5D-36-0A-07               0:8        PF        Activated  Adaptive   4
Ethernet 3   70     00-15-5D-36-0A-08               0:16       PF        Activated  Adaptive   4
Ethernet 3   71     00-15-5D-36-0A-09               1:0        PF        Activated  Adaptive   4
Ethernet 3   72     00-15-5D-36-0A-0A               0:0        PF        Activated  Adaptive   4
Ethernet 3   73     00-15-5D-36-0A-0B               0:8        PF        Activated  Adaptive   4
Ethernet 3   74     00-15-5D-36-0A-F4               0:16       PF        Activated  Adaptive   4
Ethernet 3   75     00-15-5D-36-0A-F5               1:0        PF        Activated  Adaptive   4
Ethernet 3   76     00-15-5D-36-0A-F6               0:0        PF        Activated  Adaptive   4
Ethernet 3   77     00-15-5D-36-0A-F7               0:8        PF        Activated  Adaptive   4
Ethernet 3   78     00-15-5D-36-0A-F8               0:16       PF        Activated  Adaptive   4
Ethernet 3   79     00-15-5D-36-0A-F9               1:0        PF       Activated    Adaptive   4
Ethernet 3   80     00-15-5D-36-0A-FA               0:0        PF       Activated    Adaptive   4

PS C:\Users\Administrator> get-netadaptervmq

Name         InterfaceDescription              Enabled  BaseVmqProcessor  MaxProcessors NumberOfReceive
                                                                                        Queues
----         --------------------              -------  ----------------  ------------- ---------------
Ethernet 4   QLogic FastLinQ 41xxx             False    0:0               16            1
```

Default and Maximum VMMQ Virtual NIC

According to the current implementation, a maximum quantity of 4 VMMQs is available per virtual NIC; that is, up to 16 virtual NICs.

Four default queues are available as previously set using Windows PowerShell commands. The maximum default queue can currently be set to 8. To verify the maximum default queue, use the VMswitch capability.

Enabling and Disabling VMMQ on a Management NIC

To enable or disable VMMQ on a management NIC:

- To enable VMMQ on a management NIC, issue the following command:

  ```
  PS C:\Users\Administrator> Set-VMNetworkAdapter –ManagementOS
  –vmmqEnabled $true
  ```

  The Management OS VNIC has four VMMQs.

- To disable VMMQ on a management NIC, issue the following command:

  ```
  PS C:\Users\Administrator> Set-VMNetworkAdapter –ManagementOS
  –vmmqEnabled $false
  ```

  A VMMQ will also be available for the multicast open shortest path first (MOSPF).

Monitoring Traffic Statistics

To monitor virtual function traffic in a virtual machine, issue the following Windows PowerShell command:

```
PS C:\Users\Administrator> Use get-netadapterstatistics | fl
```

Configuring VXLAN

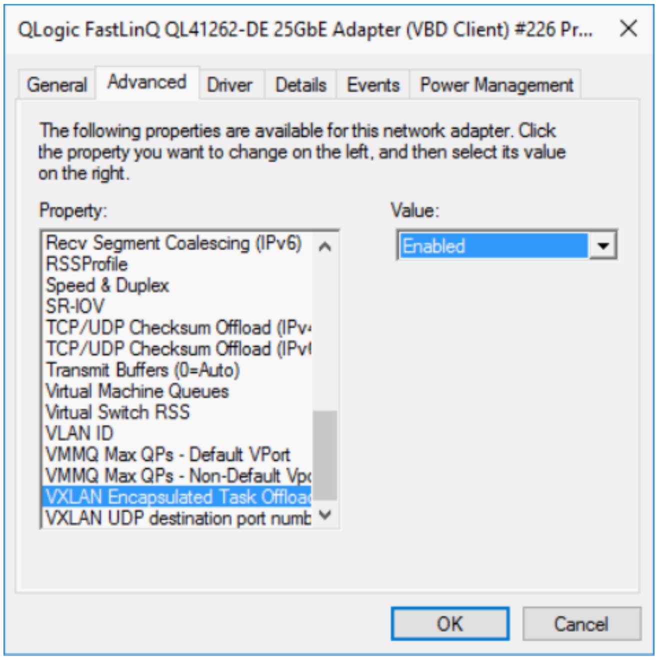
VXLAN configuration information includes:

- Enabling VXLAN Offload on the Adapter
- Deploying a Software Defined Network

Enabling VXLAN Offload on the Adapter

To enable VXLAN offload on the adapter:

1. Open the miniport window, and then click the **Advanced** tab.
2. On the Advanced Properties page (Figure 13-16) under **Property**, select **VXLAN Encapsulated Task Offload**.



QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #226 Pr...

The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

Recv Segment Coalescing (IPv6)
RSSProfile
Speed & Duplex
SR-IOV
TCP/UDP Checksum Offload (IPv4
TCP/UDP Checksum Offload (IPv6
Transmit Buffers (0=Auto)
Virtual Machine Queues
Virtual Switch RSS
VLAN ID
VMMQ Max QPs - Default VPort
VMMQ Max QPs - Non-Default Vpo
VXLAN Encapsulated Task Offload
VXLAN UDP destination port numb

Value: Enabled

Figure 13-16. Advanced Properties: Enabling VXLAN

3. Set the **Value** to **Enabled**.
4. Click **OK**.

Deploying a Software Defined Network

To take advantage of VXLAN encapsulation task offload on virtual machines, you must deploy a Software Defined Networking (SDN) stack that utilizes a Microsoft Network Controller.

For more details, refer to the following Microsoft TechNet link on Software Defined Networking:

https://technet.microsoft.com/en-us/windows-server-docs/networking/sdn/software-defined-networking--sdn-

Configuring Storage Spaces Direct

Windows Server 2016 introduces Storage Spaces Direct, which allows you to build highly available and scalable storage systems with local storage.

For more information, refer to the following Microsoft TechNet link:

https://technet.microsoft.com/en-us/windows-server-docs/storage/storage-spaces/storage-spaces-direct-windows-server-2016

Configuring the Hardware

Figure 13-17 shows an example of hardware configuration.

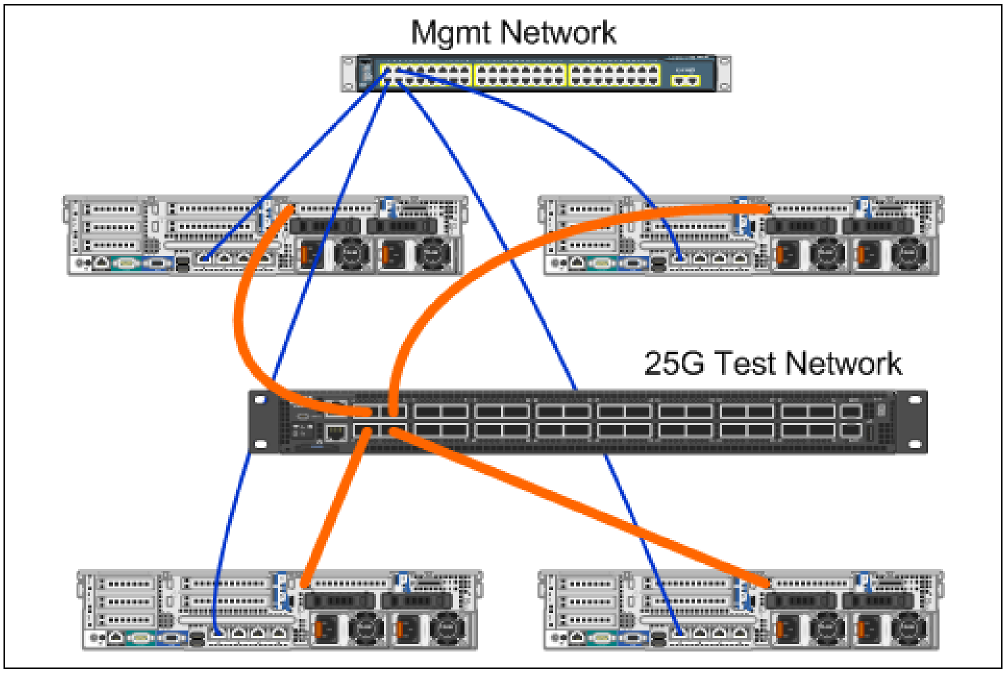


Figure 13-17. Example Hardware Configuration

NOTE

The disks used in this example are 4 × 400G NVMe™, and 12 × 200G SSD disks.

Deploying a Hyper-Converged System

This section includes instructions to install and configure the components of a Hyper-Converged system using the Windows Server 2016. The act of deploying a Hyper-Converged system can be divided into the following three high-level phases:

- Deploying the Operating System
- Configuring the Network
- Configuring Storage Spaces Direct

Deploying the Operating System

To deploy the operating systems:

1. Install the operating system.
2. Install the Windows Server roles (Hyper-V).
3. Install the following features:
   - ❑ Failover
   - ❑ Cluster
   - ❑ Data center bridging (DCB)
4. Connect the nodes to a domain and add domain accounts.

Configuring the Network

To deploy Storage Spaces Direct, the Hyper-V switch must be deployed with RDMA-enabled host virtual NICs.

NOTE

The following procedure assumes that there are four RDMA NIC ports.

To configure the network on each server:

1. Configure the physical network switch as follows:

   a. Connect all adapter NICs to the switch port.

   NOTE

   If your test adapter has more than one NIC port, you must connect both ports to the same switch.

   b. Enable the switch port and make sure that the switch port supports switch-independent teaming mode, and is also part of multiple VLAN networks.

Example Dell switch configuration:

```
no ip address
mtu 9416
portmode hybrid
switchport
dcb-map roce_S2D
protocol lldp
dcbx version cee
no shutdown
```

2. Enable **Network Quality of Service**.

NOTE

Network Quality of Service is used to ensure that the Software Defined Storage system has enough bandwidth to communicate between the nodes to ensure resiliency and performance. To configure QoS on the adapter, see "Configuring QoS for RoCE" on page 235.

3. Create a Hyper-V virtual switch with Switch Embedded Teaming (SET) and RDMA virtual NIC as follows:

   a. To identify the network adapters, issue the following command:

   ```
   Get-NetAdapter | FT
   Name,InterfaceDescription,Status,LinkSpeed
   ```

   b. To create the virtual switch connected to all of the physical network adapters, and then enable SET, issue the following command:

   ```
   New-VMSwitch -Name SETswitch -NetAdapterName
   "<port1>","<port2>","<port3>","<port4>" –
   EnableEmbeddedTeaming $true
   ```

   c. To add host virtual NICs to the virtual switch, issue the following commands:

   ```
   Add-VMNetworkAdapter –SwitchName SETswitch –Name SMB_1 –
   managementOS
   Add-VMNetworkAdapter –SwitchName SETswitch –Name SMB_2 –
   managementOS
   ```

   NOTE

   The preceding commands configure the virtual NIC from the virtual switch that you just configured for the management operating system to use.

d. To configure the host virtual NIC to use a VLAN, issue the following commands:

```
Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SMB_1"
-VlanId 5 -Access -ManagementOS
Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SMB_2"
-VlanId 5 -Access -ManagementOS
```

NOTE

These commands can be on the same or different VLANs.

e. To verify that the VLAN ID is set, issue the following command:

```
Get-VMNetworkAdapterVlan -ManagementOS
```

f. To disable and enable each host virtual NIC adapter so that the VLAN is active, issue the following commands:

```
Disable-NetAdapter "vEthernet (SMB_1)"
Enable-NetAdapter "vEthernet (SMB_1)"
Disable-NetAdapter "vEthernet (SMB_2)"
Enable-NetAdapter "vEthernet (SMB_2)"
```

g. To enable RDMA on the host virtual NIC adapters, issue the following command:

```
Enable-NetAdapterRdma "SMB1","SMB2"
```

h. To verify RDMA capabilities, issue the following command:

```
Get-SmbClientNetworkInterface | where RdmaCapable -EQ
$true
```

Configuring Storage Spaces Direct

Configuring Storage Spaces Direct in Windows Server 2016 includes the following steps:

- Step 1. Running a Cluster Validation Tool
- Step 2. Creating a Cluster
- Step 3. Configuring a Cluster Witness
- Step 4. Cleaning Disks Used for Storage Spaces Direct
- Step 5. Enabling Storage Spaces Direct
- Step 6. Creating Virtual Disks
- Step 7. Creating or Deploying Virtual Machines

Step 1. Running a Cluster Validation Tool

Run the cluster validation tool to make sure server nodes are configured correctly to create a cluster using Storage Spaces Direct.

To validate a set of servers for use as a Storage Spaces Direct cluster, issue the following Windows PowerShell command:

```
Test-Cluster -Node <MachineName1, MachineName2, MachineName3,
MachineName4> -Include "Storage Spaces Direct", Inventory,
Network, "System Configuration"
```

Step 2. Creating a Cluster

Create a cluster with the four nodes (which was validated for cluster creation) in Step 1. Running a Cluster Validation Tool.

To create a cluster, issue the following Windows PowerShell command:

```
New-Cluster -Name <ClusterName> -Node <MachineName1, MachineName2,
MachineName3, MachineName4> -NoStorage
```

The `-NoStorage` parameter is required. If it is not included, the disks are automatically added to the cluster, and you must remove them before enabling Storage Spaces Direct. Otherwise, they will not be included in the Storage Spaces Direct storage pool.

Step 3. Configuring a Cluster Witness

You should configure a witness for the cluster, so that this four-node system can withstand two nodes failing or being offline. With these systems, you can configure file share witness or cloud witness.

For more information, go to:

https://blogs.msdn.microsoft.com/clustering/2014/03/31/configuring-a-file-share-witness-on-a-scale-out-file-server/

Step 4. Cleaning Disks Used for Storage Spaces Direct

The disks intended to be used for Storage Spaces Direct must be empty and without partitions or other data. If a disk has partitions or other data, it will not be included in the Storage Spaces Direct system.

The following Windows PowerShell command can be placed in a Windows PowerShell script (`.PS1`) file and executed from the management system in an open Windows PowerShell (or Windows PowerShell ISE) console with Administrator privileges.

NOTE

Running this script helps identify the disks on each node that can be used for Storage Spaces Direct. It also removes all data and partitions from those disks.

```
icm (Get-Cluster -Name HCNanoUSClu3 | Get-ClusterNode) {
Update-StorageProviderCache

Get-StoragePool |? IsPrimordial -eq $false | Set-StoragePool
-IsReadOnly:$false -ErrorAction SilentlyContinue

Get-StoragePool |? IsPrimordial -eq $false | Get-VirtualDisk |
Remove-VirtualDisk -Confirm:$false -ErrorAction SilentlyContinue

Get-StoragePool |? IsPrimordial -eq $false | Remove-StoragePool
-Confirm:$false -ErrorAction SilentlyContinue

Get-PhysicalDisk | Reset-PhysicalDisk -ErrorAction
SilentlyContinue

Get-Disk |? Number -ne $null |? IsBoot -ne $true |? IsSystem -ne
$true |? PartitionStyle -ne RAW |% {
$_ | Set-Disk -isoffline:$false
$_ | Set-Disk -isreadonly:$false
$_ | Clear-Disk -RemoveData -RemoveOEM -Confirm:$false
$_ | Set-Disk -isreadonly:$true
$_ | Set-Disk -isoffline:$true
}
Get-Disk |? Number -ne $null |? IsBoot -ne $true |? IsSystem -ne
$true |? PartitionStyle -eq RAW | Group -NoElement -Property
FriendlyName

} | Sort -Property PsComputerName,Count
```

Step 5. Enabling Storage Spaces Direct

After creating the cluster, issue the `Enable-ClusterStorageSpacesDirect` Windows PowerShell cmdlet. The cmdlet places the storage system into the Storage Spaces Direct mode and automatically does the following:

- Creates a single, large pool that has a name such as *S2D on Cluster1*.
- Configures Storage Spaces Direct cache. If there is more than one media type available for Storage Spaces Direct use, it configures the most efficient type as cache devices (in most cases, read and write).
- Creates two tiers—**Capacity** and **Performance**—as default tiers. The cmdlet analyzes the devices and configures each tier with the mix of device types and resiliency.

Step 6. Creating Virtual Disks

If the Storage Spaces Direct was enabled, it creates a single pool using all of the disks. It also names the pool (for example *S2D on Cluster1*), with the name of the cluster that is specified in the name.

The following Windows PowerShell command creates a virtual disk with both mirror and parity resiliency on the storage pool:

```
New-Volume -StoragePoolFriendlyName "S2D*" -FriendlyName
<VirtualDiskName> -FileSystem CSVFS_ReFS -StorageTierfriendlyNames
Capacity,Performance -StorageTierSizes <Size of capacity tier in
size units, example: 800GB>, <Size of Performance tier in size
units, example: 80GB> -CimSession <ClusterName>
```

Step 7. Creating or Deploying Virtual Machines

You can provision the virtual machines onto the nodes of the hyper-converged S2D cluster. Store the virtual machine's files on the system's CSV namespace (for example, `c:\ClusterStorage\Volume1`), similar to clustered virtual machines on failover clusters.

Deploying and Managing a Nano Server

Windows Server 2016 offers Nano Server as a new installation option. Nano Server is a remotely administered server operating system optimized for private clouds and data centers. It is similar to Windows Server in Server Core mode, but is significantly smaller, has no local logon capability, and supports only 64-bit applications, tools, and agents. The Nano Server takes less disk space, sets up faster, and requires fewer updates and restarts than Windows Server. When it does restart, it restarts much faster.

Roles and Features

Table 13-1 shows the roles and features that are available in this release of Nano Server, along with the Windows PowerShell options that will install the packages for them. Some packages are installed directly with their own Windows PowerShell options (such as `-Compute`). Others are installed as extensions to the `-Packages` option, which you can combine in a comma-separated list.

Table 13-1. Roles and Features of Nano Server

Role or Feature	Options
Hyper-V role	`-Compute`
Failover Clustering	`-Clustering`
Hyper-V guest drivers for hosting the Nano Server as a virtual machine	`-GuestDrivers`

Table 13-1. Roles and Features of Nano Server (Continued)

Role or Feature	Options
Basic drivers for a variety of network adapters and storage controllers. This is the same set of drivers included in a Server Core installation of Windows Server 2016 Technical Preview.	`-OEMDrivers`
File Server role and other storage components	`-Storage`
Windows Defender Antimalware, including a default signature file	`-Defender`
Reverse forwarders for application compatibility; for example, common application frameworks such as Ruby, Node.js, and others.	`-ReverseForwarders`
DNS Server Role	`-Packages Microsoft-NanoServer-DNS-Package`
Desired State Configuration (DSC)	`-Packages Microsoft-NanoServer-DSC-Package`
Internet Information Server (IIS)	`-Packages Microsoft-NanoServer-IIS-Package`
Host Support for Windows Containers	`-Containers`

System Center Virtual Machine Manager Agent	`-Packages Microsoft-Windows-Server-SCVMM-Package`
	`-Packages Microsoft-Windows-Server-SCVMM-Compute-Package`
	Note: Use this package only if you are monitoring Hyper-V. If you install this package, do not use the `-Compute` option for the Hyper-V role; instead use the `-Packages` option to install `-Packages Microsoft-NanoServer-Compute-Package, Microsoft-Windows-Server-SCVMM-Compute-Package`.

Network Performance Diagnostics Service (NPDS)	`-Packages Microsoft-NanoServer-NPDS-Package`
Data Center Bridging	`-Packages Microsoft-NanoServer-DCB-Package`

The next sections describe how to configure a Nano Server image with the required packages, and how to add additional device drivers specific to Cavium QLogic devices. They also explain how to use the Nano Server Recovery Console, how to manage a Nano Server remotely, and how to run NTttcp traffic from a Nano Server.

Deploying a Nano Server on a Physical Server

Follow these steps to create a Nano Server virtual hard disk (VHD) that will run on a physical server using the preinstalled device drivers.

To deploy the Nano Server:

1. Download the Windows Server 2016 OS image.
2. Mount the ISO.
3. Copy the following files from the `NanoServer` folder to a folder on your hard drive:

   ❑ `NanoServerImageGenerator.psm1`
   ❑ `Convert-WindowsImage.ps1`

4. Start Windows PowerShell as an administrator.
5. Change directory to the folder where you pasted the files from Step 3.
6. Import the NanoServerImageGenerator script by issuing the following command:

```
Import-Module .\NanoServerImageGenerator.psm1 -Verbose
```

7. To create a VHD that sets a computer name and includes the OEM drivers and Hyper-V, issue the following Windows PowerShell command:

NOTE

This command will prompt you for an administrator password for the new VHD.

```
New-NanoServerImage -DeploymentType Host -Edition
<Standard/Datacenter> -MediaPath <path to root of media>
-BasePath
.\Base -TargetPath .\NanoServerPhysical\NanoServer.vhd
-ComputerName
<computer name> -Compute -Storage -Cluster -OEMDrivers –
Compute
-DriversPath "<Path to Qlogic Driver sets>"
```

Example:

```
New-NanoServerImage -DeploymentType Host -Edition Datacenter
-MediaPath C:\tmp\TP4_iso\Bld_10586_iso
-BasePath ".\Base" -TargetPath
"C:\Nano\PhysicalSystem\Nano_phy_vhd.vhd" -ComputerName
"Nano-server1" –Compute -Storage -Cluster -OEMDrivers
-DriversPath
"C:\Nano\Drivers"
```

In the preceding example, `C:\Nano\Drivers` is the path for Cavium QLogic drivers. This command takes about 10 to 15 minutes to create a VHD file. A sample output for this command is shown here:

```
Windows(R) Image to Virtual Hard Disk Converter for Windows(R) 10
Copyright (C) Microsoft Corporation.  All rights reserved.
Version 10.0.14300.1000.amd64fre.rs1_release_svc.160324-1723
INFO   : Looking for the requested Windows image in the WIM file
INFO   : Image 1 selected (ServerDatacenterNano)...
INFO   : Creating sparse disk...
INFO   : Mounting VHD...
INFO   : Initializing disk...
INFO   : Creating single partition...
INFO   : Formatting windows volume...
INFO   : Windows path (I:) has been assigned.
INFO   : System volume location: I:
INFO   : Applying image to VHD. This could take a while...
INFO   : Image was applied successfully.
INFO   : Making image bootable...
INFO   : Fixing the Device ID in the BCD store on VHD...
INFO   : Drive is bootable.  Cleaning up...
INFO   : Dismounting VHD...
INFO   : Closing Windows image...
INFO   : Done.
Done. The log is at:
C:\Users\ADMINI~1\AppData\Local\Temp\2\NanoServerImageGenerator.log
```

8. Log in as an administrator on the physical server where you want to run the Nano Server VHD.

9. To copy the VHD to the physical server and configure it to boot from the new VHD:

   a. Open the **Computer Management** utility.

   b. In the left pane, expand the **Storage** node, right-click **Disk Management**, and then select **Attach VHD**.

   c. In the **Location** box, enter the VHD file path.

   d. Click **OK**.

   e. Run **`bcdboot d:\windows`**.

   NOTE

   In this example, the VHD is attached under `D:\`.

   f. Right-click **Disk Management** and select **Detach VHD**.

10. Reboot the physical server into the Nano Server VHD.

11. Log in to the Recovery Console using the administrator and password you supplied while running the script in Step 7.

12. Obtain the IP address of the Nano Server computer.

13. Use Windows PowerShell remoting (or other remote management) tool to connect and remotely manage the server.

Deploying a Nano Server in a Virtual Machine

To create a Nano Server virtual hard drive (VHD) to run in a virtual machine:

1. Download the Windows Server 2016 OS image.

2. Go to the `NanoServer` folder from the downloaded file in Step 1.

3. Copy the following files from the `NanoServer` folder to a folder on your hard drive:

   ❑ `NanoServerImageGenerator.psm1`

   ❑ `Convert-WindowsImage.ps1`

4. Start Windows PowerShell as an administrator.

5. Change directory to the folder where you pasted the files from Step 3.

6. Import the NanoServerImageGenerator script by issuing the following command:

```
Import-Module .\NanoServerImageGenerator.psm1 -Verbose
```

7. Issue the following Windows PowerShell command to create a VHD that sets a computer name and includes the Hyper-V guest drivers:

NOTE

This following command will prompt you for an administrator password for the new VHD.

```
New-NanoServerImage -DeploymentType Guest -Edition
<Standard/Datacenter> -MediaPath <path to root of media>
-BasePath
.\Base -TargetPath .\NanoServerPhysical\NanoServer.vhd
-ComputerName
<computer name> -GuestDrivers
```

Example:

```
New-NanoServerImage -DeploymentType Guest -Edition Datacenter
-MediaPath C:\tmp\TP4_iso\Bld_10586_iso
-BasePath .\Base -TargetPath .\Nano1\VM_NanoServer.vhd
-ComputerName
Nano-VM1 -GuestDrivers
```

The preceding command takes about 10 to 15 minutes to create a VHD file. A sample output for this command follows:

```
PS C:\Nano> New-NanoServerImage -DeploymentType Guest -Edition
Datacenter -MediaPath
C:\tmp\TP4_iso\Bld_10586_iso -BasePath .\Base -TargetPath
.\Nano1\VM_NanoServer.vhd -ComputerName Nano-VM1 -GuestDrivers
cmdlet New-NanoServerImage at command pipeline position 1
Supply values for the following parameters:
Windows(R) Image to Virtual Hard Disk Converter for Windows(R) 10
Copyright (C) Microsoft Corporation.  All rights reserved.
Version 10.0.14300. 1000.amd64fre.rs1_release_svc.160324-1723
INFO   : Looking for the requested Windows image in the WIM file
INFO   : Image 1 selected (ServerTuva)...
INFO   : Creating sparse disk...
INFO   : Attaching VHD...
INFO   : Initializing disk...
INFO   : Creating single partition...
INFO   : Formatting windows volume...
INFO   : Windows path (G:) has been assigned.
INFO   : System volume location: G:
```

```
INFO   : Applying image to VHD. This could take a while...
INFO   : Image was applied successfully.
INFO   : Making image bootable...
INFO   : Fixing the Device ID in the BCD store on VHD...
INFO   : Drive is bootable.  Cleaning up...
INFO   : Closing VHD...
INFO   : Deleting pre-existing VHD : Base.vhd...
INFO   : Closing Windows image...
INFO   : Done.
Done. The log is at:
C:\Users\ADMINI~1\AppData\Local\Temp\2\NanoServerImageGenerator.log
```

8. Create a new virtual machine in Hyper-V Manager, and use the VHD created in Step 7.
9. Boot the virtual machine.
10. Connect to the virtual machine in Hyper-V Manager.
11. Log in to the Recovery Console using the administrator and password you supplied while running the script in Step 7.
12. Obtain the IP address of the Nano Server computer.
13. Use Windows PowerShell remoting (or other remote management) tool to connect and remotely manage the server.

Managing a Nano Server Remotely

Options for managing Nano Server remotely include Windows PowerShell, Windows Management Instrumentation (WMI), Windows Remote Management, and Emergency Management Services (EMS). This section describes how to access Nano Server using Windows PowerShell remoting.

Managing a Nano Server with Windows PowerShell Remoting

To manage Nano Server with Windows PowerShell remoting:

1. Add the IP address of the Nano Server to your management computer's list of trusted hosts.

   NOTE

   Use the recovery console to find the server IP address.

2. Add the account you are using to the Nano Server's administrators.
3. (Optional) Enable **CredSSP** if applicable.

Adding the Nano Server to a List of Trusted Hosts

At an elevated Windows PowerShell prompt, add the Nano Server to the list of trusted hosts by issuing the following command:

```
Set-Item WSMan:\localhost\Client\TrustedHosts "<IP address of Nano
Server>"
```

Examples:

```
Set-Item WSMan:\localhost\Client\TrustedHosts "172.28.41.152"
Set-Item WSMan:\localhost\Client\TrustedHosts "*"
```

NOTE

The preceding command sets all host servers as trusted hosts.

Starting the Remote Windows PowerShell Session

At an elevated local Windows PowerShell session, start the remote Windows PowerShell session by issuing the following commands:

```
$ip = "<IP address of Nano Server>"
$user = "$ip\Administrator"
Enter-PSSession -ComputerName $ip -Credential $user
```

You can now run Windows PowerShell commands on the Nano Server as usual. However, not all Windows PowerShell commands are available in this release of Nano Server. To see which commands are available, issue the command `Get-Command -CommandType Cmdlet`. To stop the remote session, issue the command `Exit-PSSession`.

For more details on Nano Server, go to:

https://technet.microsoft.com/en-us/library/mt126167.aspx

Managing Cavium Adapters on a Windows Nano Server

To manage Cavium QLogic adapters in Nano Server environments, refer to the Windows QConvergeConsole GUI and Windows QLogic Control Suite CLI management tools and associated documentation, available from the Cavium Web site.

RoCE Configuration

To manage the Nano Server with Windows PowerShell remoting:

1. Connect to the Nano Server through Windows PowerShell Remoting from another machine. For example:

```
PS C:\Windows\system32> $1p="172.28.41.152"
PS C:\Windows\system32> $user="172.28.41.152\Administrator"
PS C:\Windows\system32> Enter-PSSession -ComputerName $ip
-Credential $user
```

NOTE

In the preceding example, the Nano Server IP address is `172.28.41.152` and the user name is `Administrator`.

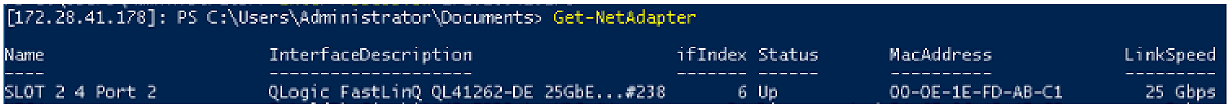
If the Nano Server connects successfully, the following is returned:

```
[172.28.41.152]: PS C:\Users\Administrator\Documents>
```

2. To determine if the drivers are installed and the link is up, issue the following Windows PowerShell command:

```
[172.28.41.152]: PS C:\Users\Administrator\Documents>
Get-NetAdapter
```

Figure 13-18 shows example output.



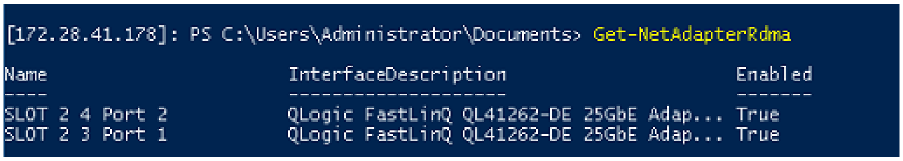
```
[172.28.41.178]: PS C:\Users\Administrator\Documents> Get-NetAdapter

Name                      InterfaceDescription                    ifIndex Status       MacAddress             LinkSpeed
----                      --------------------                    ------- ------       ----------             ---------
SLOT 2 4 Port 2           QLogic FastLinQ QL41262-DE 25GbE...#238       6 Up           00-0E-1E-FD-AB-C1        25 Gbps
```

Figure 13-18. Windows PowerShell Command: Get-NetAdapter

3. To verify whether the RDMA is enabled on the adapter, issue the following Windows PowerShell command:

```
[172.28.41.152]: PS C:\Users\Administrator\Documents>
Get-NetAdapterRdma
```

Figure 13-19 shows example output.



```
[172.28.41.178]: PS C:\Users\Administrator\Documents> Get-NetAdapterRdma

Name                       InterfaceDescription                     Enabled
----                       --------------------                     -------
SLOT 2 4 Port 2            QLogic FastLinQ QL41262-DE 25GbE Adap... True
SLOT 2 3 Port 1            QLogic FastLinQ QL41262-DE 25GbE Adap... True
```

Figure 13-19. Windows PowerShell Command: Get-NetAdapterRdma

4. To assign an IP address and VLAN ID to all interfaces of the adapter, issue the following Windows PowerShell commands:

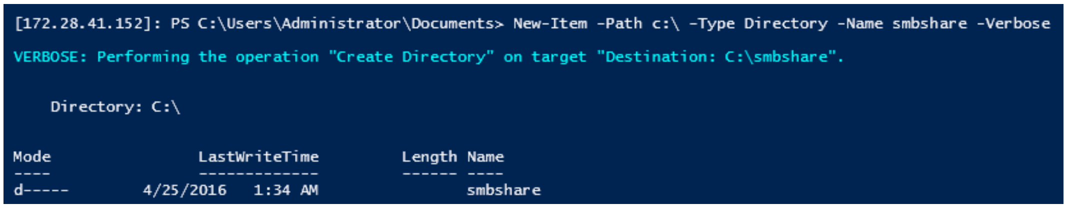
```
[172.28.41.152]: PS C:\> Set-NetAdapterAdvancedProperty
-InterfaceAlias "slot 1 port 1" -RegistryKeyword vlanid
-RegistryValue 5
```

```
[172.28.41.152]: PS C:\> netsh interface ip set address
name="SLOT 1 Port 1" static 192.168.10.10 255.255.255.0
```

5. To create SMBShare on the Nano Server, issue the following Windows PowerShell command:

```
[172.28.41.152]: PS C:\Users\Administrator\Documents>
New-Item -Path c:\ -Type Directory -Name smbshare -Verbose
```

Figure 13-20 shows example output.

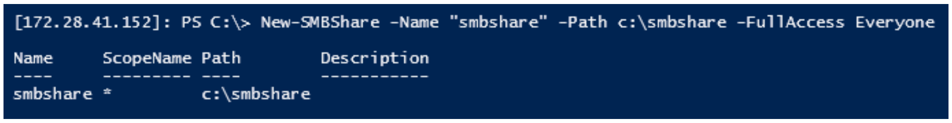


```
[172.28.41.152]: PS C:\Users\Administrator\Documents> New-Item -Path c:\ -Type Directory -Name smbshare -Verbose
VERBOSE: Performing the operation "Create Directory" on target "Destination: C:\smbshare".

    Directory: C:\

Mode                LastWriteTime         Length Name
----                -------------         ------ ----
d-----        4/25/2016   1:34 AM                smbshare
```

Figure 13-20. Windows PowerShell Command: New-Item

```
[172.28.41.152]: PS C:\> New-SMBShare -Name "smbshare" -Path
c:\smbshare -FullAccess Everyone
```

Figure 13-21 shows example output.



```
[172.28.41.152]: PS C:\> New-SMBShare -Name "smbshare" -Path c:\smbshare -FullAccess Everyone

Name     ScopeName Path        Description
----     --------- ----        -----------
smbshare *         c:\smbshare
```

Figure 13-21. Windows PowerShell Command: New-SMBShare

6. To map the SMBShare as a network drive in the client machine, issue the following Windows PowerShell command:

NOTE

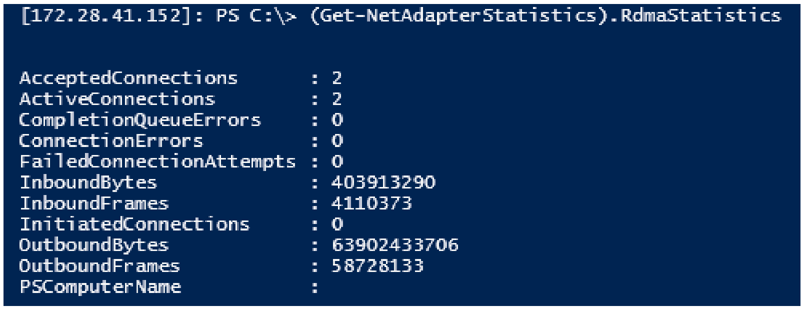
The IP address of an interface on the Nano Server is 192.168.10.10.

```
PS C:\Windows\system32> net use z: \\192.168.10.10\smbshare
This command completed successfully.
```

7. To perform read/write on SMBShare and check RDMA statistics on the Nano Sever, issue the following Windows PowerShell command:

```
[172.28.41.152]: PS C:\>
(Get-NetAdapterStatistics).RdmaStatistics
```

Figure 13-22 shows the command output.



```
[172.28.41.152]: PS C:\> (Get-NetAdapterStatistics).RdmaStatistics


AcceptedConnections       : 2
ActiveConnections         : 2
CompletionQueueErrors     : 0
ConnectionErrors          : 0
FailedConnectionAttempts  : 0
InboundBytes              : 403913290
InboundFrames             : 4110373
InitiatedConnections      : 0
OutboundBytes             : 63902433706
OutboundFrames            : 58728133
PSComputerName            :
```

Figure 13-22. Windows PowerShell Command: Get-NetAdapterStatistics

14 Troubleshooting

This chapter provides the following troubleshooting information:

Troubleshooting Checklist

CAUTION

Before you open the server cabinet to add or remove the adapter, review the "Safety Precautions" on page 5.

The following checklist provides recommended actions to resolve problems that may arise while installing the 41*xxx* Series Adapter or running it in your system.

- Inspect all cables and connections. Verify that the cable connections at the network adapter and the switch are attached properly.
- Verify the adapter installation by reviewing "Installing the Adapter" on page 6. Ensure that the adapter is properly seated in the slot. Check for specific hardware problems, such as obvious damage to board components or the PCI edge connector.
- Verify the configuration settings and change them if they are in conflict with another device.
- Verify that your server is using the latest BIOS.
- Try inserting the adapter in another slot. If the new position works, the original slot in your system may be defective.

- Replace the failed adapter with one that is known to work properly. If the second adapter works in the slot where the first one failed, the original adapter is probably defective.
- Install the adapter in another functioning system, and then run the tests again. If the adapter passes the tests in the new system, the original system may be defective.
- Remove all other adapters from the system, and then run the tests again. If the adapter passes the tests, the other adapters may be causing contention.

Verifying that Current Drivers Are Loaded

Ensure that the current drivers are loaded for your Windows, Linux, or VMware system.

Verifying Drivers in Windows

See the Device Manager to view vital information about the adapter, link status, and network connectivity.

Verifying Drivers in Linux

To verify that the qed.ko driver is loaded properly, issue the following command:

```
# lsmod | grep -i <module name>
```

If the driver is loaded, the output of this command shows the size of the driver in bytes. The following example shows the drivers loaded for the qed module:

```
# lsmod | grep -i qed
qed                  199238  1
qede                1417947  0
```

If you reboot after loading a new driver, you can issue the following command to verify that the currently loaded driver is the correct version:

```
modinfo qede
```

Or, you can issue the following command:

```
[root@test1]# ethtool -i eth2
driver: qede
version: 8.4.7.0
firmware-version: mfw 8.4.7.0 storm 8.4.7.0
bus-info: 0000:04:00.2
```

If you loaded a new driver, but have not yet rebooted, the `modinfo` command will not show the updated driver information. Instead, issue the following `dmesg` command to view the logs. In this example, the last entry identifies the driver that will be active upon reboot.

```
# dmesg | grep -i "Cavium" | grep -i "qede"
```

```
[   10.097526] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[   23.093526] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[   34.975396] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[   34.975896] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[ 3334.975896] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
```

Verifying Drivers in VMware

To verify that the VMware ESXi drivers are loaded, issue the following command:

```
# esxcli software vib list
```

Testing Network Connectivity

This section provides procedures for testing network connectivity in Windows and Linux environments.

NOTE

When using forced link speeds, verify that both the adapter and the switch are forced to the same speed.

Testing Network Connectivity for Windows

Test network connectivity using the `ping` command.

To determine if the network connection is working:

1. Click **Start**, and then click **Run**.
2. In the **Open** box, type **`cmd`**, and then click **OK**.
3. To view the network connection to be tested, issue the following command:

   ```
   ipconfig /all
   ```

4. Issue the following command, and then press ENTER.

   ```
   ping <ip_address>
   ```

The displayed ping statistics indicate whether or not the network connection is working.

Testing Network Connectivity for Linux

To verify that the Ethernet interface is up and running:

1. To check the status of the Ethernet interface, issue the `ifconfig` command.
2. To check the statistics on the Ethernet interface, issue the `netstat -i` command.

To verify that the connection has been established:

1. Ping an IP host on the network. From the command line, issue the following command:

   ```
   ping <ip_address>
   ```

2. Press ENTER.

The displayed ping statistics indicate whether or not the network connection is working.

The adapter link speed can be forced to 10Gbps or 25Gbps using either the operating system GUI tool or the ethtool command, `ethtool -s ethX speed SSSS`.

Microsoft Virtualization with Hyper-V

Microsoft Virtualization is a hypervisor virtualization system for Windows Server 2012 R2. For more information on Hyper-V, go to:

https://technet.microsoft.com/en-us/library/Dn282278.aspx

Linux-specific Issues

Problem: Errors appear when compiling driver source code.

Solution: Some installations of Linux distributions do not install the development tools and kernel sources by default. Before compiling driver source code, ensure that the development tools for the Linux distribution that you are using are installed.

Miscellaneous Issues

Problem: The 41*xxx* Series Adapter has shut down, and an error message appears indicating that the fan on the adapter has failed.

Solution: The 41*xxx* Series Adapter may intentionally shut down to prevent permanent damage. Contact Cavium Technical Support for assistance.

Problem: In an ESXi environment, with the iSCSI driver (qedil) installed, sometimes, the VI-client cannot access the host. This is due to the termination of the hostd daemon, which affects connectivity with the VI-client.

Solution: Contact VMware technical support.

Collecting Debug Data

Use the commands in Table 14-1 to collect debug data.

Table 14-1. Collecting Debug Data Commands

Debug Data	Description
`demesg-T`	Kernel logs
`ethtool-d`	Register dump
`sys_info.sh`	System information; available in the driver bundle

A Adapter LEDS

Table A-1 lists the LED indicators for the state of the adapter port link and activity.

Table A-1. Adapter Port Link and Activity LEDs

Port LED	LED Appearance	Network State
Link LED	Off Continuously illuminated	No link (cable disconnected) Link
Activity LED	Off Blinking	No port activity Port activity

B Cables and Optical Modules

This appendix provides the following information for the supported cables and optical modules:

- Supported Specifications
- "Tested Cables and Optical Modules" on page 274
- "Tested Switches" on page 278

Supported Specifications

The 41*xxx* Series Adapters support a variety of cables and optical modules that comply with SFF8024. Specific form factor compliance is as follows:

- SFPs:
  - SFF8472 (for memory map)
  - SFF8419 or SFF8431 (low speed signals and power)
- Optical modules electrical input/output, active copper cables (ACC), and active optical cables (AOC):
  - 10G—SFF8431 limiting interface
  - 25G—IEEE 802.3by Annex 109B (25GAUI) (does not support RS-FEC)

Tested Cables and Optical Modules

Cavium does not guarantee that every cable or optical module that satisfies the compliance requirements will operate with the 41*xxx* Series Adapters. Cavium has tested the components listed in Table B-1 and presents this list for your convenience.

Table B-1. Tested Cables and Optical Modules

Speed/Form Factor	Manufac-turer	Part Number	Type	Cable Length[a]	Gauge
Cables					
10G DAC [b]	Brocade®	1539W	SFP+10G-to-SFP+10G	1	26
		V239T	SFP+10G-to-SFP+10G	3	26
		48V40	SFP+10G-to-SFP+10G	5	26
	Cisco	H606N	SFP+10G-to-SFP+10G	1	26
		K591N	SFP+10G-to-SFP+10G	3	26
		G849N	SFP+10G-to-SFP+10G	5	26
	Dell	V250M	SFP+10G-to-SFP+10G	1	26
		53HVN	SFP+10G-to-SFP+10G	3	26
		358VV	SFP+10G-to-SFP+10G	5	26
		407-BBBK	SFP+10G-to-SFP+10G	1	30
		407-BBBI	SFP+10G-to-SFP+10G	3	26
		407-BBBP	SFP+10G-to-SFP+10G	5	26
25G DAC	Amphenol®	NDCCGF0001	SFP28-25G-to-SFP28-25G	1	30
		NDCCGF0003	SFP28-25G-to-SFP28-25G	3	30
		NDCCGJ0003	SFP28-25G-to-SFP28-25G	3	26
		NDCCGJ0005	SFP28-25G-to-SFP28-25G	5	26
	Dell	2JVDD	SFP28-25G-to-SFP28-25G	1	26
		D0R73	SFP28-25G-to-SFP28-25G	2	26
		OVXFJY	SFP28-25G-to-SFP28-25G	3	26
		9X8JP	SFP28-25G-to-SFP28-25G	5	26

Table B-1. Tested Cables and Optical Modules (Continued)

Speed/Form Factor	Manufac-turer	Part Number	Type	Cable Length[a]	Gauge
40G Copper QSFP Splitter (4 × 10G)	Dell	TCPM2	QSFP+40G-to-4xSFP+10G	1	30
		27GG5	QSFP+40G-to-4xSFP+10G	3	30
		P8T4W	QSFP+40G-to-4xSFP+10G	5	26
1G Copper RJ45 Transceiver	Dell	8T47V	SFP+ to 1G RJ	1G RJ45	N/A
		XK1M7	SFP+ to 1G RJ	1G RJ45	N/A
		XTY28	SFP+ to 1G RJ	1G RJ45	N/A
10G Copper RJ45 Transceiver	Dell	PGYJT	SFP+ to 10G RJ	10G RJ45	N/A
40G DAC Splitter (4 × 10G)	Dell	470-AAVO	QSFP+40G-to-4xSFP+10G	1	26
		470-AAXG	QSFP+40G-to-4xSFP+10G	3	26
		470-AAXH	QSFP+40G-to-4xSFP+10G	5	26
100G DAC Splitter (4 × 25G)	Amphenol	NDAQGJ-0001	QSFP28-100G-to-4xSFP28-25G	1	26
		NDAQGF-0002	QSFP28-100G-to-4xSFP28-25G	2	30
		NDAQGF-0003	QSFP28-100G-to-4xSFP28-25G	3	30
		NDAQGJ-0005	QSFP28-100G-to-4xSFP28-25G	5	26
	Dell	026FN3 Rev A00	QSFP28-100G-to-4XSFP28-25G	1	26
		0YFNDD Rev A00	QSFP28-100G-to-4XSFP28-25G	2	26
		07R9N9 Rev A00	QSFP28-100G-to-4XSFP28-25G	3	26
	FCI	10130795-4050LF	QSFP28-100G-to-4XSFP28-25G	5	26

Table B-1. Tested Cables and Optical Modules (Continued)

Speed/Form Factor	Manufac-turer	Part Number	Type	Cable Length[a]	Gauge
Optical Solutions					
10G Optical Transceiver	Avago®	AFBR-703SMZ	SFP+ SR	N/A	N/A
		AFBR-701SDZ	SFP+ LR	N/A	N/A
	Dell	Y3KJN	SFP+ SR	1G/10G	N/A
		WTRD1	SFP+ SR	10G	N/A
		3G84K	SFP+ SR	10G	N/A
		RN84N	SFP+ SR	10G-LR	N/A
	Finisar®	FTLX8571D3BCL-QL	SFP+ SR	N/A	N/A
		FTLX1471D3BCL-QL	SFP+ LR	N/A	N/A
25G Optical Transceiver	Dell	P7D7R	SFP28 Optical Transceiver SR	25G SR	N/A
	Finisar	FTLF8536P4BCL	SFP28 Optical Transceiver SR	N/A	N/A
		FTLF8538P4BCL	SFP28 Optical Transceiver SR no FEC	N/A	N/A

Table B-1. Tested Cables and Optical Modules (Continued)

Speed/Form Factor	Manufacturer	Part Number	Type	Cable Length[a]	Gauge
10G AOC [c]	Dell	470-ABLV	SFP+ AOC	2	N/A
		470-ABLZ	SFP+ AOC	3	N/A
		470-ABLT	SFP+ AOC	5	N/A
		470-ABML	SFP+ AOC	7	N/A
		470-ABLU	SFP+ AOC	10	N/A
		470-ABMD	SFP+ AOC	15	N/A
		470-ABMJ	SFP+ AOC	20	N/A
		YJF03	SFP+ AOC	2	N/A
		P9GND	SFP+ AOC	3	N/A
		T1KCN	SFP+ AOC	5	N/A
		1DXKP	SFP+ AOC	7	N/A
		MT7R2	SFP+ AOC	10	N/A
		K0T7R	SFP+ AOC	15	N/A
		W5G04	SFP+ AOC	20	N/A
25G AOC	Dell	X5DH4	SFP28 AOC	20	N/A
	InnoLight®	TF-PY003-N00	SFP28 AOC	3	N/A
		TF-PY020-N00	SFP28 AOC	20	N/A

[a] Cable length is indicated in meters.

[b] DAC is direct attach cable.

[c] AOC is active optical cable.

Tested Switches

Table B-2 lists the switches that have been tested for interoperability with the 41*xxx* Series Adapters. This list is based on switches that are available at the time of product release, and is subject to change over time as new switches enter the market or are discontinued.

Table B-2. Switches Tested for Interoperability

Manufacturer	Ethernet Switch Model
Arista	7060X 7160
Cisco	Nexus 3132 Nexus 3232C Nexus 5548 Nexus 5596T Nexus 6000
Dell EMC	S6100 Z9100
HPE	FlexFabric 5950
Mellanox	SN2410 SN2700

C Dell Z9100 Switch Configuration

The 41*xxx* Series Adapters support connections with the Dell Z9100 Ethernet Switch. However, until the auto-negotiation process is standardized, the switch must be explicitly configured to connect to the adapter at 25Gbps.

To configure a Dell Z9100 switch port to connect to the 41*xxx* Series Adapter at 25Gbps:

1. Establish a serial port connection between your management workstation and the switch.
2. Open a command line session, and then log in to the switch as follows:

   ```
   Login: admin
   Password: admin
   ```

3. Enable configuration of the switch port:

   ```
   Dell> enable
   Password: xxxxxx
   Dell# config
   ```

4. Identify the module and port to be configured. The following example uses module 1, port 5:

   ```
   Dell(conf)#stack-unit 1 port 5 ?
   portmode                Set portmode for a module
   Dell(conf)#stack-unit 1 port 5 portmode ?
   dual                    Enable dual mode
   quad                    Enable quad mode
   single                  Enable single mode
   Dell(conf)#stack-unit 1 port 5 portmode quad ?
   speed                   Each port speed in quad mode
   Dell(conf)#stack-unit 1 port 5 portmode quad speed ?
   10G                     Quad port mode with 10G speed
   ```

```
25G                      Quad port mode with 25G speed
Dell(conf)#stack-unit 1 port 5 portmode quad speed 25G
```

For information about changing the adapter link speed, see "Testing Network Connectivity" on page 269.

5. Verify that the port is operating at 25Gbps:

```
Dell# Dell#show running-config | grep "port 5"
stack-unit 1 port 5 portmode quad speed 25G
```

6. To disable auto-negotiation on switch port 5, follow these steps:

   a. Identify the switch port interface (module 1, port 5, interface 1) and confirm the auto-negotiation status:

   ```
   Dell(conf)#interface tw 1/5/1

   Dell(conf-if-tf-1/5/1)#intf-type cr4 ?
   autoneg                  Enable autoneg
   ```

   b. Disable auto-negotiation:

   ```
   Dell(conf-if-tf-1/5/1)#no intf-type cr4 autoneg
   ```

   c. Verify that auto-negotiation is disabled.

   ```
   Dell(conf-if-tf-1/5/1)#do show run interface tw 1/5/1
   !
   interface twentyFiveGigE 1/5/1
   no ip address
   mtu 9416
   switchport
   flowcontrol rx on tx on
   no shutdown
   no intf-type cr4 autoneg
   ```

For more information about configuring the Dell Z9100 switch, refer to the *Dell Z9100 Switch Configuration Guide* on the Dell Support Web site:

support.dell.com

D Feature Constraints

This appendix provides information about feature constraints implemented in the current release.

These feature coexistence constraints may be removed in a future release. At that time, you should be able to use the feature combinations without any additional configuration steps beyond what would be usually required to enable the features.

Concurrent FCoE and iSCSI Is Not Supported on the Same Port in NPAR Mode

The current release does not support configuration of both FCoE and iSCSI on PFs belonging to the same physical port when in NPAR Mode (concurrent FCoE and iSCSI *is* only supported on the same port in Default Mode). Either FCoE or iSCSI is allowed on a physical port in NPAR Mode.

After a PF with either iSCSI or FCoE personality has been configured on a port using either HII or Cavium QLogic management tools, configuration of the storage protocol on another PF is disallowed by those management tools.

Because storage personality is disabled by default, only the personality that has been configured using HII or Cavium QLogic management tools is written in NVRAM configuration. When this limitation is removed, users can configure additional PFs on the same port for storage in NPAR Mode.

Concurrent RoCE and iWARP Is Not Supported on the Same Port

RoCE and iWARP are not supported on the same port. HII and Cavium QLogic management tools do not allow users to configure both concurrently.

NIC and SAN Boot to Base Is Supported Only on Select PFs

Ethernet and PXE boot are currently supported only on PF0 and PF1. In NPAR configuration, other PFs do not support Ethernet and PXE boot.

- When the **Virtualization Mode** is set to **NPAR**, non-offloaded FCoE boot is supported on Partition 2 (PF2 and PF3) and iSCSI boot is supported on Partition 3 (PF4 and PF5). iSCSI and FCoE boot is limited to a single target per boot session. iSCSI boot target LUN support is limited to LUN ID 0 only.
- When the **Virtualization Mode** is set to **None** or **SR-IOV**, boot from SAN is not supported.

Glossary

ACPI

The *Advanced Configuration and Power Interface (ACPI)* specification provides an open standard for unified operating system-centric device configuration and power management. The ACPI defines platform-independent interfaces for hardware discovery, configuration, power management, and monitoring. The specification is central to operating system-directed configuration and Power Management (OSPM), a term used to describe a system implementing ACPI, which therefore removes device management responsibilities from legacy firmware interfaces.

adapter

The board that interfaces between the host system and the target devices. Adapter is synonymous with Host Bus Adapter, host adapter, and board.

adapter port

A port on the adapter board.

Advanced Configuration and Power Interface

See ACPI.

bandwidth

A measure of the volume of data that can be transmitted at a specific transmission rate. A 1Gbps or 2Gbps Fibre Channel port can transmit or receive at nominal rates of 1 or 2Gbps, depending on the device to which it is connected. This corresponds to actual bandwidth values of 106MB and 212MB, respectively.

BAR

Base address register. Used to hold memory addresses used by a device, or offsets for port addresses. Typically, memory address BARs must be located in physical RAM while I/O space BARs can reside at any memory address (even beyond physical memory).

base address register

See BAR.

basic input output system

See BIOS.

BIOS

Basic input output system. Typically in Flash PROM, the program (or utility) that serves as an interface between the hardware and the operating system and allows booting from the adapter at startup.

challenge-handshake authentication protocol

See CHAP.

CHAP

Challenge-handshake authentication protocol (CHAP) is used for remote logon, usually between a client and server or a Web browser and Web server. A challenge/response is a security mechanism for verifying the identity of a person or process without revealing a secret password that is shared by the two entities. Also referred to as a *three-way handshake*.

data center bridging

See DCB.

data center bridging exchange

See DCBX.

DCB

Data center bridging. Provides enhancements to existing 802.1 bridge specifications to satisfy the requirements of protocols and applications in the data center. Because existing high-performance data centers typically comprise multiple application-specific networks that run on different link layer technologies (Fibre Channel for storage and Ethernet for network management and LAN connectivity), DCB enables 802.1 bridges to be used for the deployment of a converged network where all applications can be run over a single physical infrastructure.

DCBX

Data center bridging exchange. A protocol used by DCB devices to exchange configuration information with directly connected peers. The protocol may also be used for misconfiguration detection and for configuration of the peer.

device

A target, typically a disk drive. Hardware such as a disk drive, tape drive, printer, or keyboard that is installed in or connected to a system. In Fibre Channel, a *target device*.

DHCP

Dynamic host configuration protocol. Enables computers on an IP network to extract their configuration from servers that have information about the computer only after it is requested.

driver

The software that interfaces between the file system and a physical data storage device or network media.

dynamic host configuration protocol

See DHCP.

eCore

A layer between the OS and the hardware and firmware. It is device-specific and OS-agnostic. When eCore code requires OS services (for example, for memory allocation, PCI configuration space access, and so on) it calls an abstract OS function that is implemented in OS-specific layers. eCore flows may be driven by the hardware (for example, by an interrupt) or by the OS-specific portion of the driver (for example, loading and unloading the load and unload).

EEE

Energy-efficient Ethernet. A set of enhancements to the twisted-pair and backplane Ethernet family of computer networking standards that allows for less power consumption during periods of low data activity. The intention was to reduce power consumption by 50 percent or more, while retaining full compatibility with existing equipment. The Institute of Electrical and Electronics Engineers (IEEE), through the IEEE 802.3az task force, developed the standard.

EFI

Extensible firmware interface. A specification that defines a software interface between an operating system and platform firmware. EFI is a replacement for the older BIOS firmware interface present in all IBM PC-compatible personal computers.

energy-efficient Ethernet

See EEE.

enhanced transmission selection

See ETS.

Ethernet

The most widely used LAN technology that transmits information between computers, typically at speeds of 10 and 100 million bits per second (Mbps).

ETS

Enhanced transmission selection. A standard that specifies the enhancement of transmission selection to support the allocation of bandwidth among traffic classes. When the offered load in a traffic class does not use its allocated bandwidth, enhanced transmission selection allows other traffic classes to use the available bandwidth. The bandwidth-allocation priorities coexist with strict priorities. ETS includes managed objects to support bandwidth allocation. For more information, refer to:

http://ieee802.org/1/pages/802.1az.html

extensible firmware interface

See EFI.

FCoE

Fibre Channel over Ethernet. A new technology defined by the T11 standards body that allows traditional Fibre Channel storage networking traffic to travel over an Ethernet link by encapsulating Fibre Channel frames inside Layer 2 Ethernet frames. For more information, visit www.fcoe.com.

Fibre Channel over Ethernet

See FCoE.

file transfer protocol

See FTP.

FTP

File transfer protocol. A standard network protocol used to transfer files from one host to another host over a TCP-based network, such as the Internet. FTP is required for out-of-band firmware uploads that will complete faster than in-band firmware uploads.

human interface infrastructure

See HII.

HII

Human interface infrastructure. A specification (part of UEFI 2.1) for managing user input, localized strings, fonts, and forms, that allows OEMs to develop graphical interfaces for preboot configuration.

IEEE

Institute of Electrical and Electronics Engineers. An international nonprofit organization for the advancement of technology related to electricity.

Internet Protocol

See IP.

Internet small computer system interface

See iSCSI.

Internet wide area RDMA protocol

See iWARP.

IP

Internet protocol. A method by which data is sent from one computer to another over the Internet. IP specifies the format of packets, also called *datagrams*, and the addressing scheme.

IQN

iSCSI qualified name. iSCSI node name based on the initiator manufacturer and a unique device name section.

iSCSI

Internet small computer system interface. Protocol that encapsulates data into IP packets to send over Ethernet connections.

iSCSI qualified name

See IQN.

iWARP

Internet wide area RDMA protocol. A networking protocol that implements RDMA for efficient data transfer over IP networks. iWARP is designed for multiple environments, including LANs, storage networks, data center networks, and WANs.

jumbo frames

Large IP frames used in high-performance networks to increase performance over long distances. Jumbo frames generally means 9,000 bytes for Gigabit Ethernet, but can refer to anything over the IP MTU, which is 1,500 bytes on an Ethernet.

large send offload

See LSO.

Layer 2

Refers to the data link layer of the multilayered communication model, Open Systems Interconnection (OSI). The function of the data link layer is to move data across the physical links in a network, where a switch redirects data messages at the Layer 2 level using the destination MAC address to determine the message destination.

Link Layer Discovery Protocol

See LLDP.

LLDP

A vendor-neutral Layer 2 protocol that allows a network device to advertise its identity and capabilities on the local network. This protocol supersedes proprietary protocols like Cisco Discovery Protocol, Extreme Discovery Protocol, and Nortel Discovery Protocol (also known as SONMP).

Information gathered with LLDP is stored in the device and can be queried using SNMP. The topology of a LLDP-enabled network can be discovered by crawling the hosts and querying this database.

LSO

Large send offload. LSO Ethernet adapter feature that allows the TCP\IP network stack to build a large (up to 64KB) TCP message before sending it to the adapter. The adapter hardware segments the message into smaller data packets (frames) that can be sent over the wire: up to 1,500 bytes for standard Ethernet frames and up to 9,000 bytes for jumbo Ethernet frames. The segmentation process frees up the server CPU from having to segment large TCP messages into smaller packets that will fit inside the supported frame size.

maximum transmission unit

See MTU.

message signaled interrupts

See MSI, MSI-X.

MSI, MSI-X

Message signaled interrupts. One of two PCI-defined extensions to support message signaled interrupts (MSIs), in PCI 2.2 and later and PCI Express. MSIs are an alternative way of generating an interrupt through special messages that allow emulation of a pin assertion or deassertion.

MSI-X (defined in PCI 3.0) allows a device to allocate any number of interrupts between 1 and 2,048 and gives each interrupt separate data and address registers. Optional features in MSI (64-bit addressing and interrupt masking) are mandatory with MSI-X.

MTU

Maximum transmission unit. Refers to the size (in bytes) of the largest packet (IP datagram) that a specified layer of a communications protocol can transfer.

network interface card

See NIC.

NIC

Network interface card. Computer card installed to enable a dedicated network connection.

NIC partitioning

See NPAR.

non-volatile random access memory

See NVRAM.

non-volatile memory express

See NVMe.

NPAR

NIC partitioning. The division of a single NIC port into multiple physical functions or partitions, each with a user-configurable bandwidth and personality (interface type). Personalities include NIC, FCoE, and iSCSI.

NVRAM

Non-volatile random access memory. A type of memory that retains data (configuration settings) even when power is removed. You can manually configure NVRAM settings or restore them from a file.

NVMe

A storage access method designed for sold-state drives (SSDs).

OFED™

OpenFabrics Enterprise Distribution. An open source software for RDMA and kernel bypass applications.

PCI™

Peripheral component interface. A 32-bit local bus specification introduced by Intel®.

PCI Express (PCIe)

A third-generation I/O standard that allows enhanced Ethernet network performance beyond that of the older peripheral component interconnect (PCI) and PCI extended (PCI-X) desktop and server slots.

QoS

Quality of service. Refers to the methods used to prevent bottlenecks and ensure business continuity when transmitting data over virtual ports by setting priorities and allocating bandwidth.

quality of service

See QoS.

PF

Physical function.

RDMA

Remote direct memory access. The ability for one node to write directly to the memory of another (with address and size semantics) over a network. This capability is an important feature of VI networks.

reduced instruction set computer

See RISC.

remote direct memory access

See RDMA.

RISC

Reduced instruction set computer. A computer microprocessor that performs fewer types of computer instructions, thereby operating at higher speeds.

RDMA over Converged Ethernet

See RoCE.

RoCE

RDMA over Converged Ethernet. A network protocol that allows remote direct memory access (RDMA) over a converged or a non-converged Ethernet network. RoCE is a link layer protocol that allows communication between any two hosts in the same Ethernet broadcast domain.

SCSI

Small computer system interface. A high-speed interface used to connect devices, such as hard drives, CD drives, printers, and scanners, to a computer. The SCSI can connect many devices using a single controller. Each device is accessed by an individual identification number on the SCSI controller bus.

SerDes

Serializer/deserializer. A pair of functional blocks commonly used in high-speed communications to compensate for limited input/output. These blocks convert data between serial data and parallel interfaces in each direction.

serializer/deserializer

See SerDes.

single root input/output virtualization

See SR-IOV.

small computer system interface

See SCSI.

SR-IOV

Single root input/output virtualization. A specification by the PCI SIG that enables a single PCIe device to appear as multiple, separate physical PCIe devices. SR-IOV permits isolation of PCIe resources for performance, interoperability, and manageability.

target

The storage-device endpoint of a SCSI session. Initiators request data from targets. Targets are typically disk-drives, tape-drives, or other media devices. Typically a SCSI peripheral device is the target but an adapter may, in some cases, be a target. A target can contain many LUNs.

A target is a device that responds to a requested by an initiator (the host system). Peripherals are targets, but for some commands (for example, a SCSI COPY command), the peripheral may act as an initiator.

TCP

Transmission control protocol. A set of rules to send data in packets over the Internet protocol.

TCP/IP

Transmission control protocol/Internet protocol. Basic communication language of the Internet.

TLV

Type-length-value. Optional information that may be encoded as an element inside of the protocol. The type and length fields are fixed in size (typically 1–4 bytes), and the value field is of variable size. These fields are used as follows:

- Type—A numeric code that indicates the kind of field that this part of the message represents.
- Length—The size of the value field (typically in bytes).
- Value—Variable-sized set of bytes that contains data for this part of the message.

transmission control protocol

See TCP.

transmission control protocol/Internet protocol

See TCP/IP.

type-length-value

See TLV.

UDP

User datagram protocol. A connectionless transport protocol without any guarantee of packet sequence or delivery. It functions directly on top of IP.

UEFI

Unified extensible firmware interface. A specification detailing an interface that helps hand off control of the system for the preboot environment (that is, after the system is powered on, but before the operating system starts) to an operating system, such as Windows or Linux. UEFI provides a clean interface between operating systems and platform firmware at boot time, and supports an architecture-independent mechanism for initializing add-in cards.

unified extensible firmware interface

See UEFI.

user datagram protocol

See UDP.

VF

Virtual function.

VI

Virtual interface. An initiative for remote direct memory access across Fibre Channel and other communication protocols. Used in clustering and messaging.

virtual interface

See VI.

virtual logical area network

See VLAN.

virtual machine

See VM.

VLAN

Virtual logical area network (LAN). A group of hosts with a common set of requirements that communicate as if they were attached to the same wire, regardless of their physical location. Although a VLAN has the same attributes as a physical LAN, it allows for end stations to be grouped together even if they are not located on the same LAN segment. VLANs enable network reconfiguration through software, instead of physically relocating devices.

VM

Virtual machine. A software implementation of a machine (computer) that executes programs like a real machine.

wake on LAN

See WoL.

WoL

Wake on LAN. An Ethernet computer networking standard that allows a computer to be remotely switched on or awakened by a network message sent usually by a simple program executed on another computer on the network.

Corporate Headquarters Cavium, Inc. 2315 N. First Street San Jose, CA 95131 408-943-7100

International Offices UK | Ireland | Germany | France | India | Japan | China | Hong Kong | Singapore | Taiwan | Israel

This document is provided for informational purposes only and may contain errors. Cavium reserves the right, without notice, to make changes to this document or in product design or specifications. Cavium disclaims any warranty of any kind, expressed or implied, and does not guarantee that any results or performance described in the document will be achieved by you. All statements regarding Cavium's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.