User Manual for Mellanox ConnectX®-3, ConnectX®-3 Pro, ConnectX®-4, ConnectX®-4 Lx and ConnectX®-5 Ex Ethernet Adapters for Dell EMC PowerEdge Servers

Rev 1.8

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Table of Contents

List of Tables

List of Figures

Revision History

This document was printed on October 25, 2018.

Table 1 - Revision History Table

Date	Rev	Comments/Changes
September 2018	1.8	• Added Mellanox ConnectX®-5 Ex Dual Port 100GbE QSFP cards support across document. • Updated Functional Description on page 16 • Updated Features on page 17 • Updated Adapter Card Interfaces on page 22 • Updated LED Assignments and Bracket Mechanical Drawings on page 23 • Updated Linux Driver Features on page 41 • Updated WinOF / WinOF-2 Features on page 95 • Added Mellanox ConnectX-5 Ex Dual Port 100GbE QSFP Network Adapter Specifications on page 202. • Updated Main Configuration Page - NIC Configuration on page 210
June 2018	1.7	• Added ConnectX®-4 LX Dual Port 25 GbE KR Mezzanine Card support across document. • Updated Functional Description on page 16 • Updated Features on page 17 • Updated Adapter Card Interfaces on page 22 • Updated Uninstalling Mellanox WinOF / WinOF-2 Driver on page 94 • Updated Data Center Bridging Exchange (DCBX) on page 107 • Added Mellanox ConnectX-4 Lx Dual Port 25GbE KR Mezzanine Card Specifications on page 201. • Updated Linux on page 189.
December 2017	1.6	• Updated "Linux Driver Features" with the following: • Added Enabling/Disabling RoCE on VFs (ConnectX-4, ConnectX-4 Lx and Connect X-5 Ex) on page 42. • Added Flow Steering Dump Tool on page 65. • Added the following sections in "WinOF / WinOF-2 Features": • Performance Tuning and Counters on page 123. • Differentiated Services Code Point (DSCP) on page 101. • Configuring the Ethernet Driver on page 106. • Receive Segment Coalescing (RSC) on page 106. • Receive Side Scaling (RSS) on page 106. • Wake on LAN (WoL) on page 107. • Data Center Bridging Exchange (DCBX) on page 107. • Receive Path Activity Monitoring on page 110. • Head of Queue Lifetime Limit on page 110. • Threaded DPC on page 111. • Performance Tuning and Counters on page 123. • Updated the following specification tables: • Mellanox ConnectX-4 Dual Port 100 GbE QSFP Network Adapter Specifications on page 198. • Mellanox ConnectX-4 Lx Dual Port SFP28 25GbE for Dell Rack NDC on page 199 • Mellanox ConnectX-4 Lx Dual 25GbE SFP28 Network Adapter Specifications on page 200 • Updated Troubleshooting on page 188. • Added Wake on LAN Configuration on page 219.

Date	Rev	Comments/Changes
May 2016	1.5	• Added ConnectX-4 support across document. • Updated the document's title. • Updated About this Manual on page 15. • Updated Functional Description on page 16. • Updated Features on page 17. • Updated Adapter Card Interfaces on page 22. • Updated Installing the Hardware on page 32. • Updated Driver Installation and Configuration on page 35 and Linux Driver Features on page 41. • Updated Remote Boot on page 164 • Updated Firmware on page 186 • Updated Troubleshooting on page 188 • Added Mellanox ConnectX-4 Dual Port 100 GbE QSFP Network Adapter Specifications on page 198. • Updated Mellanox ConnectX-4 Lx Dual 25GbE SFP28 Network Adapter Specifications on page 200 • Updated Configuration for Mellanox Adapters through System Setup on page 206
July 2016	1.4	• Added ConnectX-4 Lx support across document. • Updated the document's title. • Updated About this Manual on page 15. • Updated Functional Description on page 16. • Updated Features on page 17. • Updated Adapter Card Interfaces on page 22. • Updated Installing the Hardware on page 32. • Updated Linux Driver on page 35 and Linux Driver Features on page 41. • Updated VMware Driver for ConnectX-3 and ConnectX-3 Pro on page 89 • Added VMware Driver for ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex on page 91. • Updated Linux Driver on page 35 and WinOF / WinOF-2 Features on page 95. • Updated Booting Windows from an iSCSI Target on page 171. • Added Mellanox ConnectX-4 Lx Dual Port SFP28 25GbE for Dell Rack NDC on page 199. • Added Mellanox ConnectX-4 Lx Dual 25GbE SFP28 Network Adapter Specifications on page 200. • Updated Mellanox ConnectX-4 Lx Dual 25GbE SFP28 Network Adapter Specifications on page 200. • Updated Remote Boot on page 164. • Updated Configuration for Mellanox Adapters through System Setup on page 206.
August 2015	1.3	• Added ConnectX-3 Pro support across document. • Added VXLAN Hardware Stateless Offloads on page 65 • Added SectionNetwork Virtualization Generic Routing Encapsulation (NVGRE),. • Updated Performance Tuning and Counters on page 123 • Updated iSCSI Boot on page 164 • Added Mellanox ConnectX-3 Pro Dual 40GbE QSFP+ Network Adapter Specifications on page 195, Mellanox ConnectX-3 Pro Dual 10GbE SFP+ Network Adapter Specifications on page 196 and Mellanox ConnectX-3 Pro Dual 10GbE KR Blade Mezzanine Card Specifications on page 197 • Added Network Adapter Certification for ConnectX-3 Pro. See Ethernet Network Adapter Certifications on page 203.

Date	Rev	Comments/Changes
March 2015	1.2	• Updated installation script in Installation Procedure on page 37. • Updated SR-IOV VFs recommendation to less than 63. See Setting Up SR-IOV. • Updated Configuration for Mellanox Adapters through System Setup on page 206.
August 2014	1.1	• Added Linux Driver Features on page 41 • Added WinOF / WinOF-2 Features on page 95 • Added Remote Boot on page 127 • Added Configuration for Mellanox Adapters through System Setup on page 206
November 2013	1.0	Initial Release

About this Manual

This *User Manual* describes Mellanox Technologies ConnectX-3/ConnectX-3 Pro 10/40GbE, ConnectX-4 100GbE, ConnectX-4 Lx 25GbE and ConnectX-5 Ex 100GbE adapter cards for Dell EMC PowerEdge Servers. It provides details as to the interfaces of the board, specifications, required software and firmware for operating the board, and relevant documentation.

Intended Audience

This manual is intended for the installer and user of these cards.

The manual assumes the user has basic familiarity with Ethernet networks and architecture specifications.

Related Documentation

Table 2 - Documents List

IEEE Std 802.3 Specification	This is the IEEE Ethernet specification http://standards.ieee.org/getieee802
PCI Express 3.0 Specifications	Industry Standard PCI Express 3.0 Base and PCI_Express_CE-M_r3.0

Document Conventions

This document uses the following conventions:

• MB and MBytes are used to mean size in mega Bytes. The use of Mb or Mbits (small b) indicates size in mega bits.

• PCIe is used to mean PCI Express

Technical Support

Dell Support site: http://www.dell.com/support

1 Introduction

1.1 Functional Description

Mellanox Ethernet adapters utilizing IBTA RoCE technology provide efficient RDMA services, delivering high performance to bandwidth and latency sensitive applications. Applications utilizing TCP/UDP/IP transport can achieve industry-leading throughput over 10, 25, 40 or 100GbE. The hardware-based stateless offload and flow steering engines in Mellanox adapters reduce the CPU overhead of IP packet transport, freeing more processor cycles to work on the application. Sockets acceleration software further increases performance for latency sensitive applications. Table 3 lists Dell EMC PowerEdge Products covered in this User Manual.

> The following products are customized products for use in Dell EMC PowerEdge servers.

Table 3 - Dell EMC PowerEdge Adapter Cards

ConnectX-3 Products
Mellanox ConnectX®-3 Dual Port 40GbE QSFP Network Adapter with Full Height Bracket
Mellanox ConnectX®-3 Dual Port 40GbE QSFP Network Adapter with Low Profile Bracket
Mellanox ConnectX®-3 Dual Port 10GbE SFP+ Network Adapter with Full Height Bracket
Mellanox ConnectX®-3 Dual Port 10GbE SFP+ Network Adapter with Low Profile Bracket
Mellanox ConnectX®-3 Dual Port 10GbE KR Blade Mezzanine Card
ConnectX-3 Pro Products
Mellanox ConnectX®-3 Pro Dual Port QSFP 40GbE Adapter Card with Full Height Bracket
Mellanox ConnectX®-3 Pro Dual Port QSFP 40GbE Adapter Card with Low Profile Bracket
Mellanox ConnectX®-3 Pro Dual Port 10GbE SFP+ Adapter Card with Low Profile Bracket
Mellanox ConnectX®-3 Pro Dual Port 10GbE Mezzanine card
ConnectX-4 Products
Mellanox ConnectX®-4 Dual Port 100GbE QSFP28 Network Adapter Card with Low Profile Bracket
Mellanox ConnectX®-4 Dual Port 100GbE QSF28P Network Adapter Card with Full Height Profile Bracket
ConnectX-4 Lx Products
Mellanox ConnectX®-4 Lx Dual Port 25GbE SFP28 Network Adapter Card with Low Profile Bracket
Mellanox ConnectX®-4 Lx Dual Port 25GbE SFP28 Network Adapter Card with Full Height Bracket
Mellanox ConnectX®-4 Lx Dual Port 25GbE SFP28 Dell Rack NDC
Mellanox ConnectX®-4 Lx Dual Port 25GbE KR Mezzanine Card
ConnectX-5 Ex Products
Mellanox ConnectX®-5 Ex Dual Port 100GbE QSFP Network Adapter with Full Height Bracket
Mellanox ConnectX®-5 Ex Dual Port 100GbE QSFP Network Adapter with Low Profile Bracket

1.2   Features

The adapter cards described in this manual support the following features:

Table 4 - Features

Feature	Sub-Feature	Supported Adapters
Low latency RDMA over Ethernet		ConnectX-3 / ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
Traffic steering across multiple cores		
Intelligent interrupt coalescence		
Advanced Quality of Service		
Dual Ethernet ports		
CPU offload of transport operations		
Application Offload		
End-to-end QoS and congestion control		

Feature	Sub-Feature	Supported Adapters
Ethernet	100GbE / 50GbE	ConnectX-4 / ConnectX-5 Ex
	40GbE / 10GbE / 1GbE	ConnectX-3/ ConnectX-3 Pro / ConnectX-4/ ConnectX-5 Ex
	25GbE / 10GbE / 1GbE	ConnectX-4 / ConnectX-4 Lx/ ConnectX-5 Ex
	25G Ethernet Consortium 25	ConnectX-4 / ConnectX-4 Lx/ ConnectX-5 Ex
	IEEE 802.3ba 40 Gigabit Ethernet	ConnectX-3 / ConnectX-3 Pro Dual Port 40GbE QSFP+ Network Adapter / ConnectX-4 / ConnectX-5 Ex
	IEEE 802.3by 25 Gigabit Ethernet	ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
	IEEE 802.3ae 10 Gigabit Ethernet	ConnectX-3 / ConnectX-3 Pro / ConnectX-4/ ConnectX-4 Lx / ConnectX-5 Ex
	IEEE 802.3x Flow Control	
	IEEE 802.3ap based auto-negotiation and KR startup	
	IEEE 802.3ad, 802.1AX Link Aggregation	
	IEEE 802.1Q, 802.1P VLAN tags and priority	
	IEEE 802.1Qau (QCN) – Congestion Notification	
	IEEE 802.1Qaz (ETS)	
	IEEE 802.1Qbb (PFC)	
	IEEE 802.1Qbg	
	IEEE 802.1Qbh	
	IEEE P802.1Qbb D1.0 Priority-based Flow Control	
	IEEE 1588v2	
	Jumbo frame support (9.6KB)	ConnectX-3/ ConnectX-3 Pro
	128 MAC/VLAN addresses per port	
	127 MAC/VLAN addresses per port	ConnectX-4 / ConnectX-4 Lx/ ConnectX-5 Ex
	Wake on LAN (WoL) supported on Mellanox	ConnectX-3 / ConnectX-3 Pro Dual Port 10GbE KR Blade Mezzanine Card and ConnectX-4 Lx Dual Port 25GbE SFP Rack NDC / Mellanox ConnectX®-4 Lx Dual Port 25GbE KR Mezzanine Card / ConnectX-5 Ex

Feature	Sub-Feature	Supported Adapters
PCI Express Interface	PCIe Base 3.0 compliant, 1.1 and 2.0 compatible	ConnectX-3 / ConnectX-3 Pro/ConnectX-4/ ConnectX-4 Lx / ConnectX-5 Ex
	2.5, 5.0, or 8.0GT/s link rate x1	ConnectX-3 / ConnectX-3 Pro / ConnectX-4/ ConnectX-4 Lx / ConnectX-5 Ex
	Auto-negotiates to x8, x4, or x1	ConnectX-3 / ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
	Auto-negotiates to x16, x8, x4, or x1	ConnectX-4 / ConnectX-5 Ex
	Support for MSI/MSI-X mechanisms	ConnectX-3 / ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
Hardware-based I/O Virtualization	Single Root IOV (SR-IOV)	ConnectX-3 / ConnectX-3 Pro / ConnectX-4/ ConnectX-4 Lx / ConnectX-5 Ex
	Address translation and protection	
	Dedicated adapter resources	
	Multiple queues per virtual machine	
	Enhanced QoS for vNICs	
	VMware NetQueue support	
Additional CPU Offloads	RDMA over Converged Ethernet TCP/ UDP/IP stateless offload	ConnectX-3 / ConnectX-3 Pro / ConnectX-4/ ConnectX-4 Lx / ConnectX-5 Ex
	Intelligent interrupt coalescence	
FlexBoot™ Technology	Remote boot over Ethernet	ConnectX-3 / ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
	iSCSI boot	
	PXE boot	
Overlay Networks	Stateless offloads for overlay networks and tunneling protocols	ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
	Hardware offload of encapsulation and decapsulation of NVGRE and VXLAN overlay networks	

Feature	Sub-Feature	Supported Adapters
Connectivity	Interoperable with 100/50GbE switches	ConnectX-4/ ConnectX-5 Ex
	Interoperable with 1/10/40GbE switches	ConnectX-3 / ConnectX-3 Pro / ConnectX-4/ ConnectX-5 Ex
	Interoperable with 1/10/25bE switches	ConnectX-4 / ConnectX-4 Lx/ ConnectX-5 Ex
	QSFP28 connectors	ConnectX-4/ ConnectX-5 Ex
	QSFP+ connectors	ConnectX-3 / ConnectX-3 Pro Dual Port 40GbE QSFP+ Network Adapter only
	SFP+ connectors	ConnectX-3 / ConnectX-3 Pro Dual Port 10GbE SFP+ Network Adapter only
	SFP28 connectors	ConnectX-4 Lx 25GbE SFP28 Network Adapters only
	Passive copper cable	ConnectX-3 / ConnectX-3 Pro/ConnectX-4/ConnectX-4 Lx/ ConnectX-5 Ex
	Powered connectors for optical and active cable support	ConnectX-3 / ConnectX-3 Pro/ConnectX-4/ConnectX-4 Lx/ ConnectX-5 Ex
	Two IMPEL connectors connected to Two PTMs or Switch Modules	ConnectX®-4 Lx Dual Port 25GbE KR Mezzanine Card
Management and Tools	MIB, MIB-II, MIB-II Extensions, RMON, RMON 2	ConnectX-3 / ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex
	Configuration and diagnostic tools	
RoHS-R6 compliant		ConnectX-3 / ConnectX-3 Pro / ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Ex

1.2.1 Single Root IO Virtualization (SR-IOV)

Single Root IO Virtualization (SR-IOV) is a technology that allows a physical PCIe device to present itself multiple times through the PCIe bus. This technology enables multiple virtual instances of the device with separate resources. ConnectX-3 and ConnectX-3 Pro Mellanox adapters are capable of exposing up to 63 virtual instances called Virtual Functions (VFs). ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards are capable of exposing up to 127 VFs. These virtual functions can then be provisioned separately. Each VF can be seen as an additional device connected to the Physical Function. It shares the same resources with the Physical Function, and its number of ports equals those of the Physical Function. SR-IOV is commonly used in conjunction with an SR-IOV enabled hypervisor to provide virtual machines direct hardware access to network resources, hence increasing its performance.

1.2.2   Remote Direct Memory Access

Remote Direct Memory Access (RDMA) is the remote memory management capability that allows server to server data -movement directly between application memory without any CPU involvement. RDMA over Converged Ethernet (RoCE) is a mechanism which provides this efficient data transfer with very low latencies on loss-less Ethernet networks. With advances in data center convergence over reliable Ethernet, ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex with RoCE use the proven and efficient RDMA transport to provide the platform for deploying RDMA technology in mainstream data center application at 10GigE, 25GigbE, 40GigE and 100GigE link-speed. ConnectX-3, ConnectX-3 Pro, ConnectX-4 ConnectX-4 Lx and ConnectX-5 Ex with their hardware offload support take advantage of this efficient RDMA transport services over Ethernet to deliver ultra low latency for performance-critical and transaction intensive applications such as financial, database, storage, and content delivery networks. RoCE encapsulates IB transport and GRH headers in Ethernet packets bearing a dedicated ether type. While the use of GRH is optional within InfiniBand subnets, it is mandatory when using RoCE. Applications written over IB verbs should work seamlessly, but they require provisioning of GRH information when creating address vectors. The library and driver are modified to provide mapping from GID to MAC addresses required by the hardware.

1.3   Supported Operating Systems/Distributions

- RedHat Enterprise Linux (RHEL)
- SuSe Linux Enterprise Server (SLES)
- OpenFabrics Enterprise Distribution (OFED)
- Microsoft Windows Server Family of Operating Systems
- VMware ESX

> For the list of the specific supported operating systems and distributions, please refer to the release notes for the applicable software downloads on the Dell support site: http://www.dell.com/support.

2   Adapter Card Interfaces

2.1   I/O Interfaces

Each adapter card includes the following interfaces:

- High speed port:
  - QSFP28 for 100GbE Network Adapters
  - QSFP+ for 40GbE Network Adapters
  - SFP28 for 25GbE Network Adapters
  - SFP+ for 10GbE Network Adapters
  - Backplane connection to the M1000e chassis for the 10GbE KR Blade Mezzanine Card
  - Two IMPEL connectors to Two PTMs or Switch Modules for the ConnectX-4 Lx 25GbE Mezzanine Card
- PCI Express (PCIe) x8 edge connector (Applies to ConnectX-3/ConnectX-3 Pro and ConnectX-4 Lx adapter cards)
- PCI Express (PCIe) x16 edge connector (Applies to ConnectX-4 adapter and / ConnectX-5 Ex cards)
- I/O panel LEDs *(does not apply to Mellanox ConnectX-3/ConnectX-3 Pro Dual Port 10GbE KR Blade Mezzanine Card and ConnectX®-4 LX Dual Port 25 GbE KR Mezzanine Card)*

2.1.1   Ethernet QSFP+/QSFP28/SFP+/SFP28 Interface

Note: This section does not apply to *Mellanox ConnectX-3/ConnectX-3 Pro Dual Port 10GbE KR Blade Mezzanine Card and ConnectX®-4 LX Dual Port 25 GbE KR Mezzanine Card.*

The network ports of ConnectX-3, ConnectX-3 Pro, ConnectX-4 ConnectX-4 Lx and ConnectX-5 Ex adapter cards are compliant with the IEEE 802.3 Ethernet standards. The QSFP+ and QSFP28 port has four Tx/Rx pairs of SerDes. The SFP+ and SFP28 ports have one Tx/Rx pair of SerDes. Ethernet traffic is transmitted through the cards' QSFP+, SFP+, SFP28 and QSFP28 connectors.

2.1.2 LED Assignments and Bracket Mechanical Drawings

There is a one bi-color link LED, green and yellow, and a green color activity LED located on the I/O panel. Link LED color is determined by link speed. The below tables detail the different LED functions per adapter card.

2.1.2.1 ConnectX-3/ConnectX-3 Pro 10GbE SFP+ Network Adapter Card

Note: This section does not apply to Mellanox *ConnectX-3/ConnectX-3 Pro Dual Port 10GbE KR Blade Mezzanine Card*.

Table 5 - LED Assignment for 10GbE SFP+ Network Adapters

Link LED (Bicolor - Green and Yellow)	Activity LED (Green)	Function
Off	Off	No link present
Yellow	Off	1 Gb/s link is present[a]
Green	Off	10 Gb/s link is present
Yellow	Blinking Green	Speed lower than the maximum is active
Green	Blinking Green	Maximum supported speed is active

a. 1 Gb/s Link Speed is only supported with 1 Gb/s optics. No 1 Gb/s optics are currently supported.

Mellanox ConnectX-3/ConnectX-3 Pro Dual Port 10GbE SFP+ Network Adapter Full Height Bracket
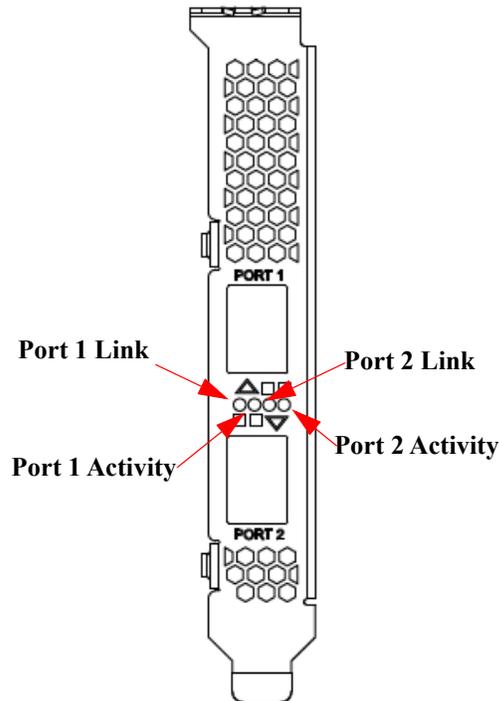
2.1.2.2 ConnectX-3/ConnectX-3 Pro 40GbE QSFP+ Network Adapter Card

Table 6 - LED Assignment for 40GbE QSFP+ Network Adapter

Link LED (Bicolor - Green and Yellow)	Activity LED (Green)	Function
Off	Off	No link present
Yellow	Off	10 Gb/s link is present[a]
Green	Off	40 Gb/s link is present
Yellow	Blinking Green	Speed lower than the maximum is active
Green	Blinking Green	Maximum supported speed is active

a. 10 Gb/s Link Speed is only supported with the Mellanox Quad to Serial Small Form Factor Pluggable Adapter (QSFP+ to SFP+ adapter or QSA).

Mellanox ConnectX-3/ConnectX-3 Pro Dual Port 40GbE QSFP+ Network Adapter Full Height Bracket

Port 1 Link

Port 2 Link

Port 1 Activity

Port 2 Activity
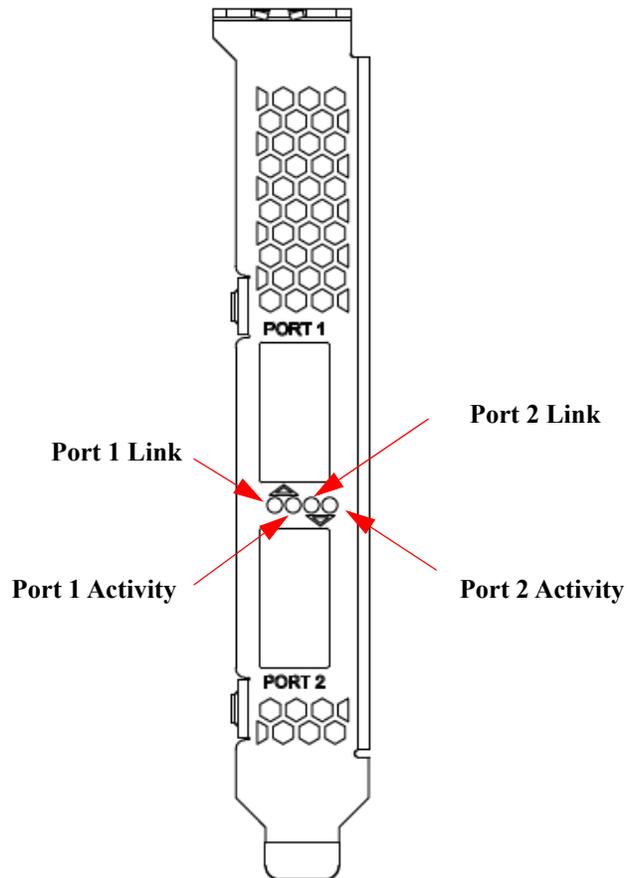
2.1.2.3 ConnectX-4 100GbE QSFP28 Network Adapter Card

Table 7 - LED Assignment for 100GbE QSFP28 Network Adapters

Link LED (Bicolor - Green and Yellow)	Activity LED (Green)	Function
Off	Off	No link present
Yellow	Off	40 Gb/s link is present
Green	Off	100 Gb/s link is present
Yellow	Blinking Green	40Gb/s speed is Active
Green	Blinking Green	100Gb/s speed is Active

Figure 1: Mellanox ConnectX-4 Dual Port QSFP28 Network Adapter Full Height Bracket

Port 1
Bi-color Link
and Activity LED

Port 2
Bi-color Link
and Activity LED
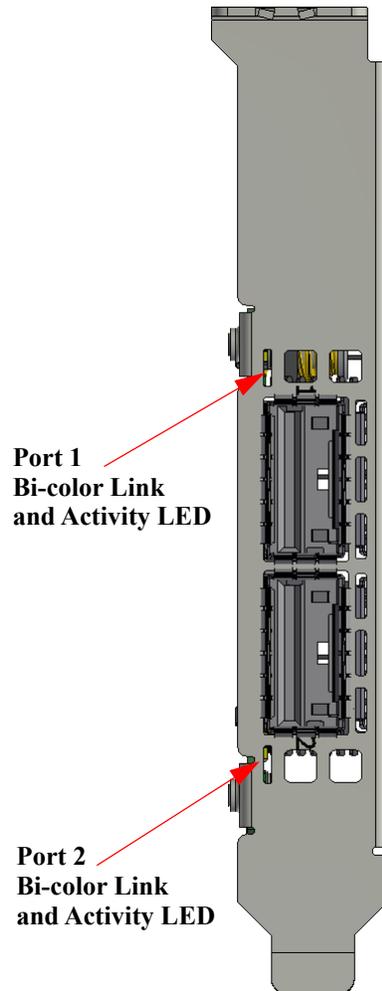
Figure 2: Mellanox ConnectX-4 Dual Port QSFP28 Network Adapter Low Profile Bracket

Port 1
Bi-Color Link
and Activity LED

Port 2
Bi-color Link
and Activity LED
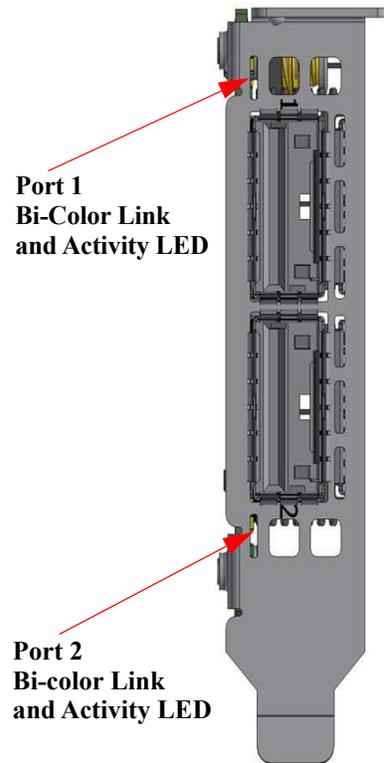
2.1.2.4 ConnectX-4 Lx 25GbE SFP28 Network Adapter Card

Table 8 - LED Assignment for 25GbE SFP28 Network Adapters

Link LED (Bicolor - Green and Yellow)	Activity LED (Green)	Function
Off	Off	No link present
Yellow	Off	10 Gb/s link is present
Green	Off	25 Gb/s link is present
Yellow	Blinking Green	Speed lower than the maximum is active
Green	Blinking Green	Maximum supported speed is active

Figure 3: Mellanox ConnectX-4 Lx Dual Port 25GbE SFP28 Network Adapter Full Height Bracket
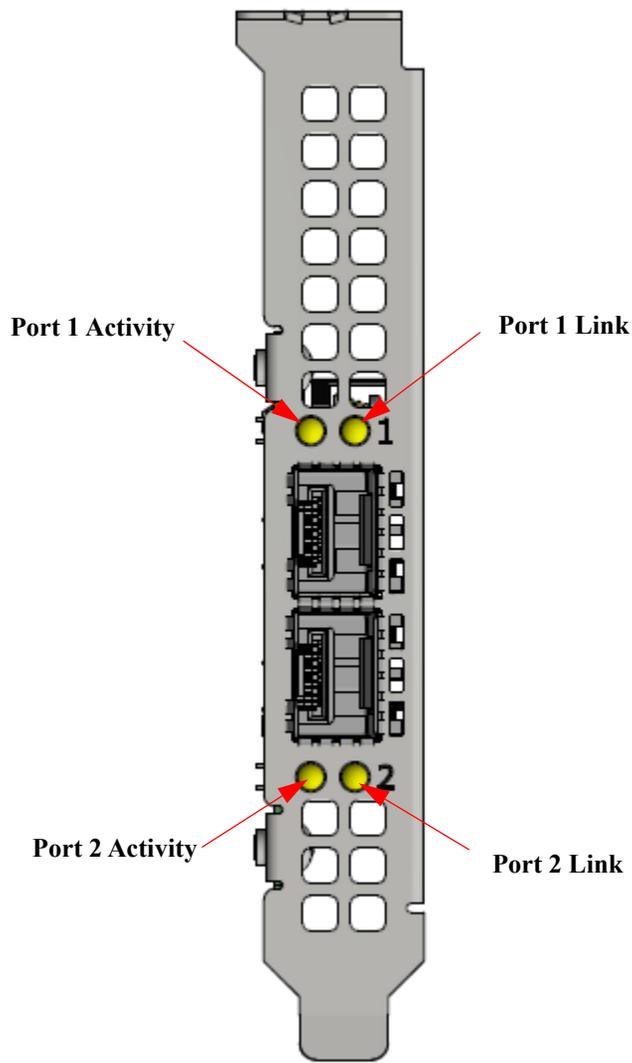
Figure 4: Mellanox ConnectX-4 Lx Dual Port 25GbE SFP28 Network Adapter Low Profile Bracket
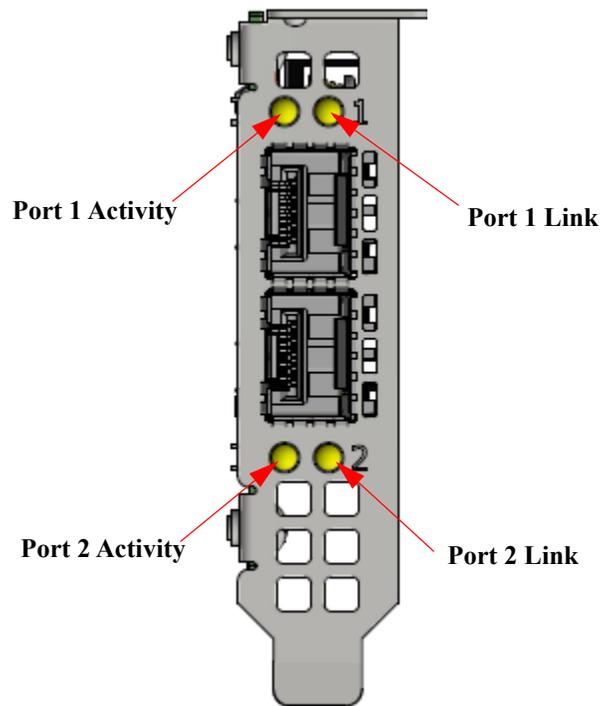
2.1.2.5 ConnectX-4 Lx 25GbE SFP28 for Dell Rack NDC Network Adapter Card

Table 9 - LED Assignment for 25GbE SFP28 for Dell Rack NDC Network Adapters

Link LED (Bicolor - Green and Yellow)	Activity LED (Green)	Function
Off	Off	No link present
Yellow	Off	10 Gb/s link is present
Green	Off	25 Gb/s link is present
Yellow	Blinking Green	Speed lower than the maximum is active
Green	Blinking Green	Maximum supported speed is active

Mellanox ConnectX-4 Lx Dual Port SFP28 25GbE for Dell rack NDC Faceplate
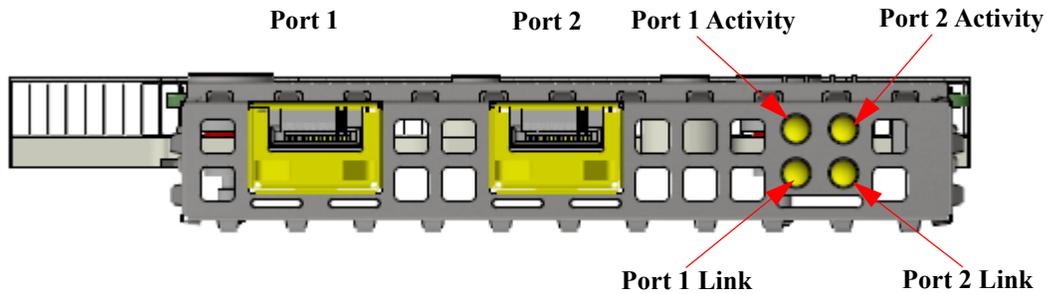
2.1.2.6 ConnectX-5 Ex Dual Port 100 GbE QSFP Network Adapter

Table 10 - LED Assignment for 100GbE QSFP28 Network Adapters

Link LED (Bicolor - Green and Yellow)	Activity LED (Green)	Function
Off	Off	No link present
Yellow	Off	40 Gb/s link is present
Green	Off	100 Gb/s link is present
Yellow	Blinking Green	40Gb/s speed is Active
Green	Blinking Green	100Gb/s speed is Active

Figure 5: **ConnectX-5 Ex Dual Port 100GbE QSFP28 Network Adapter Full Height Bracket**

Port 1 Bi-color Link and Activity LED

Port 2 Bi-color Link and Activity LED
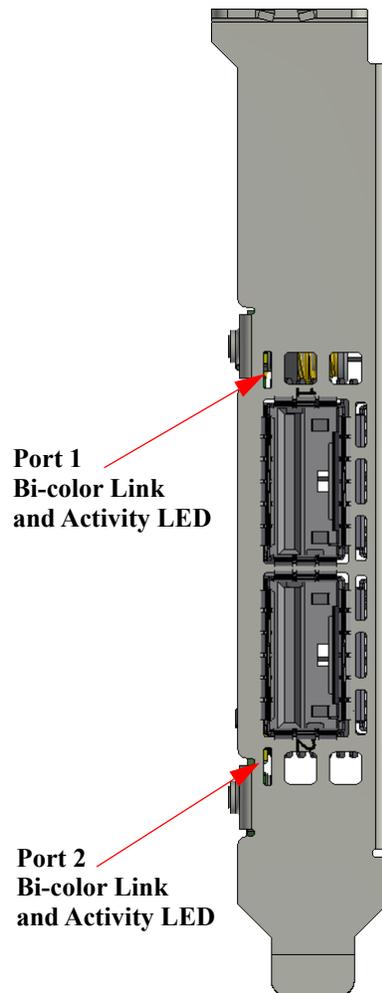
Figure 6: ConnectX-5 Ex Dual Port 100GbE QSFP28 Network Adapter Low Profile Bracket

Port 1
Bi-Color Link
and Activity LED

Port 2
Bi-color Link
and Activity LED
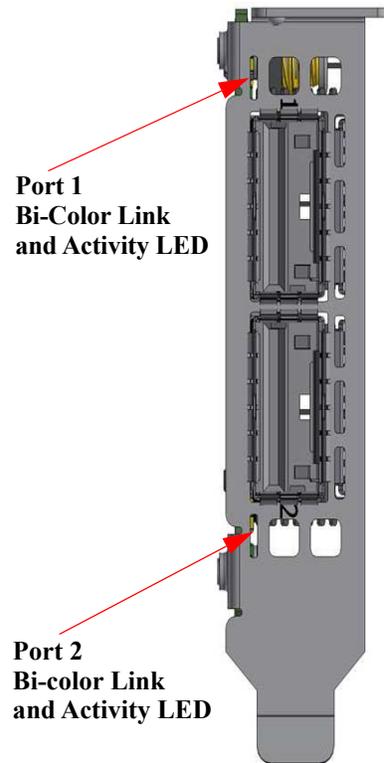
3   Installing the Hardware

3.1   System Requirements

3.1.1   Hardware

To install ConnectX-3/ConnectX-3 Pro ConnectX-4 Lx network adapter cards, a Dell EMC PowerEdge Server with an available PCI Express Gen 3.0 x8 slot is required.

To install ConnectX-4 and ConnectX-5 Ex Network adapter cards, a Dell EMC PowerEdge Server with an available PCI Express Gen 3.0 x16 slot is required.

> For the list of supported Dell EMC PowerEdge Servers please refer to the release notes for the applicable software and firmware downloads on the Dell support site: http://www.dell.com/support.

> For installation of Dell rNDC, please refer to Dell support site: http://www.dell.com/support.

3.1.2   Operating Systems/Distributions

Please refer to .

> For the list of the specific supported operating systems and distributions, please refer to the release notes for the applicable software downloads on the Dell support site: http://www.dell.com/support.

3.1.3   Software Stacks

Mellanox OpenFabric software package - MLNX_OFED for Linux and VMware, WinOF and WinOF-2 for Windows.

3.1.4   Co-requisites

For full functionality including manageability support, minimum versions of Server BIOS, Integrated Dell Remote Access Controller (iDRAC), and Dell Lifecycle Controller are required.

> For the list of co-requisites, please refer to the release notes for the applicable software and firmware downloads on the Dell support site: http://www.dell.com/support.

3.2 Safety Precautions

> ⚠️ The adapter is being installed in a system that operates with voltages that can be lethal. Before opening the case of the system, observe the following precautions to avoid injury and prevent damage to system components.

1. Remove any metallic objects from your hands and wrists.
2. Make sure to use only insulated tools.
3. Verify that the system is powered off and is unplugged.
4. It is required to use an ESD strap or other antistatic devices.

3.3 Pre-installation Checklist

1. Verify that your system meets the hardware and software requirements stated above.
2. Shut down your system if active.
3. After shutting down the system, turn off power and unplug the cord.
4. Remove the card from its package. Please note that the card must be placed on an antistatic surface.
5. Check the card for visible signs of damage. Do not attempt to install the card if damaged.

3.4 Installation Instructions

Please refer to the Dell EMC PowerEdge Server User Manual for your server system for instructions on installing add-in cards, Mezzanine cards, or Rack Network Daughter Cards into the server.

3.5 Connecting the Network Cables

3.5.1 Inserting a Cable into the Adapter Card

1. Support the weight of the cable before connecting it to the adapter card. Do this by using a cable holder or tying the cable to the rack.
2. Determine the correct orientation of the connector to the card before inserting the connector. Do not try and insert the connector upside down. This may damage the adapter card.
3. Insert the connector into the adapter card. Be careful to insert the connector straight into the cage. Do not apply any torque, up or down, to the connector cage in the adapter card.
4. Make sure that the connector locks in place.

3.5.2 Removing a Cable from the Adapter Card

1. Pull on the latch release mechanism thereby unlatching the connector and pull the connector out of the cage.
2. Do not apply torque to the connector when removing it from the adapter card.

3. Remove any cable supports that were used to support the cable's weight.

3.6 Identifying the Card in A System

3.6.1 On Linux

Get the device location on the PCI bus by running lspci and locating lines with the string "Mellanox Technologies":

```
> lspci |grep -i Mellanox
27:00.0 Network controller: Mellanox Technologies MT27500 Family [ConnectX-3]
27:00.0 Network controller: Mellanox Technologies MT27520 Family [ConnectX-3 Pro]
27:00.0 Network controller: Mellanox Technologies MT27700 Family [ConnectX-4]
27:00.0 Network controller: Mellanox Technologies MT27630 Family [ConnectX-4 Lx]
27:00.0 Network controller: Mellanox Technologies MT28800 Family [ConnectX-5 Ex]
```

4 Driver Installation and Configuration

4.1 Linux Driver

For Linux, download and install the latest Linux Drivers for Mellanox ConnectX-3, ConnectX-3 Pro, ConnectX-4 ConnectX-4 Lx and ConnectX-5 Ex Ethernet adapters software package available at Dell's support site http://www.dell.com/support. For driver installation instructions, please refer to Dell documentation at http://www.dell.com/support.

4.1.1 Installation Requirements

Required Disk Space for Installation
- 100 MB

Software Requirements
- Linux operating system

> For the list of supported operating system distributions, kernels and release notes for the applicable softwares, please refer to Dell's support site: http://www.dell.com/support.

Installer Privileges
- The installation requires administrator privileges on the target machine

4.1.2 Downloading Mellanox OFED

Step 1. Verify that the system has a Mellanox network adapter (NIC) installed by ensuring that you can see ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex in the display. The following example shows a system with an installed Mellanox NIC:

```
host1# lspci -v | grep Mellanox

27:00.0 Network controller: Mellanox Technologies MT27500 Family [ConnectX-3]
27:00.0 Network controller: Mellanox Technologies MT27520 Family [ConnectX-3 Pro]
27:00.0 Network controller: Mellanox Technologies MT27700 Family [ConnectX-4]
86:00.1 Network controller: Mellanox Technologies MT27630 Family [ConnectX-4 Lx]
86:00.1 Network controller: Mellanox Technologies MT28800 Family [ConnectX-5 Ex]
```

Step 2. Download the software release to your host.

The software release name has the format `MLNX_OFED_LINUX-<ver>.tar.gz`

Step 3. Use the md5sum utility to confirm the file integrity of your software release. Run the following command and compare the result to the value provided on the download page.

```
host1$ md5sum MLNX_OFED_LINUX-<ver>.tar.gz
```

4.1.3 Installing Mellanox OFED

The installation script, `install.sh`, performs the following:

* Discovers the currently installed kernel
* Uninstalls any software stacks that are part of the standard operating system distribution or another vendor's commercial stack
* Installs the MLNX_OFED_LINUX binary RPMs (if they are available for the current kernel)

4.1.3.1 Pre-installation Notes

The installation script removes all previously installed Mellanox OFED packages and installs the software release.

4.1.3.2 Installation Script

Within each distribution specific subdirectory there is an installation script called `install.sh`. Its usage is described below. You will use it during the installation procedure described in .

4.1.3.3 mlnxofedinstall Return Codes

The table below lists the `install.sh` script return codes and their meanings.

Table 11 - install.sh Return Codes

Return Code	Meaning
0	The Installation ended successfully
1	The installation failed
2	No firmware was found for the adapter device
22	Invalid parameter
28	Not enough free space
171	Not applicable to this system configuration. This can occur when the required hardware is not present on the system.
172	Prerequisites are not met. For example, missing the required software installed or the hardware is not configured correctly.
173	Failed to start the `mst` driver

4.1.4 Installation Procedure

Step 1. Login to the installation machine as root.

Step 2. Copy the software release on your machine

> For specific installation instructions, please refer to the applicable software download on the Dell support site http://www.dell.com/support.

Step 3. Un-tar the software release.

```
host1# tar -xvf MLNX_OFED_LINUX-<ver>.tar.gz
```

Step 4. Change directory to the distribution specific subdirectory.

```
host1# cd /MLNX_OFED_LINUX-<ver>/rhel6/rhel6.4
```

Step 5. Run the installation script (example).

```
../install.sh

Installing mlnx-ofa_kernel RPM
Preparing...                    ################################################
mlnx-ofa_kernel                 ################################################
Installing kmod-mlnx-ofa_kernel RPM
Preparing...                    ################################################
kmod-mlnx-ofa_kernel            ################################################
Installing mlnx-ofa_kernel-devel RPM
Preparing...                    ################################################
mlnx-ofa_kernel-devel           ################################################
Installing user level RPMs:
Preparing...                    ################################################
ofed-scripts                    ################################################
Preparing...                    ################################################
libibverbs                      ################################################
Preparing...                    ################################################
libibverbs-devel                ################################################
Preparing...                    ################################################
libibverbs-devel-static         ################################################
Preparing...                    ################################################
libibverbs-utils                ################################################
Preparing...                    ################################################
libmlx4                         ################################################
Preparing...                    ################################################
libmlx4-devel                   ################################################
Preparing...                    ################################################
libibumad                       ################################################
Preparing...                    ################################################
libibumad-devel                 ################################################
Preparing...                    ################################################
```

```
libibumad-static           #################################################
Preparing...               #################################################
libibmad                   #################################################
Preparing...               #################################################
libibmad-devel             #################################################
Preparing...               #################################################
libibmad-static            #################################################
Preparing...               #################################################
librdmacm                  #################################################

Preparing...               #################################################
librdmacm-utils            #################################################
Preparing...               #################################################
librdmacm-devel            #################################################
Preparing...               #################################################
perftest                   #################################################
Device (02:00.0):
                02:00.0 Ethernet controller: Mellanox Technologies MT27500 Family
[ConnectX-3]

                Link Width: 8x
                PCI Link Speed: Unknown


Installation finished successfully.
```

Step 6. The script adds the following lines to `/etc/security/limits.conf` for the user-space components such as MPI:

```
* soft memlock unlimited
* hard memlock unlimited
```

These settings unlimit the amount of memory that can be pinned by a user space application. If desired, tune the value unlimited to a specific amount of RAM.

4.1.5   Installation Results

- The OFED package is installed under the /usr directory.

- The kernel modules are installed under:

    - mlx4 driver:

```
/lib/modules/<kernel_version>/extra/mlnx-ofa_kernel/drivers/net/ethernet/mellanox/
mlx4/
```

    - RDS:

```
/lib/modules/`uname -r`/updates/kernel/net/rds/rds.ko
/lib/modules/`uname -r`/updates/kernel/net/rds/rds_rdma.ko
/lib/modules/`uname -r`/updates/kernel/net/rds/rds_tcp.ko
```

> Kernel's modules location may vary depending on the kernel's configuration.
> For example: /lib/modules/`uname -r`/extra/kernel/drivers/net/ethernet/mellanox/mlx4/mlx4_core

- The script `openibd` is installed under `/etc/init.d/`. This script can be used to load and unload the software stack.
  - `/etc/sysconfig/network/` on a SuSE machine
- The installation process unlimits the amount of memory that can be pinned by a user space application. See Step 6.
- Man pages will be installed under `/usr/share/man/`

4.1.6 Post-installation Notes

Most of the Mellanox OFED components can be configured or reconfigured after the installation by modifying the relevant configuration files.

4.1.7 Uninstalling Mellanox OFED

Either use the distribution specific uninstall.sh script or use the script `/usr/sbin/ofed_un-install.sh` to uninstall the Mellanox OFED package. The ofed_uninstall.sh is part of the `ofed-scripts` RPM.

4.1.8 UEFI Secure Boot

All kernel modules included in MLNX_OFED for RHEL7 and SLES12 are signed with x.509 key to support loading the modules when Secure Boot is enabled.

4.1.8.1 Enrolling Mellanox's x.509 Public Key On your Systems

In order to support loading MLNX_OFED drivers when an OS supporting Secure Boot boots on a UEFI-based system with Secure Boot enabled, the Mellanox x.509 public key should be added to the UEFI Secure Boot key database and loaded onto the system key ring by the kernel.

Follow these steps below to add the Mellanox's x.509 public key to your system:

> Prior to adding the Mellanox's x.509 public key to your system, please make sure:
> - the 'mokutil' package is installed on your system
> - the system is booted in UEFI mode

Step 1. Download the x.509 public key.

```
# wget http://www.mellanox.com/downloads/ofed/mlnx_signing_key_pub.der
```

Step 2. Add the public key to the MOK list using the mokutil utility.

You will be asked to enter and confirm a password for this MOK enrollment request.

```
# mokutil --import mlnx_signing_key_pub.der
```

Step 3. Reboot the system.

The pending MOK key enrollment request will be noticed by `shim.efi` and it will launch `Mok-Manager.efi` to allow you to complete the enrollment from the UEFI console. You will need to enter the password you previously associated with this request and confirm the enrollment. Once done, the public key is added to the MOK list, which is persistent. Once a key is in the MOK list, it will be automatically propagated to the system key ring and subsequently will be booted when the UEFI Secure Boot is enabled.

> To see what keys have been added to the system key ring on the current boot, install the 'keyutils' package and run: `#keyctl list %:.system_keyring`

4.1.8.2 Removing Signature from Kernel Modules

The signature can be removed from a signed kernel module using the 'strip' utility which is provided by the 'binutils' package.

```
# strip -g my_module.ko
```

The strip utility will change the given file without saving a backup. The operation can be undone only by resigning the kernel module. Hence, we recommend backing up a copy prior to removing the signature.

➢ *To remove the signature from the MLNX_OFED kernel modules:*

Step 1. Remove the signature.

```
# rpm -qa | grep -E "kernel-ib|mlnx-ofa_kernel|iser|srp|knem" | xargs rpm -ql | grep "\.ko$" | xargs strip -g
```

After the signature has been removed, a message as the below will no longer be presented upon module loading:

```
"Request for unknown module key 'Mellanox Technologies signing key:
61feb074fc7292f958419386ffdd9d5ca999e403' err -11"
```

However, please note that a similar message as the following will still be presented:

```
"my_module: module verification failed: signature and/or required key missing - taint-
ing kernel"
```

This message is presented only once for each boot of the first module which has no signature or whose key is not in the kernel key ring. Therefore, it is easy to miss this message. You will not be able to see it on repeated tests where you unload and reload a kernel module, unless you reboot. It is not possible to eliminate this message.

Step 2. Update the initramfs on RHEL systems with the stripped modules.

```
mkinitrd /boot/initramfs-$(uname -r).img $(uname -r) --force
```

4.2 Linux Driver Features

4.2.1 iSCSI Extensions for RDMA (iSER)

iSCSI Extensions for RDMA (iSER) extends the iSCSI protocol to RDMA. It permits the transfer of data into and out of SCSI buffers without intermediate data copies.

iSER uses the RDMA protocol suite to supply higher bandwidth for block storage transfers (zero time copy behavior). To that fact, it eliminates the TCP/IP processing overhead while preserving compatibility with iSCSI protocol.
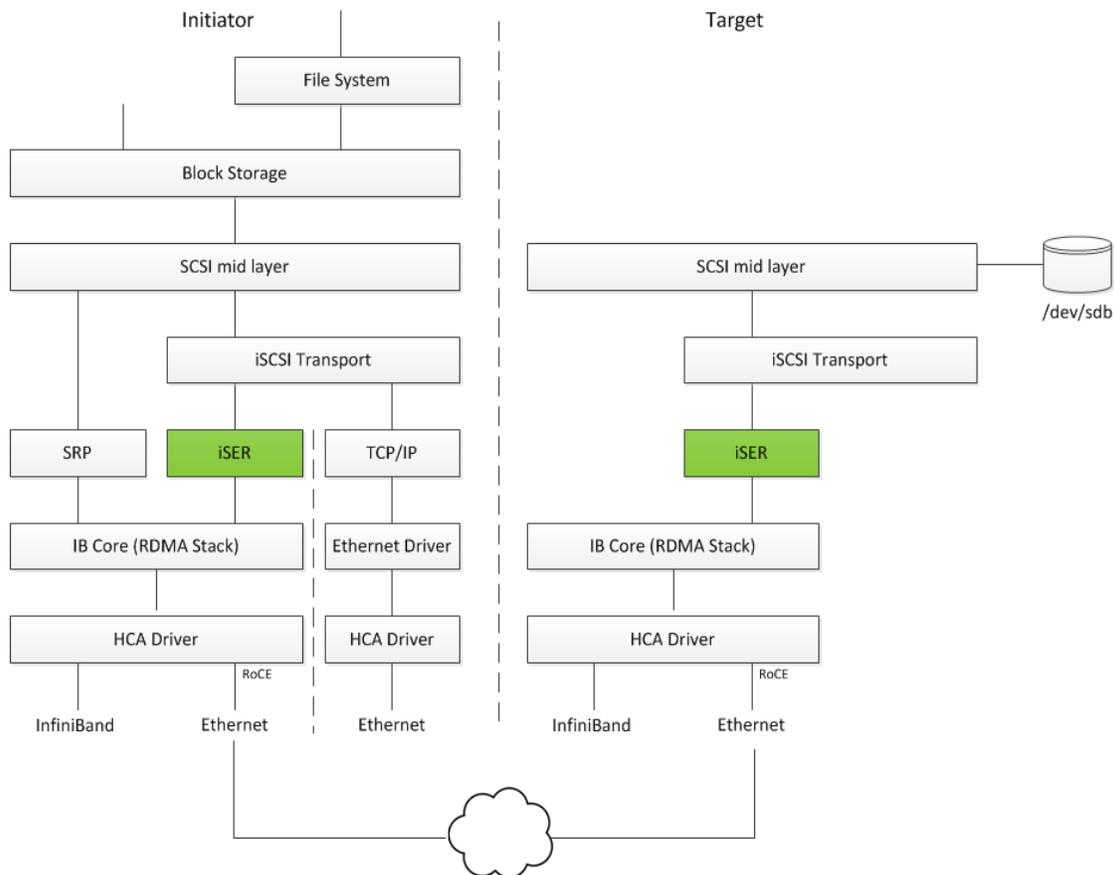
There are three target implementations of iSER:

- Linux SCSI target framework (tgt)

- Linux-IO target (LIO)

- Generic SCSI target subsystem for Linux (SCST)

Each one of those targets can work in TCP or iSER transport modes.

iSER also supports RoCE without any additional required configuration. To bond the RoCE interfaces, set the `fail_over_mac` option in the bonding driver.

RDMA/RoCE is located below the iSER block on the network stack. In order to run iSER, the RDMA layer should be configured and validated (over Ethernet). For troubleshooting RDMA, please refer to "How To Enable, Verify and Troubleshoot RDMA" on Mellanox Community (https://community.mellanox.com).

4.2.2 Enabling/Disabling RoCE on VFs (ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex)

By default, when configuring several VFs on the hypervisor, all VFs will be enabled with RoCE. This means that they require more OS memory comparing to Ethernet only VFs. In case you are only interested in Ethernet (no RDMA) on the VF, and you wish to save the hypervisor memory, you can disable RoCE on the VF from the hypervisor. By doing this, the VF will request less host memory from hypervisor.

For details on how to enable/disable RoCE on a VF, refer to HowTo Enable/Disable RoCE on VFs Community post.

4.2.2.1 RoCE LAG (ConnectX-3/ConnectX-3 Pro)

RoCE Link Aggregation (RoCE LAG) provides failover and link aggregation capabilities for mlx4 device physical ports. In this mode, only one IB port, that represents the two physical ports, is exposed to the application layer. Kernel 4.0 is a requirement for this feature to properly function.

4.2.2.1.1 Enabling RoCE Link Aggregation

To enter the Link Aggregation mode, a bonding master that enslaves the two net devices on the mlx4 ports is required. Then, the mlx4 device re-registers itself in the IB stack with a single port. If the requirement is not met, the device re-registers itself again with two ports.

For the device to enter the Link Aggregation mode, the following prerequisites must exist:

- Exactly 2 slaves must be under the bonding master
- The bonding master has to be in one of the following modes:
  - (1) active-backup mode
  - (2) static active-active mode
  - (4) dynamic active-active mode

Restarting the device, when entering or leaving Link Aggregation mode, invalidates the open resources (QPs, MRs, etc.) on the device.

4.2.2.1.1.1 Link Aggregation in active-backup Mode

When the bonding master works in active-backup mode, RoCE packets are transmitted and received from the active port that the bonding master reports. The logic of fail over is done solely in the bonding driver and the mlx4 driver only polls it.

4.2.2.1.1.2 Link Aggregation in active-active Mode

In this mode, RoCE packets are transmitted and received from both physical ports. While the mlx4 driver has no influence on the port on which packets are being received from, it can determine the port packets are transmitted to.

If user application does not set a preference, the mlx4 driver chooses a port in a round robin fashion when QP is modified from RESET to INIT. This is necessary because application sees only one port to use so it will always state `port_num` 1 in the QP attributes. With that, the theoretical bandwidth of the system will be kept as the sum of the two ports.

Application that prefers to send packets on the specific port for a specific QP, should set `flow_entropy` when modifying a QP from RESET to INIT. Values for the `flow_entropy` parameter are interpreted by the mlx4 driver as a hint to associate the SQ of the QP to "1" while odd values associate the SQ with port 2.

The code example below shows how to set `flow_entropy` for a QP.

```
struct ibv_exp_qp_attr attr = {
                                .comp_mask              = IBV_EXP_QP_ATTR_FLOW_ENTROPY,
                                .qp_state               = IBV_QPS_INIT,
                                .pkey_index             = 0,
                                .port_num               = port,
                                .qp_access_flags        = 0,
                                .flow_entropy           = 1
        };
          if (ibv_exp_modify_qp(ctx->qp, &attr,
                                 IBV_QP_STATE                |
                                IBV_QP_PKEY_INDEX           |
                                IBV_QP_PORT                 |
                                IBV_EXP_QP_FLOW_ENTROPY     |
                                IBV_QP_ACCESS_FLAGS)) {
                                fprintf(stderr, "Failed to modify QP to INIT\n"); goto
                                clean_qp;
        }
```

4.2.2.1.2 Link Aggregation for Virtual Functions

When ConnectX®-3 Virtual Functions are present, High Availability behaves differently. Nonetheless, its configuration process remain the same and is performed in the Hypervisor. However, since the mlx4 device in the Hypervisor does not re-register, the two ports remain exposed to the upper layer. Therefore, entering the LAG mode does not invalidate the open resources although applications that run in the Hypervisor are still protected from a port failure.

When Virtual Functions are present and RoCE Link Aggregation is configured in the Hypervisor, a VM with an attached ConnectX-3 Virtual Function is protected from a Virtual Function port failure. For example, if the Virtual Function is bounded to `port #1` and this port fails, the Virtual Function will be redirected to `port #2`. Once `port #1` comes up, the Virtual Function is redirected back to `port #1`.

When the Hypervisor enters the LAG mode, it checks for the requirements below. If they are met, the Hypervisor enables High Availability also for the Virtual Functions.

The requirements are:

- Only single port VFs are configured, on either port (See Section 4.2.11.2.1, "Configuring SR-IOV for ConnectX-3/ConnectX-3 Pro", on page 78)

- Flow steering is enabled

- Total number of VFs is smaller than 64

4.2.2.2 RoCE LAG (ConnectX-4/ConnectX-4 Lx/ConnectX-5 Ex)

RoCE LAG is a feature meant for mimicking Ethernet bonding for IB devices, and is available for dual port cards only.

RoCE LAG mode is entered when both Ethernet interfaces are configured as a bond in one of the following modes:

- active-backup (mode 1)

- balance-xor (mode 2)

- 802.3ad (LACP) (mode 4)

 Any change of bonding configuration that negates one of the above rules (i.e, bonding mode is

not 1, 2 or 4, or both Ethernet interfaces that belong to the same card are not the only slaves

of the bond interface), will result in exiting RoCE LAG mode, and the return to normal IB device

per port configuration.

For further information on RoCE LAG for ConnectX-4 and ConnectX-4 Lx, refer to HowTo Test RoCE over LAG (ConnectX-4) Community post.

4.2.3 iSER Initiator

The iSER initiator is controlled through the iSCSI interface available at the iscsi-initiator-utils package.

To discover and log into iSCSI targets, as well as access and manage the open-iscsi database, use the `iscasiadm` utility, a command-line tool.

To enable iSER as a transport protocol use "`-I iser`" as a parameter of the `iscasiadm` command.

Example for discovering and connecting targets over iSER:

```
iscsiadm -m discovery -o new -o old -t st -I iser -p <ip:port> -l
```

Note that the target implementation (e.g. LIO, SCST, TGT) does not affect the initiation process and configuration.

4.2.3.1 iSER Targets

Setting the iSER target is out of scope of this manual. For guidelines on how to do so, please refer to the relevant target documentation (e.g. stgt, clitarget).

Target settings such as `timeouts` and `retries` are set the same as any other iSCSI targets.

If targets are set to auto connect on boot, and targets are unreachable, it may take a long time to continue the boot process if `timeouts` and `max retries` are set too high.

For various configuration, troubleshooting and debugging examples, please refer to Storage Solutions on Mellanox Community (https://community.mellanox.com).

4.2.4 Quality of Service (QoS) Ethernet

4.2.4.1 Mapping Traffic to Traffic Classes

Mapping traffic to TCs consists of several actions which are user controllable, some controlled by the application itself and others by the system/network administrators.

The following is the general mapping traffic to Traffic Classes flow:

1. The application sets the required Type of Service (ToS).

2. The ToS is translated into a Socket Priority (`sk_prio`).

3. The `sk_prio` is mapped to a User Priority (UP) by the system administrator (some applications set `sk_prio` directly).

4. The UP is mapped to TC by the network/system administrator.

5. TCs hold the actual QoS parameters.

QoS can be applied on the following types of traffic. However, the general QoS flow may vary:

- **Plain Ethernet** - Applications use regular inet sockets and the traffic passes via the kernel Ethernet driver

- **RoCE** - Applications use the RDMA API to transmit using QPs

- **Raw Ethernet QP** - Applications use the VERB API to transmit using a Raw Ethernet QP

4.2.4.2 Plain Ethernet Quality of Service Mapping

Applications use regular inet sockets and the traffic passes via the kernel Ethernet driver.

The following is the Plain Ethernet QoS mapping flow:

1. The application sets the ToS of the socket using `setsockopt` (`IP_TOS`, value).

2. ToS is translated into the `sk_prio` using a fixed translation:

```
TOS 0 <=> sk_prio 0
TOS 8 <=> sk_prio 2
TOS 24 <=> sk_prio 4
TOS 16 <=> sk_prio 6
```

3. The Socket Priority is mapped to the UP:

   • If the underlying device is a VLAN device, `egress_map` is user controlled by the `vconfig` command. This is per VLAN mapping.

   • If the underlying device is not a VLAN device, the `tc` command is used. In this case, even though `tc` manual states that the mapping is from the `sk_prio` to the TC number, the `mlx-4_en` driver interprets this as a `sk_prio` to UP mapping.

   Mapping the sk_prio to the UP is done by using `tc_wrap.py -i <dev name> -u 0,1,2,3,4,5,6,7`

4. The UP is mapped to the TC as configured by the `mlnx_qos` tool or by the `lldpad` daemon if DCBX is used.

> Socket applications can use `setsockopt` (`SK_PRIO`, value) to directly set the `sk_prio` of the socket. In this case, the ToS to `sk_prio` fixed mapping is not needed. This allows the application and the administrator to utilize more than the 4 possible values via ToS.

> In the case of VLAN interface, the UP obtained according to the above mapping is also used in the VLAN tag of the traffic

4.2.4.3 RoCE Quality of Service Mapping

Applications use RDMA-CM API to create and use QPs.

The following is RoCE QoS mapping flow:

1. The application sets the ToS of the QP using the `rdma_set_option` option (`RDMA_OP-TION_ID_TOS`, value).

2. ToS is translated into the Socket Priority (`sk_prio`) using a fixed translation:

```
TOS 0 <=> sk_prio 0
TOS 8 <=> sk_prio 2
TOS 24 <=> sk_prio 4
TOS 16 <=> sk_prio 6
```

3. The Socket Priority is mapped to the User Priority (UP) using the `tc` command.

   In the case of a VLAN device, the parent real device is used for the purpose of this mapping.

4. The UP is mapped to the TC as configured by the `mlnx_qos` tool or by the `lldpad` daemon if DCBX is used.

> With RoCE, there can only be 4 predefined ToS values for the purpose of QoS mapping.

4.2.4.4 Raw Ethernet QP Quality of Service Mapping

Applications directly open a Raw Ethernet QP using VERBs.

The following is the Raw Ethernet QoS mapping flow:

1. The application sets the UP of the Raw Ethernet QP during the INIT to RTR state transition of the QP:

    • Sets `qp_attrs.ah_attrs.sl = up`

    • Calls `modify_qp` with `IB_QP_AV` set in the mask

2. The UP is mapped to the TC as configured by the `mlnx_qos` tool or by the `lldpad` daemon if DCBX is used.

> When using Raw Ethernet QP mapping, the TOS/sk_prio to UP mapping is lost.

> Performing the Raw Ethernet QP mapping forces the QP to transmit using the given UP.
> If packets with VLAN tag are transmitted, UP in the VLAN tag will be overwritten with the given UP.

4.2.4.5 Map Priorities with tc_wrap.py/mlnx_qos

Network flow that can be managed by QoS attributes is described by a User Priority (UP). A user's `sk_prio` is mapped to UP which in turn is mapped into TC.

• Indicating the UP

  • When the user uses `sk_prio`, it is mapped into a UP by the 'tc' tool. This is done by the `tc_wrap.py` tool which gets a list of <= 16 comma separated UP and maps the `sk_prio` to the specified UP.
  For example, `tc_wrap.py -ieth0 -u 1,5` maps `sk_prio 0` of `eth0` device to UP 1 and `sk_prio 1` to UP 5.

  • Setting `set_egress_map` in VLAN, maps the `skb_priority` of the VLAN to a `vlan_qos`. The `vlan_qos` represents a UP for the VLAN device.

  • In RoCE, `rdma_set_option` with `RDMA_OPTION_ID_TOS` could be used to set the UP

  • When creating QPs, the `sl` field in `ibv_modify_qp` command represents the UP

- Indicating the TC

  - After mapping the `skb_priority` to UP, one should map the UP into a TC. This assigns the user priority to a specific hardware traffic class. In order to do that, `mlnx_qos` should be used. `mlnx_qos` gets a list of mappings between UPs to TCs. For example, `mlnx_qos -ieth0 -p 0,0,0,0,1,1,1,1` maps UPs 0-3 to `TC0`, and Ups 4-7 to `TC1`.

4.2.4.6 Quality of Service Properties

The different QoS properties that can be assigned to a TC are:

- Strict Priority (see "Strict Priority")

- Enhanced Transmission Selection (ETS) (see "Enhanced Transmission Selection (ETS)")

- Rate Limit (see "Rate Limit")

4.2.4.6.1 Strict Priority

When setting a TC's transmission algorithm to be 'strict', then this TC has absolute (strict) priority over other TC strict priorities coming before it (as determined by the TC number: TC 7 is highest priority, TC 0 is lowest). It also has an absolute priority over non strict TCs (ETS).

This property needs to be used with care, as it may easily cause starvation of other TCs.

A higher strict priority TC is always given the first chance to transmit. Only if the highest strict priority TC has nothing more to transmit, will the next highest TC be considered.

Non strict priority TCs will be considered last to transmit.

This property is extremely useful for low latency and low bandwidth traffic that needs to get immediate service when it exists, but is not of high volume to starve other transmitters in the system.

4.2.4.6.2 Enhanced Transmission Selection (ETS)

Enhanced Transmission Selection standard (ETS) exploits the time periods in which the offered load of a particular Traffic Class (TC) is less than its minimum allocated bandwidth by allowing the difference to be available to other traffic classes.

After servicing the strict priority TCs, the amount of bandwidth (BW) left on the wire may be split among other TCs according to a minimal guarantee policy.

If, for instance, TC0 is set to 80% guarantee and TC1 is set to 20% (the TCs sum must be 100), then the BW left after servicing all strict priority TCs will be split according to this ratio.

Since this is a minimal guarantee, there is no maximum enforcement. This means, in the same example, if TC1 did not use its share of 20%, the remainder will be used by TC0.

ETS is configured using the mlnx_qos tool ("mlnx_qos") which allows you to:

- Assign a transmission algorithm to each TC (strict or ETS)

- Set minimal BW guarantee to ETS TCs

Usage:

```
mlnx_qos -i [options]
```

4.2.4.6.3 Rate Limit

Rate limit defines a maximum bandwidth allowed for a TC. Please note that 10% deviation from the requested values is considered acceptable.

4.2.4.7 Quality of Service Tools

4.2.4.7.1 mlnx_qos

`mlnx_qos` is a centralized tool used to configure QoS features of the local host. It communicates directly with the driver, therefore not requiring setting up a DCBX daemon on the system.

The `mlnx_qos` tool enables the system administrator to:

*   Inspect the current QoS mappings and configuration

    The tool will also display maps configured by TC and `vconfig set_egress_map` tools, in order to give a centralized view of all QoS mappings.

*   Set UP to TC mapping

*   Assign a transmission algorithm to each TC (strict or ETS)

*   Set minimal BW guarantee to ETS TCs

*   Set rate limit to TCs

For unlimited ratelimit set the ratelimit to 0.

Usage:

```
mlnx_qos -i <interface> [options]
```

Options:

```
--version           show program's version number and exit
-h, --help          show this help message and exit
-p LIST, --prio_tc=LIST
                    maps UPs to TCs. LIST is 8 comma seperated TC numbers.
                    Example: 0,0,0,0,1,1,1,1 maps UPs 0-3 to TC0, and UPs
                    4-7 to TC1
-s LIST, --tsa=LIST  Transmission algorithm for each TC. LIST is comma
                    seperated algorithm names for each TC. Possible
                    algorithms: strict, etc. Example: ets,strict,ets sets
                    TC0,TC2 to ETS and TC1 to strict. The rest are
                    unchanged.
-t LIST, --tcbw=LIST  Set minimal guaranteed %BW for ETS TCs. LIST is comma
                    seperated percents for each TC. Values set to TCs that
                    are not configured to ETS algorithm are ignored, but
                    must be present. Example: if TC0,TC2 are set to ETS,
                    then 10,0,90 will set TC0 to 10% and TC2 to 90%.
                    Percents must sum to 100.
-r LIST, --ratelimit=LIST
                    Rate limit for TCs (in Gbps). LIST is a comma
                    seperated Gbps limit for each TC. Example: 1,8,8 will
                    limit TC0 to 1Gbps, and TC1,TC2 to 8 Gbps each.
-i INTF, --interface=INTF
                    Interface name
-a                  Show all interface's TCs
```

Get Current Configuration:

```
tc: 0 ratelimit: unlimited, tsa: strict
        up:   0
                skprio: 0
                skprio: 1
                skprio: 2 (tos: 8)
                skprio: 3
                skprio: 4 (tos: 24)
                skprio: 5
                skprio: 6 (tos: 16)
                skprio: 7
                skprio: 8
                skprio: 9
                skprio: 10
                skprio: 11
                skprio: 12
                skprio: 13
                skprio: 14
                skprio: 15
        up:   1
        up:   2
        up:   3
        up:   4
        up:   5
        up:   6
        up:   7
```

Set ratelimit- 3Gbps for tc0, 4Gbps for tc1 and 2Gbps for tc2:

```
tc: 0 ratelimit: 3 Gbps, tsa: strict
        up:  0
                skprio: 0
                skprio: 1
                skprio: 2 (tos: 8)
                skprio: 3
                skprio: 4 (tos: 24)
                skprio: 5
                skprio: 6 (tos: 16)
                skprio: 7
                skprio: 8
                skprio: 9
                skprio: 10
                skprio: 11
                skprio: 12
                skprio: 13
                skprio: 14
                skprio: 15
        up:  1
        up:  2
        up:  3
        up:  4
        up:  5
        up:  6
        up:  7
```

Configure QoS. Map UP 0,7 to tc0, 1,2,3 to tc1 and 4,5,6 to tc 2. Set tc0,tc1 as ets and tc2 as strict. Divide ets 30% for tc0 and 70% for tc1:

```
mlnx_qos -i eth3 -s ets,ets,strict -p 0,1,1,1,2,2,2 -t 30,70
tc: 0 ratelimit: 3 Gbps, tsa: ets, bw: 30%
        up:  0
                skprio: 0
                skprio: 1
                skprio: 2 (tos: 8)
                skprio: 3
                skprio: 4 (tos: 24)
                skprio: 5
                skprio: 6 (tos: 16)
                skprio: 7
                skprio: 8
                skprio: 9
                skprio: 10
                skprio: 11
                skprio: 12
                skprio: 13
                skprio: 14
                skprio: 15
    up:  7
tc: 1 ratelimit: 4 Gbps, tsa: ets, bw: 70%
```

```
        up:  1
        up:  2
        up:  3
tc: 2 ratelimit: 2 Gbps, tsa: strict
        up:  4
        up:  5
        up:  6
```

4.2.4.7.2 tc and tc_wrap.py

The 'tc' tool is used to setup `sk_prio` to UP mapping, using the `mqprio` queue discipline.

In kernels that do not support `mqprio` (such as 2.6.34), an alternate mapping is created in `sysfs` when using ConnectX®-3/ConnectX®-3 Pro adapter cards. The 'tc_wrap.py' tool will use either the `sysfs` or the 'tc' tool to configure the `sk_prio` to UP mapping. In ConnectX®-4 Lx adapter cards, the `'tc_wrap.py'` tool is used only in kernels that support `mqprio`.

Usage:

```
tc_wrap.py -i <interface> [options]
```

Options:

```
--version                             show program's version number and exit
-h, --help                            show this help message and exit
-u SKPRIO_UP, --skprio_up=SKPRIO_UP   maps sk_prio to UP. LIST is <=16 comma separated
                                      UP. index of element is sk_prio.
-i INTF, --interface=INTF             Interface name
```

Set skprio 0-2 to UP0, and skprio 3-7 to UP1 on the Ethernet Interface

```
h-qa-le014:~ # tc_wrap.py -i eth4 -u 0,0,0,1,1,1,1,1
UP  0
        skprio: 0
        skprio: 1
        skprio: 2 (tos: 8)
        skprio: 7
        skprio: 8
        skprio: 9
        skprio: 10
        skprio: 11
        skprio: 12
        skprio: 13
        skprio: 14
        skprio: 15
UP  1
        skprio: 3
        skprio: 4 (tos: 24)
        skprio: 5
        skprio: 6 (tos: 16)
        skprio: 7
```

```
UP  2
UP  3
UP  4
UP  5
UP  6
UP  7
```

Additional Tools

tc tool compiled with the `sch_mqprio` module is required to support kernel v2.6.32 or higher. This is a part of `iproute2` package v2.6.32-19 or higher. Otherwise, an alternative custom sysfs interface is available.

- `mlnx_qos tool` (package: ofed-scripts) requires python >= 2.5
- `tc_wrap.py` (package: ofed-scripts) requires python >= 2.5

4.2.5   Ethernet Timestamping

Timestamping is the process of keeping track of a packet creation. A timestamping service supports assertions of proof that a datum existed before a particular time. Incoming packets are time-stamped before they are distributed on the PCI depending on the congestion in the PCI buffers. Outgoing packets are time-stamped very close to placing them on the wire.

4.2.5.1   Enabling Timestamping

Timestamping is off by default and should be enabled before use.

➢ *To enable timestamping for a socket:*

- Call setsockopt() with SO_TIMESTAMPING and the following flags:

```
SOF_TIMESTAMPING_TX_HARDWARE:  try to obtain send timestamp in hardware
SOF_TIMESTAMPING_TX_SOFTWARE:  if SOF_TIMESTAMPING_TX_HARDWARE is off or
                               fails, then do it in software
SOF_TIMESTAMPING_RX_HARDWARE:  return the original, unmodified timestamp
                               as generated by the hardware
SOF_TIMESTAMPING_RX_SOFTWARE:  if SOF_TIMESTAMPING_RX_HARDWARE is off or
                               fails, then do it in software
SOF_TIMESTAMPING_RAW_HARDWARE: return original raw hardware timestamp
SOF_TIMESTAMPING_SYS_HARDWARE: return hardware timestamp transformed to
                               the system time base
SOF_TIMESTAMPING_SOFTWARE:     return system timestamp generated in
                               software

SOF_TIMESTAMPING_TX/RX determine how timestamps are generated.
SOF_TIMESTAMPING_RAW/SYS determine how they are reported
```

➢ *To enable timestamping for a net device:*

Admin privileged user can enable/disable timestamping through calling ioctl (sock, SIOCSH-WTSTAMP, &ifreq) with the following values:

Send side time sampling:

- Enabled by ifreq.hwtstamp_config.tx_type when:

```
/* possible values for hwtstamp_config->tx_type */
enum hwtstamp_tx_types {
        /*
         * No outgoing packet will need hardware timestamping;
         * should a packet arrive which asks for it, no hardware
         * timestamping will be done.
         */
        HWTSTAMP_TX_OFF,

        /*
         * Enables hardware timestamping for outgoing packets;
         * the sender of the packet decides which are to be
         * timestamped by setting %SOF_TIMESTAMPING_TX_SOFTWARE
         * before sending the packet.
         */
        HWTSTAMP_TX_ON,
   /*
         * Enables timestamping for outgoing packets just as
         * HWTSTAMP_TX_ON does, but also enables timestamp insertion
         * directly into Sync packets. In this case, transmitted Sync
         * packets will not received a timestamp via the socket error
         * queue.
         */
        HWTSTAMP_TX_ONESTEP_SYNC,
};
Note: for send side timestamping currently only HWTSTAMP_TX_OFF and
HWTSTAMP_TX_ON are supported.
```

Receive side time sampling:

- Enabled by ifreq.hwtstamp_config.rx_filter when:

```
/* possible values for hwtstamp_config->rx_filter */
enum hwtstamp_rx_filters {
        /* timestamp no incoming packet at all */
        HWTSTAMP_FILTER_NONE,

        /* timestamp any incoming packet */
        HWTSTAMP_FILTER_ALL,
      /* return value: timestamp all packets requested plus some others */
        HWTSTAMP_FILTER_SOME,

        /* PTP v1, UDP, any kind of event packet */
        HWTSTAMP_FILTER_PTP_V1_L4_EVENT,
        /* PTP v1, UDP, Sync packet */
        HWTSTAMP_FILTER_PTP_V1_L4_SYNC,
        /* PTP v1, UDP, Delay_req packet */
        HWTSTAMP_FILTER_PTP_V1_L4_DELAY_REQ,
        /* PTP v2, UDP, any kind of event packet */
        HWTSTAMP_FILTER_PTP_V2_L4_EVENT,
        /* PTP v2, UDP, Sync packet */
        HWTSTAMP_FILTER_PTP_V2_L4_SYNC,
        /* PTP v2, UDP, Delay_req packet */
        HWTSTAMP_FILTER_PTP_V2_L4_DELAY_REQ,

        /* 802.AS1, Ethernet, any kind of event packet */
        HWTSTAMP_FILTER_PTP_V2_L2_EVENT,
        /* 802.AS1, Ethernet, Sync packet */
        HWTSTAMP_FILTER_PTP_V2_L2_SYNC,
        /* 802.AS1, Ethernet, Delay_req packet */
        HWTSTAMP_FILTER_PTP_V2_L2_DELAY_REQ,

        /* PTP v2/802.AS1, any layer, any kind of event packet */
        HWTSTAMP_FILTER_PTP_V2_EVENT,
        /* PTP v2/802.AS1, any layer, Sync packet */
        HWTSTAMP_FILTER_PTP_V2_SYNC,
        /* PTP v2/802.AS1, any layer, Delay_req packet */
        HWTSTAMP_FILTER_PTP_V2_DELAY_REQ,
};
Note: for receive side timestamping currently only HWTSTAMP_FILTER_NONE and
HWTSTAMP_FILTER_ALL are supported.
```

4.2.5.2 Getting Timestamping

Once timestamping is enabled, a timestamp is placed in the socket Ancillary data. recvmsg() can be used to get a control message for regular incoming packets. For sending timestamps, the outgoing packet is looped back to the socket's error queue with the send timestamp(s) attached. It can be received with recvmsg(flags=MSG_ERRQUEUE). The call returns the original outgoing packet data including all headers prepended down to include the link layer, the scm_timestamping control message, a sock_extended_err control message with ee_errno==ENOMSG and ee_origin==SO_EE_ORIGIN_TIMESTAMPING. A socket with such a pending bounced packet is ready for reading as far as select() is concerned. If the outgoing packet has to be fragmented, then only the first fragment is timestamped and returned to the sending socket.

> When timestamping is enabled, VLAN stripping is disabled. For more information, please refer to Documentation/networking/timestamping.txt in kernel.org

4.2.5.3 Querying Timestamping Capabilities via ethtool

➢ *To display timestamping capabilities via ethtool:*

- Show timestamping capabilities

```
ethtool -T eth<x>
```

Example:

```
ethtool -T eth0
Timestamping parameters for p2p1:
Capabilities:
        hardware-transmit     (SOF_TIMESTAMPING_TX_HARDWARE)
        software-transmit     (SOF_TIMESTAMPING_TX_SOFTWARE)
        hardware-receive      (SOF_TIMESTAMPING_RX_HARDWARE)
        software-receive      (SOF_TIMESTAMPING_RX_SOFTWARE)
        software-system-clock (SOF_TIMESTAMPING_SOFTWARE)
        hardware-raw-clock    (SOF_TIMESTAMPING_RAW_HARDWARE)
PTP Hardware Clock: none
Hardware Transmit Timestamp Modes:
        off                   (HWTSTAMP_TX_OFF)
        on                    (HWTSTAMP_TX_ON)
Hardware Receive Filter Modes:
        none                  (HWTSTAMP_FILTER_NONE)
        all                   (HWTSTAMP_FILTER_ALL)
```

For more details on PTP Hardware Clock, please refer to:

https://www.kernel.org/doc/Documentation/ptp/ptp.txt

4.2.6 RoCE Timestamping

RoCE timestamping allows you to stamp packets when they are sent to the wire / received from the wire. The timestamp is given in raw hardware cycles, but could be easily converted into hardware referenced nanoseconds based time. Additionally, it enables you to query the hardware for the hardware time, thus stamp other application's event and compare time.

4.2.6.1 Query Capabilities

Timestamping is available only if the hardware reports are capable of doing so. To verify if RoCE timestamping is available, run `ibv_exp_query_device`.

For example:

```
struct ibv_exp_device_attr attr;
ibv_exp_query_device(context, &attr);
if (attr.comp_mask & IBV_EXP_DEVICE_ATTR_WITH_TIMESTAMP_MASK) {
if (attr.timestamp_mask) {
                /* Timestamping is supported with mask attr.timestamp_mask */
        }
}
if (attr.comp_mask & IBV_EXP_DEVICE_ATTR_WITH_HCA_CORE_CLOCK) {
        if (attr.hca_core_clock) {
                /* reporting the device's clock is supported. */
                /* attr.hca_core_clock is the frequency in MHZ */
        }
}
```

4.2.6.2 Creating Timestamping Completion Queue

To get timestamps, a suitable extended Completion Queue (CQ) must be created via a special call to `ibv_exp_create_cq` verb.

```
cq_init_attr.flags = IBV_EXP_CQ_TIMESTAMP;
cq_init_attr.comp_mask = IBV_CQ_INIT_ATTR_FLAGS;
cq = ibv_exp_create_cq(context, cqe, node, NULL, 0, &cq_init_attr);
```

In ConnectX-3 adapter cards, this CQ cannot report SL or SLID information. The value of `sl` and `sl_id` fields in `struct ibv_exp_wc` are invalid. Only the fields indicated by the `exp_wc_flags` field in `struct ibv_exp_wc` contains a valid and usable value.

In ConnectX-3 adapter cards, when using timestamping, several fields of `struct ibv_-exp_wc` are not available resulting in RoCE UD / RoCE traffic with VLANs failure.

In ConnectX-4 and ConnectX-4 Lx adapter cards, timestamping in not available when CQE zipping is used.

4.2.6.3 Polling a Completion Queue

Polling a CQ for timestamp is done via the `ibv_exp_poll_cq` verb.

```
ret = ibv_exp_poll_cq(cq, 1, &wc_ex, sizeof(wc_ex));
if (ret > 0) {
                /* CQ returned a wc */
                if (wc_ex.wc_flags & IBV_EXP_WC_WITH_TIMESTAMP) {
                                /* This wc contains a timestamp */
                                timestamp = wc_ex.timestamp;
                                /* Timestamp is given in raw hardware time */
                }
}
```

> CQs that are opened with the `ibv_exp_create_cq` verbs should always be polled with the `ibv_exp_poll_cq` verb.

4.2.6.4 Querying the Hardware Time

Querying the hardware for time is done via the `ibv_exp_query_values` verb.

For example:

```
ret = ibv_exp_query_values(context, IBV_VALUES_HW_CLOCK, &queried_values);
if (!ret && queried_values.comp_mask & IBV_VALUES_HW_CLOCK)
queried_time = queried_values.hwclock;
```

To change the queried time in nanoseconds resolution, use the `IBV_VALUES_HW_CLOCK_NS` flag along with the `hwclock_ns` field.

```
ret = ibv_exp_query_values(context, IBV_EXP_VALUES_HW_CLOCK, &queried_values);
if (!ret && queried_values.comp_mask & IBV_EXP_VALUES_HW_CLOCK)
queried_time = queried_values.hwclock;
```

> In ConnectX-3 and ConnectX-3 Pro adapter cards, querying the Hardware Time is available only on physical functions / native machines.

4.2.7 Flow Steering

> Flow Steering is supported in ConnectX®-3, ConnectX®-3 Pro, ConnectX®-4, ConnectX®-4 Lx and ConnectX®-5 Ex adapter cards.

Flow steering is a new model which steers network flows based on flow specifications to specific QPs. Those flows can be either unicast or multicast network flows. In order to maintain flexibility, domains and priorities are used. Flow steering uses a methodology of flow attribute, which is

a combination of L2-L4 flow specifications, a destination QP and a priority. Flow steering rules may be inserted using ethtool. The verbs abstraction uses a different terminology from the flow attribute (`ibv_exp_flow_attr`), defined by a combination of specifications (`struct ibv_exp_-flow_spec_*`).

4.2.7.1 Enable/Disable Flow Steering

> Applicable to ConnectX®-3 and ConnectX®-3 Pro adapter cards only.
> In ConnectX®-4, ConnectX®-4 Lx and ConnectX®-5 Ex adapter cards, Flow Steering is automatically enabled as of MLNX_OFED v3.1-x.0.0.

Flow steering is generally enabled when the `log_num_mgm_entry_size` module parameter is non positive (e.g., `-log_num_mgm_entry_size`), which means the absolute value of the parameter is a bit field. Every bit indicates a condition or an option regarding the flow steering mechanism:

reserved	b5	b4	b3	b2	b1	b0

bit	Operation	Description
b0	Force device managed Flow Steering	When set to 1, it forces NIC to be enabled regardless of whether NC-SI Flow Steering is supported or not.
b1	Disable Flow Steering	When set to 1, it disables the support of Flow Steering. This bit should be set to 1 when "b2- Enable A0 static DMFS steering" is used.
b2	Enable A0 static DMFS steering	When set to 1, A0 static DMFS steering is enabled. This bit should be set to 0 when "b1- Disable Flow Steering" is 0.
b3	Enable DMFS only if the NIC supports more than 64QPs per MCG entry	When set to 1, DMFS is enabled only if the NIC supports more than 64 QPs attached to the same rule. For example, attaching 64VFs to the same multicast address causes 64QPs to be attached to the same MCG. If the NIC supports less than 64 QPs per MCG, B0 is used.
b4	Optimize steering table for non source IP rules when possible	When set to 1, steering table will be optimized to support rules ignoring source IP check. This optimization is available only when Flow Steering is set.
b5	Optimize steering table for non source IP rules when possible	When set to 1, steering table will be optimized to support rules ignoring source IP check. This optimization is possible only when DMFS mode is set.

For example, a value of (-7) means:

• Forcing Flow Steering regardless of NC-SI Flow Steering support

• Enabling A0 static DMFS steering

• Steering table is not optimized for rules ignoring source IP check

The default value of `log_num_mgm_entry_size` is -10 which means Ethernet Flow Steering is enabled by default if NC-SI DMFS is supported and the NIC supports at least 64 QPs per MCG entry. Otherwise, L2 steering (B0) is used.

When using SR-IOV, flow steering is enabled if there is an adequate amount of space to store the flow steering table for the guest/master.

➢ *To enable Flow Steering:*

Step 1. Open the `/etc/modprobe.d/mlnx.conf` file.

Step 2. Set the parameter `log_num_mgm_entry_size` to non positive value by writing the option `mlx4_core log_num_mgm_entry_size=<value>`.

Step 3. Restart the driver

➢ *To disable Flow Steering:*

Step 1. Open the `/etc/modprobe.d/mlnx.conf` file.

Step 2. Remove the `options mlx4_core log_num_mgm_entry_size== <value>`.

Step 3. Restart the driver

4.2.7.2 Flow Steering Support

➢ *[For ConnectX®-3 and ConnectX®-3 Pro only] To determine which Flow Steering features are supported:*

```
ethtool --show-priv-flags eth4
```

The following output is shown:

```
mlx4_flow_steering_ethernet_l2: on      Creating Ethernet L2 (MAC) rules is supported
mlx4_flow_steering_ipv4: on             Creating IPv4 rules is supported
mlx4_flow_steering_tcp: on              Creating TCP/UDP rules is supported
```

> For ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards, all supported features are enabled.

4.2.7.3 A0 Static Device Managed Flow Steering

> A0 Static Device Managed Flow Steering is supported in ConnectX®-3 and ConnectX®-3 Pro only.

This mode enables fast steering, however it may impact flexibility. Using it increases the packet rate performance by ~30%, with the following limitations for Ethernet link-layer unicast QPs:

• Limits the number of opened RSS Kernel QPs to 96. MACs should be unique (1 MAC per 1 QP). The number of VFs is limited.

- When creating Flow Steering rules for user QPs, only MAC--> QP rules are allowed. Both MACs and QPs should be unique between rules. Only 62 such rules could be created.

- When creating rules with Ethtool, MAC--> QP rules can be used, where the QP must be the indirection (RSS) QP. Creating rules that indirect traffic to other rings is not allowed. Ethtool MAC rules to drop packets (action -1) are supported.

- RFS is not supported in this mode.

- VLAN is not supported in this mode.

4.2.7.4 Flow Domains and Priorities

> ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards support only User Verbs domain with struct ibv_exp_flow_spec_eth flow specification using 4 priorities.

Flow steering defines the concept of a domain and its priority. Each domain represents a user agent that can attach a flow. The domains are prioritized. A higher priority domain will always supersede a lower priority domain when their flow specifications overlap. Setting a lower priority value will result in higher priority.

In addition to the domain, there is priority within each of the domains. Each domain can have at most 2^12 priorities in accordance with its needs.

The following are the domains at a descending order of priority:

- **User Verb** allows a user application QP to be attached into a specified flow when using `ibv_exp_create_flow` and `ibv_exp_destroy_flow` verbs

  - `ibv_create_flow`

    ```
    struct ibv_exp_flow *ibv_exp_create_flow(struct ibv_qp *qp, struct ibv_exp_flow_attr
    *flow)
    ```

    Input parameters:

    - `struct ibv_qp` - the attached QP.

    - `struct ibv_exp_flow_attr` - attaches the QP to the specified flow. The flow contains mandatory control parameters and optional L2, L3 and L4 headers. The optional headers are detected by setting the size and `num_of_specs` fields:

      `struct ibv_exp_flow_attr` can be followed by the optional flow headers structs:

      ```
      struct ibv_flow_spec_ib
      struct ibv_flow_spec_eth
      struct ibv_flow_spec_ipv4
      struct ibv_flow_spec_tcp_udp
      ```

      For further information, please refer to the `ibv_exp_create_flow` man page.

Be advised that from MLNX_OFED v2.0-3.0.0 and higher, the parameters (both the value and the mask) should be set in big-endian format.

Each header struct holds the relevant network layer parameters for matching. To enforce the match, the user sets a mask for each parameter.

The mlx5 driver supports partial masks. The mlx4 driver supports the following masks:

- All one mask - include the parameter value in the attached rule
  Note: Since the VLAN ID in the Ethernet header is 12bit long, the following parameter should be used: `flow_spec_eth.mask.vlan_tag = htons(0x0fff)`.

- All zero mask - ignore the parameter value in the attached rule

When setting the flow type to NORMAL, the incoming traffic will be steered according to the rule specifications. ALL_DEFAULT and MC_DEFAULT rules options are valid only for Ethernet link type.

For further information, please refer to the relevant man pages.

- `ibv_exp_destroy_flow`

```
int ibv_exp_destroy_flow(struct ibv_exp_flow *flow_id)
```

Input parameters:

`ibv_exp_destroy_flow` requires `struct ibv_exp_flow` which is the return value of `ibv_exp_create_flow` in the case of success.

Output parameters:

Returns the value of 0 on success, or the value of errno on failure.

For further information, please refer to the `ibv_exp_destroy_flow` man page.

- **Ethtool**

Ethtool domain is used to attach an RX ring, specifically its QP to a specified flow.

Please refer to the most recent ethtool manpage for all the ways to specify a flow.

Examples:

- ethtool –U eth5 flow-type ether dst 00:11:22:33:44:55 loc 5 action 2.

  All packets that contain the above destination MAC address are to be steered into rx-ring 2 (its underlying QP), with priority 5 (within the ethtool domain).

- ethtool –U eth5 flow-type tcp4 src-ip 1.2.3.4 dst-port 8888 loc 5 action 2.

  All packets that contain the above destination IP address and source port are to be steered into rx-ring 2. When destination MAC is not given, the user's destination MAC is filled automatically.

- ethtool –u eth5.

  Shows all of ethtool's steering rules.

When configuring two rules with the same priority, the second rule will overwrite the first one, so this ethtool interface is effectively a table. Inserting Flow Steering rules in the kernel requires support from both the ethtool in the user space and in kernel (v2.6.28).

- **MLX4 Driver Support**

The mlx4 driver supports only a subset of the flow specification the ethtool API defines. Asking for an unsupported flow specification will result with an "invalid value" failure.

The following are flow specific parameters:

Table 12 - Flow Specific Parameters

	ether	tcp4/udp4	ip4
Mandatory	dst		src-ip/dst-ip
Optional	vlan	src-ip, dst-ip, src-port, dst-port, vlan	src-ip, dst-ip, vlan

* **RFS**

    RFS is an in-kernel-logic responsible for load balancing between CPUs by attaching flows to CPUs that are used by flow's owner applications. This domain allows the RFS mechanism to use flow steering infrastructure to support the RFS logic by implementing the `ndo_rx_-flow_steer`, which, in turn, calls the underlying flow steering mechanism with the RFS domain.

    Enabling the RFS requires enabling the '`ntuple`' flag via ethtool.

    For example, to enable ntuple for eth0, run:

    ```
    ethtool -K eth0 ntuple on
    ```

    RFS requires the kernel to be compiled with the `CONFIG_RFS_ACCEL` option. This option is available in kernels 2.6.39 and above. Furthermore, RFS requires Device Managed Flow Steering support.

    RFS cannot function if LRO is enabled. LRO can be disabled via ethtool.

* **All of the rest**

    The lowest priority domain serves the following user:

    * **The mlx4 Ethernet driver** attaches its unicast and multicast MACs addresses to its QP using L2 flow specifications

    Fragmented UDP traffic cannot be steered. It is treated as 'other' protocol by hardware (from the first packet) and not considered as UDP traffic.

    We recommend using libibverbs v2.0-3.0.0 and libmlx4 v2.0-3.0.0 and higher as of MLNX_OFED v2.0-3.0.0 due to API changes.

4.2.7.5 Flow Steering Dump Tool

> Available in ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

The mlx_fs_dump is a python tool that prints the steering rules in a readable manner. Python v2.7 or above, as well as pip, anytree and termcolor libraries are required to be installed on the host.

Running example:

```
./ofed_scripts/utils/mlx_fs_dump -d /dev/mst/mt4115_pciconf0
FT: 9 (level: 0x18, type: NIC_RX)
+-- FG: 0x15 (MISC)
    |-- FTE: 0x0 (FWD) to (TIR:0x7e) out.ethtype:IPv4 out.ip_prot:UDP out.udp_dport:0x140
    +-- FTE: 0x1 (FWD) to (TIR:0x7e) out.ethtype:IPv4 out.ip_prot:UDP out.udp_dport:0x13f
...
```

For further information on the mlx_fs_dump tool, please refer to mlx_fs_dump Community post.

4.2.8 VXLAN Hardware Stateless Offloads

> Available in ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

VXLAN technology provides scalability and security solutions. It requires extension of the traditional stateless offloads to avoid performance drop. ConnectX-3 Pro, ConnectX-4 and ConnectX-4 Lx adapter cards offer the following stateless offloads for a VXLAN packet, similar to the ones offered to non-encapsulated packets. VXLAN protocol encapsulates its packets using outer UDP header.

Available hardware stateless offloads:

- Checksum generation (Inner IP and Inner TCP/UDP)
- Checksum validation (Inner IP and Inner TCP/UDP). This will allow the use of GRO (in ConnectX-3 Pro card only) for inner TCP packets.
- TSO support for inner TCP packets.
- RSS distribution according to inner packets attributes.
- Receive queue selection - inner frames may be steered to specific QPs.

VXLAN Hardware Stateless Offloads require the following prerequisites:

- NIC and their minimum firmware required:
  - ConnectX-3 Pro - Firmware v2.32.5100
  - ConnectX-4 - Firmware v12.17.20.52

- ConnectX-4 Lx - Firmware v14.14.2312
- ConnectX-5 Ex- Firmware v16.23.10.20
- Operating Systems:
  - RHEL7 or upstream kernel 3.12.10 (or higher)
- ConnectX-3 Pro Supported Features:
  - DMFS enabled
  - A0 static mode disabled

4.2.8.1 Enabling VXLAN Hardware Stateless Offloads for ConnectX-3 Pro

> Applies to ConnectX-3 Pro adapter cards.

To enable the VXLAN offloads support load the `mlx4_core` driver with Device-Managed Flow-steering (DMFS) enabled. DMFS is the default steering mode.

➢ *To verify it is enabled by the adapter card:*

Step 1. Open the `/etc/modprobe.d/mlnx.conf` file.

Step 2. Set the parameter `debug_level` to "1".

```
options mlx4_core debug_level=1
```

Step 3. Restart the driver.

Step 4. Verify in the dmesg that the tunneling mode is: `vxlan`.

The net-device will advertise the `tx-udp-tnl-segmentation` flag shown when running "`etht-hool -k $DEV | grep udp`" only when VXLAN is configured in the OpenvSwitch (OVS) with the configured UDP port.

For example:

```
$ ethtool -k eth0 | grep udp_tnl
tx-udp_tnl-segmentation: on
```

As of firmware version 2.31.5050, VXLAN tunnel can be set on any desired UDP port. If using previous firmware versions, set the VXLAN tunnel over UDP port 4789.

➢ *To add the UDP port to /etc/modprobe.d/vxlan.conf:*

```
options vxlan udp_port=<number decided above>
```

4.2.8.2 Enabling VXLAN Hardware Stateless Offloads for ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex Adapter Cards

> Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

VXLAN offload is enabled by default for ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards running the minimum required firmware version and a kernel version that includes VXLAN support.

To confirm if the current setup supports VXLAN, run:

```
ethtool -k $DEV | grep udp_tnl
```

Example:

```
# ethtool -k ens1f0 | grep udp_tnl
tx-udp_tnl-segmentation: on
```

ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards support configuring multiple UDP ports for VXLAN offload[1]. Ports can be added to the device by configuring a VXLAN device from the OS command line using the "ip" command.

Example:

```
# ip link add vxlan0 type vxlan id 10 group 239.0.0.10 ttl 10 dev ens1f0 dstport 4789
# ip addr add 192.168.4.7/24 dev vxlan0
# ip link set up vxlan0
```

The VXLAN ports can be removed by deleting the VXLAN interfaces.

Example:

```
# ip link delete vxlan0
```

> ➢ **To verify that the VXLAN ports are offloaded, use debugfs (if supported):**

Step 1. Mount debugfs.

```
# mount -t debugfs nodev /sys/kernel/debug
```

Step 2. List the offloaded ports.

```
ls /sys/kernel/debug/mlx5/$PCIDEV/VXLAN
```

Where $PCIDEV is the PCI device number of the relevant ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

Example:

```
# ls /sys/kernel/debug/mlx5/0000\:81\:00.0/VXLAN
4789
```

---

1. If you configure multiple UDP ports for offload and exceed the total number of ports supported by hardware, then those additional ports will still function properly, but will not benefit from any of the stateless offloads.

4.2.8.3 Important Notes

- VXLAN tunneling adds 50 bytes (14-eth + 20-ip + 8-udp + 8-vxlan) to the VM Ethernet frame. Please verify that either the MTU of the NIC that sends the packets (e.g. the VM virtionet NIC), or the host side veth device or the uplink takes into account the tunneling overhead. Meaning, the MTU of the sending NIC has to be decremented by 50 bytes (e.g 1450 instead of 1500), or the uplink NIC MTU has to be incremented by 50 bytes (e.g 1550 instead of 1500).

- From upstream 3.15-rc1 and onward, it is possible to use arbitrary UDP port for VXLAN. Note that this requires firmware version 2.31.2800 or higher. Additionally, you need to enable this kernel configuration option `CONFIG_MLX4_EN_VXLAN=y` (ConnectX-3 Pro only).

- On upstream kernels 3.12/3.13, GRO with VXLAN is not supported.

4.2.9 Ethtool

Ethtool is a standard Linux utility for controlling network drivers and hardware, particularly for wired Ethernet devices. It can be used to:

- Get identification and diagnostic information
- Get extended device statistics
- Control speed, duplex, autonegotiation and flow control for Ethernet devices
- Control checksum offload and other hardware offload features
- Control DMA ring sizes and interrupt moderation

The following are the ethtool supported options:

Table 13 - Ethtool Supported Options

Options	Description
ethtool --set-priv-flags eth\<x\> \<priv flag\> \<on/off\>	Enables/disables driver feature matching the given private flag.
ethtool --show-priv-flags eth\<x\>	Shows driver private flags and their states (ON/OFF) The private flag is: qcn_disable_32_14_4_e The flags below indicate the flow steering current configuration and limits. mlx4_flow_steering_ethernet_l2 mlx4_flow_steering_ipv4 mlx4_flow_steering_tcp For further information, refer to Flow Steering section. The flags below are related to *Ignore Frame Check Sequence*, and they are active when `ethtool -k` does not support them: orx-fcs orx-all

Options	Description
ethtool -a eth<x>	**Note:** Supported in ConnectX®-3/ConnectX®-3 Pro cards only. Queries the pause frame settings.
ethtool -A eth<x> [rx on|off] [tx on|off]	**Note:** Supported in ConnectX®-3/ConnectX®-3 Pro cards only. Sets the pause frame settings.
ethtool -c eth<x>	Queries interrupt coalescing settings.
ethtool -C eth<x> [pkt-rate-low N] [pkt-rate-high N] [rx-usecs-low N] [rx-usecs-high N]	**Note:** Supported in ConnectX®-3/ConnectX®-3 Pro cards only. Sets the values for packet rate limits and for moderation time high and low values. For further information, please refer to Adaptive Interrupt Moderation section.
ethtool -C eth<x> [rx-usecs N] [rx-frames N]	Sets the interrupt coalescing setting. rx-frames will be enforced immediately, rx-usecs will be enforced only when adaptive moderation is disabled. **Note:** usec settings correspond to the time to wait after the *last* packet is sent/received before triggering an interrupt.
ethtool -C eth<x> adaptive-rx on|off	**Note:** Supported in ConnectX®-3/ConnectX®-3 Pro cards only. Enables/disables adaptive interrupt moderation. By default, the driver uses adaptive interrupt moderation for the receive path, which adjusts the moderation time to the traffic pattern. For further information, please refer to Adaptive Interrupt Moderation section.
ethtool -g eth<x>	Queries the ring size values.
ethtool -G eth<x> [rx <N>] [tx <N>]	Modifies the rings size.

ethtool -i eth<x>	Checks driver and device information.
	 For example:
	`#> ethtool -i eth2`
	` driver: mlx4_en (MT_0DD0120009_CX3)`
	` version: 2.1.6 (Aug 2013)`
	` firmware-version: 2.30.3000`
	` bus-info: 0000:1a:00.0`

ethtool -k eth<x>	Queries the stateless offload status.

Options	Description
ethtool -K eth<x> [rx on|off] [tx on|off] [sg on|off] [tso on|off] [lro on|off] [gro on|off] [gso on|off] [rxvlan on|off] [txvlan on|off] [ntuple on/off] [rxhash on/off] [rx-all on/off] [rx-fcs on/off]	Sets the stateless offload status. TCP Segmentation Offload (TSO), Generic Segmentation Offload (GSO): increase outbound throughput by reducing CPU overhead. It works by queuing up large buffers and letting the network interface card split them into separate packets. Large Receive Offload (LRO): increases inbound throughput of high-bandwidth network connections by reducing CPU overhead. It works by aggregating multiple incoming packets from a single stream into a larger buffer before they are passed higher up the networking stack, thus reducing the number of packets that have to be processed. LRO is available in kernel versions < 3.1 for untagged traffic. Hardware VLAN insertion Offload (txvlan): When enabled, the sent VLAN tag will be inserted into the packet by the hardware. **Note:** LRO will be done whenever possible. Otherwise GRO will be done. Generic Receive Offload (GRO) is available throughout all kernels. Hardware VLAN Striping Offload (rxvlan): When enabled received VLAN traffic will be stripped from the VLAN tag by the hardware. RX FCS (rx-fcs): Keeps FCS field in the received packets. RX FCS validation (rx-all): Ignores FCS validation on the received packets. **Note:** The flags below are supported in ConnectX®-3/ConnectX®-3 Pro cards only: `[rxvlan on|off] [txvlan on|off] [ntuple on/off] [rxhash on/off] [rx-all on/off] [rx-fcs on/off]`
ethtool -l eth<x>	Shows the number of channels
ethtool -L eth<x> [rx <N>] [tx <N>]	Sets the number of channels **Note:** This also resets the RSS table to its default distribution, which is uniform across the physical cores on the close numa node. **Note:** For ConnectX®-4 cards, use `ethtool -L eth<x> combined <N>` to set both RX and TX channels.
ethtool -m|--dump-module-eeprom eth<x> [raw on|off] [hex on|off] [offset N] [length N]	Queries/Decodes the cable module eeprom information.
ethtool -p|--identify DEVNAME	Enables visual identification of the port by LED blinking [TIME-IN-SECONDS]
ethtool -p|--identify eth<x> <LED duration>	Allows users to identify interface's physical port by turning the ports LED on for a number of seconds. **Note:** The limit for the LED duration is 65535 seconds.
ethtool -S eth<x>	Obtains additional device statistics.

Options	Description
ethtool -s eth<x> advertise <N> autoneg on	Changes the advertised link modes to requested link modes <N> To check the link modes' hex values, run `<man ethtool>` and to check the supported link modes, run `ethtoo eth<x>` **NOTE**: <autoneg on> only sends a hint to the driver that the user wants to modify advertised link modes and not speed.
ethtool -s eth<x> msglvl [N]	Changes the current driver message level.
ethtool -s eth<x> speed <SPEED> autoneg off	Changes the link speed to requested <SPEED>. To check the supported speeds, run `ethtool eth<x>`. **NOTE**: `<autoneg off>` does not set autoneg OFF, it only hints the driver to set a specific speed.
ethtool -t eth<x>	Performs a self diagnostics test.
ethtool -T eth<x>	**Note:** Supported in ConnectX®-3/ConnectX®-3 Pro cards only. Shows Timestamping capabilities
ethtool -x eth<x>	Retrieves the receive flow hash indirection table.
ethtool -X eth<x> equal a b c...	Sets the receive flow hash indirection table. **Note:** The RSS table configuration is reset whenever the number of channels is modified (using ethtool -L command).

4.2.10 Counters

Counters are used to provide information about how well an operating system, an application, a service, or a driver is performing. The counter data helps determine system bottlenecks and fine-tune the system and application performance. The operating system, network, and devices provide counter data that an application can consume to provide users with a graphical view of how well the system is performing.

The counter index is a QP attribute given in QP context. Multiple QPs may be associated with the same counter set. If multiple QPs share the same counter, its value represents the cumulative total.

- ConnectX®-3 supports 127 different counters which are allocated as follows:
    - 4 counters reserved for PF - 2 counters for each port
    - 2 counters reserved for VF - 1 counter for each port
    - All other counters if exist are allocated by demand

4.2.10.1 RoCE Counters

- RoCE counters are available only through sysfs located under:
    - # /sys/class/infiniband/mlx4_*/ports/*/counters/
    - # /sys/class/infiniband/mlx4_*/ports/*/counters_ext/

4.2.10.1.1ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex Custom RoCE Counters

Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

Custom port counters provide the user with a clear indication about RDMA send/receive statistics and errors. The counters are available at:

```
/sys/class/infiniband/<device name>/mlx5_ports/<port_number>/counters
```

Counter	Description
rx_write_requests	Number of received WRITE request for the associated QP.
rx_read_requests	Number of received READ request for the associated QP.
rx_atomic_requests	Number of received ATOMIC request for the associated QP.
rx_dct_connect	Number of received connection request for the associated DCTs.
out_of_buffer	Number of dropped packets occurred due to lack of WQE for the associated QPs/RQs.
out_of_sequence	Number of out of sequence packets. IB only.
duplicate_request	Number of received duplicate packets. A duplicate request is a request that had been previously executed.
rnr_nak_retry_err	Number of received RNR NAC packets. The QP retry limit did not exceed.
packet_seq_err	Number of received NAK-Sequence error packets. The QP retry limit did not exceed.
implied_nak_err	Number of times the Requestor detected an ACK with a PSN larger than expected PSN for RDMA READ or ATOMIC response.

4.2.10.2 SR-IOV Counters

- Physical Function can also read Virtual Functions' port counters through sysfs located under:
  - # /sys/class/net/eth*/vf*/statistics/

4.2.10.3 Ethtool Counters

To display the network device Ethernet statistics, you can run the following::

```
Ethtool -S <devname>
```

Counter	Description
rx_packets	Total packets successfully received.
rx_bytes	Total bytes in successfully received packets.
rx_multicast_packets	Total multicast packets successfully received.
rx_broadcast_packets	Total broadcast packets successfully received.
rx_errors	Number of receive packets that contained errors preventing them from being deliverable to a higher-layer protocol.
rx_dropped	Number of receive packets which were chosen to be discarded even though no errors had been detected to prevent their being deliverable to a higher-layer protocol.
rx_length_errors	Number of received frames that were dropped due to an error in frame length
rx_over_errors	Number of received frames that were dropped due to hardware port receive buffer overflow
rx_crc_errors	Number of received frames with a bad CRC that are not runts, jabbers, or alignment errors
rx_jabbers	Number of received frames with a length greater than MTU octets and a bad CRC
rx_in_range_length_error	Number of received frames with a length/type field value in the (decimal) range [1500:46] (42 is also counted for VLAN tagged frames)
rx_out_range_length_error	Number of received frames with a length/type field value in the (decimal) range [1535:1501]
tx_packets	Total packets successfully transmitted.
tx_bytes	Total bytes in successfully transmitted packets.
tx_multicast_packets	Total multicast packets successfully transmitted.
tx_broadcast_packets	Total broadcast packets successfully transmitted.
tx_errors	Number of frames that failed to transmit
tx_dropped	Number of transmitted frames that were dropped
rx_prio_<i>_packets	Total packets successfully received with priority i.
rx_prio_<i>_bytes	Total bytes in successfully received packets with priority i.

Counter	Description
rx_novlan_packets	Total packets successfully received with no VLAN priority.
rx_novlan_bytes	Total bytes in successfully received packets with no VLAN priority.
tx_prio_<i>_packets	Total packets successfully transmitted with priority i.
tx_prio_<i>_bytes	Total bytes in successfully transmitted packets with priority i.
tx_novlan_packets	Total packets successfully transmitted with no VLAN priority.
tx_novlan_bytes	Total bytes in successfully transmitted packets with no VLAN priority.
rx_pause[a]	The total number of PAUSE frames received from the far-end port.
rx_pause_duration[1]	The total time in microseconds that far-end port was requested to pause transmission of packets.
rx_pause_transition[1]	The number of receiver transitions from XON state (paused) to XOFF state (non-paused)
tx_pause[1]	The total number of PAUSE frames sent to the far-end port
tx_pause_duration[1]	The total time in microseconds that transmission of packets has been paused
tx_pause_transition[1]	The number of transmitter transitions from XON state (paused) to XOFF state (non-paused)
vport_rx_unicast_packets	Unicast packets received successfully
vport_rx_unicast_bytes	Unicast packet bytes received successfully
vport_rx_multicast_packets	Multicast packets received successfully
vport_rx_multicast_bytes	Multicast packet bytes received successfully
vport_rx_broadcast_packets	Broadcast packets received successfully
vport_rx_broadcast_bytes	Broadcast packet bytes received successfully
vport_rx_dropped	Received packets discarded due to lack of software receive buffers (WQEs). Important indication to weather RX completion routines are keeping up with hardware ingress packet rate
vport_rx_filtered	Received packets dropped due to packet check that failed. For example: Incorrect VLAN, incorrect Ethertype, unavailable queue/QP or loopback prevention
vport_tx_unicast_packets	Unicast packets sent successfully
vport_tx_unicast_bytes	Unicast packet bytes sent successfully
vport_tx_multicast_packets	Multicast packets sent successfully
vport_tx_multicast_bytes	Multicast packet bytes sent successfully
vport_tx_broadcast_packets	Broadcast packets sent successfully
vport_tx_broadcast_bytes	Broadcast packet bytes sent successfully
vport_tx_dropped	Packets dropped due to transmit errors
rx_lro_aggregated	Number of packets processed by the LRO mechanism
rx_lro_flushed	Number of offloaded packets the LRO mechanism passed to kernel

Counter	Description
rx_lro_no_desc	LRO mechanism has no room to receive packets from the adapter. In normal condition, it should not increase
rx_alloc_failed	Number of times failed preparing receive descriptor
rx_csum_good	Number of packets received with good checksum
rx_csum_none	Number of packets received with no checksum indication
tx_chksum_offload	Number of packets transmitted with checksum offload
tx_queue_stopped	Number of times transmit queue suspended
tx_wake_queue	Number of times transmit queue resumed
tx_timeout	Number of times transmitter timeout
xmit_more	Number of times doorbell was not triggered due to skb xmit more.
tx_tso_packets	Number of packet that were aggregated
rx<i>_packets	Total packets successfully received on ring i
rx<i>_bytes	Total bytes in successfully received packets on ring i.
tx<i>_packets	Total packets successfully transmitted on ring i.
tx<i>_bytes	Total bytes in successfully transmitted packets on ring i.

a. Pause statistics can be divided into "prio_<i>", depending on PFC configuration set.

4.2.10.3.1 Persistent Naming

To avoid network interface renaming after boot or driver restart use the "`/etc/udev/rules.d/70-persistent-net.rules`" file.

• Example for Ethernet interfaces:

```
# PCI device 0x15b3:0x1003 (mlx4_core)
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="?*", ATTR{address}=="00:02:c9:fa:c3:50",
ATTR{dev_id}=="0x0", ATTR{type}=="1", KERNEL=="eth*", NAME="eth1"
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="?*", ATTR{address}=="00:02:c9:fa:c3:51",
ATTR{dev_id}=="0x0", ATTR{type}=="1", KERNEL=="eth*", NAME="eth2"
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="?*", ATTR{address}=="00:02:c9:e9:56:a1",
ATTR{dev_id}=="0x0", ATTR{type}=="1", KERNEL=="eth*", NAME="eth3"
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="?*", ATTR{address}=="00:02:c9:e9:56:a2",
ATTR{dev_id}=="0x0", ATTR{type}=="1", KERNEL=="eth*", NAME="eth4"
```

4.2.11 Single Root IO Virtualization (SR-IOV)

> Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

4.2.11.1 System Requirements

To set up an SR-IOV environment, the following is required:

*   MLNX_OFED Driver
*   A server with an SR-IOV-capable motherboard BIOS
*   Hypervisor that supports SR-IOV such as: Red Hat Enterprise Linux Server Version 6 or higher.
*   Mellanox ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards. Ethernet adapter cards with SR-IOV capability

4.2.11.2 Setting Up SR-IOV

Depending on your system, perform the steps below to set up your BIOS. The figures used in this section are for illustration purposes only. For further information, please refer to the appropriate BIOS User Manual.

Step 1.   Enable "SR-IOV" in the system BIOS.

Step 2. Enable " Virtualization Technology".
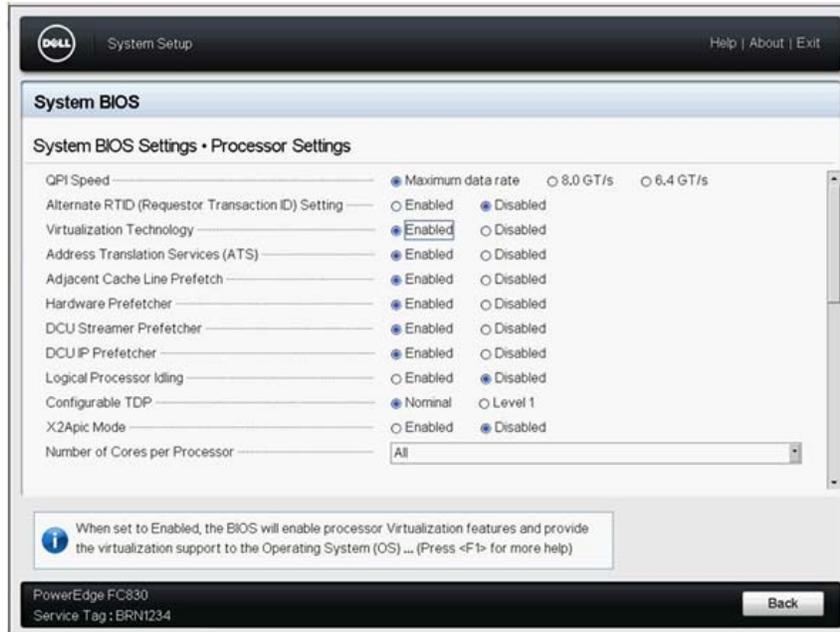
Step 3. Install a hypervisor that supports SR-IOV.

Step 4. Depending on your system, update the /boot/grub/grub.conf file to include a similar command line load parameter for the Linux kernel.

For example, to Intel systems, add:

```
default=0
timeout=5
splashimage=(hd0,0)/grub/splash.xpm.gz
hiddenmenu
title Red Hat Enterprise Linux Server (2.6.32-36.x86-645)
        root (hd0,0)
        kernel /vmlinuz-2.6.32-36.x86-64 ro root=/dev/VolGroup00/LogVol00 rhgb quiet
        intel_iommu=on[a]
        initrd /initrd-2.6.32-36.x86-64.img
```

a. Please make sure the parameter **"intel_iommu=on"** exists when updating the /boot/grub/grub.conf file, otherwise SR-IOV cannot be loaded.
Some OSes use /boot/grub2/grub.cfg file. If your server uses such file, please edit this file instead. (add "intel_iommu=on" for the relevant menu entry at the end of the line that starts with **"linux16"**).

4.2.11.2.1 Configuring SR-IOV for ConnectX-3/ConnectX-3 Pro

> Applies to ConnectX-3 and ConnectX-3 Pro adapter cards.

> For SR-IOV configuration, please refer to SR-IOV Configuration on page 216.

Step 1. Install MLNX_OFED driver for Linux that supports SR-IOV.

SR-IOV can be enabled and managed by using one of the following methods:

- Run the mlxconfig tool and set the `SRIOV_EN` parameter to "1" without re-burning the firmware

  To find the mst device run: "`mst start`" and "`mst status`"

  ```
  mlxconfig -d <mst_device> s SRIOV_EN=1
  ```

  For further information, please refer to section *"mlxconfig - Changing Device Configuration Tool"* in the MFT User Manual (www.mellanox.com > Products > Software > Firmware Tools).

- Burn firmware with SR-IOV support where the number of virtual functions (VFs) will be set to 16

  ```
  --enable-sriov
  ```

Step 2. Verify the NIC is configured to support SR-IOV.

```
# mstflint -dev <PCI Device> dc
```

1. Verify in the [NIC] section the following fields appear[1].

```
[NIC]
num_pfs = 1
total_vfs = <0-126>
sriov_en = true
```

Parameter	Recommended Value
num_pfs	1 **Note:** This field is optional and might not always appear.
total_vfs	• The recommended value is 63. **Note:** Before setting number of VFs in SR-IOV, please make sure your system can support that amount of VFs. Setting number of VFs larger than what your Hardware and Software can support may cause your system to cease working.
sriov_en	true

2. Add the above fields to the INI if they are missing.

---

1. If SR-IOV is supported, to enable SR-IOV (if it is not enabled), it is sufficient to set "sriov_en = true" in the INI.

3. Set the `total_vfs` parameter to the desired number if you need to change the number of total VFs.

4. Reburn the firmware using the mlxburn tool if the fields above were added to the INI, or the `total_vfs` parameter was modified.

If the mlxburn is not installed, please download it from Mellanox website http://www.mellanox.com => Products => Firmware tools

```
mlxburn -fw ./fw-ConnectX3-rel.mlx -dev /dev/mst/mt4099_pci_cr0 -conf ./MCX341A-
XCG_Ax.ini
```

Step 3. Create the text file /etc/modprobe.d/mlx4_core.conf if it does not exist.

Step 4. Insert an "options" line in the /etc/modprobe.d/mlx4_core.conf file to set the number of VFs, the protocol type per port, and the allowed number of virtual functions to be used by the physical function driver (probe_vf).

For example:

```
options mlx4_core num_vfs=5 port_type_array=1,2 probe_vf=1
```

Parameter	Recommended Value

num_vfs	• If absent, or zero: no VFs will be available
	• If its value is a single number in the range of 0-63: The driver will enable the `num_vfs` VFs on the NIC and this will be applied to all ConnectX® NIC on the host.
	• If its a triplet x,y,z (applies only if all ports are configured as Ethernet) the driver creates:
	  • x single port VFs on physical port 1
	  • y single port VFs on physical port 2 (applies only if such a port exist)
	  • z n-port VFs (where n is the number of physical ports on device). This applies to all ConnectX® NICs on the host

Parameter	Recommended Value

num_vfs	• If its format is a string: The string specifies the `num_vfs` parameter separately per installed NIC.
	The string format is: "bb:dd.f-v,bb:dd.f-v,…"
	  • bb:dd.f = bus:device.function of the PF of the NIC
	  • v = number of VFs to enable for that NIC which is either a single value or a triplet, as described above.
	For example:
	• `num_vfs=5` - The driver will enable 5 VFs on the NIC and this will be applied to all ConnectX® NICs on the host
	• `num_vfs=00:04.0-5,00:07.0-8` - The driver will enable 5 VFs on the NIC positioned in BDF 00:04.0 and 8 on the one in 00:07.0)
	• `num_vfs=1,2,3` - The driver will enable 1 VF on physical port 1, 2 VFs on physical port 2 and 3 dual port VFs (applies only to dual port NIC when all ports are Ethernet ports).
	• `num_vfs=00:04.0-5;6;7,00:07.0-8;9;10` - The driver will enable:
	  • NIC positioned in BDF 00:04.0
	    • 5 single VFs on port 1
	    • 6 single VFs on port 2
	    • 7 dual port VFs
	  • NIC positioned in BDF 00:07.0
	    • 8 single VFs on port 1
	    • 9 single VFs on port 2
	    • 10 dual port VFs
	Applies when all ports are configure as Ethernet in dual port NICs
	Notes:
	• PFs not included in the above list will not have SR-IOV enabled.
	• Triplets and single port VFs are only valid when all ports are configured as Ethernet.
	• The second parameter in a triplet is valid only when there are more than 1 physical port.
	In a triplet, $x+z<=63$ and $y+z<=63$, the maximum number of VFs on each physical port must be 63.

port_type_array	Specifies the protocol type of the ports. It is either one array of 2 port types `'t1,t2'` for all devices or list of BDF to `port_type_array` `'bb:dd.f-t1;t2,...'`. (string) Valid port types: 1-ib, 2-eth, 3-auto, 4-N/A If only a single port is available, use the N/A port type for port2 (e.g '1,4'). Note that this parameter is valid only when `num_vfs` is not zero (i.e., SRIOV is enabled). Otherwise, it is ignored.

Parameter	Recommended Value

probe_vf	• If absent or zero: no VF interfaces will be loaded in the Hypervisor/host
	• If num_vfs is a number in the range of 1-63, the driver running on the Hypervisor will itself activate that number of VFs. All these VFs will run on the Hypervisor. This number will apply to all ConnectX® NICs on that host.
	• If its a triplet x,y,z (applies only if all ports are configured as Ethernet), the driver probes:
	  • x single port VFs on physical port 1
	  • y single port VFs on physical port 2 (applies only if such a port exist)
	  • z n-port VFs (where n is the number of physical ports on device). Those VFs are attached to the hypervisor.
	• If its format is a string: the string specifies the `probe_vf` parameter separately per installed NIC.
	  The string format is: "bb:dd.f-v,bb:dd.f-v,…
	  • bb:dd.f = bus:device.function of the PF of the NIC
	  • v = number of VFs to use in the PF driver for that NIC which is either a single value or a triplet, as described above
	For example:
	• `probe_vfs=5` - The PF driver will activate 5 VFs on the NIC and this will be applied to all ConnectX® NICs on the host
	• `probe_vfs=00:04.0-5,00:07.0-8` - The PF driver will activate 5 VFs on the NIC positioned in BDF 00:04.0 and 8 for the one in 00:07.0)
	• `probe_vf=1,2,3` - The PF driver will activate 1 VF on physical port 1, 2 VFs on physical port 2 and 3 dual port VFs (applies only to dual port NIC when all ports are Ethernet ports).
	  This applies to all ConnectX® NICs in the host.
	• `probe_vf=00:04.0-5;6;7,00:07.0-8;9;10` - The PF driver will activate:
	  • NIC positioned in BDF 00:04.0
	    • 5 single VFs on port 1
	    • 6 single VFs on port 2
	    • 7 dual port VFs
	  • NIC positioned in BDF 00:07.0
	    • 8 single VFs on port 1
	    • 9 single VFs on port 2
	    • 10 dual port VFs
	Applies when all ports are configure as Ethernet in dual port NICs.

probe_vf	**Notes:** • PFs not included in the above list will not activate any of their VFs in the PF driver • Triplets and single port VFs are only valid when all ports are configured as Ethernet. • The second parameter in a triplet is valid only when there are more than 1 physical port • Every value (either a value in a triplet or a single value) should be less than or equal to the respective value of `num_vfs` parameter

The example above loads the driver with 5 VFs (num_vfs). The standard use of a VF is a single VF per a single VM. However, the number of VFs varies upon the working mode requirements.

Step 5. Reboot the server.

 If SR-IOV is not supported by the server, the machine might not come out of boot/load.

Step 6. Load the driver and verify SR-IOV is supported. Run:

```
lspci | grep Mellanox
03:00.0 InfiniBand: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR /
10GigE] (rev b0)
03:00.1 InfiniBand: Mellanox Technologies MT27500 Family [ConnectX-3 Virtual Function] (rev b0)
03:00.2 InfiniBand: Mellanox Technologies MT27500 Family [ConnectX-3 Virtual Function] (rev b0)
03:00.3 InfiniBand: Mellanox Technologies MT27500 Family [ConnectX-3 Virtual Function] (rev b0)
03:00.4 InfiniBand: Mellanox Technologies MT27500 Family [ConnectX-3 Virtual Function] (rev b0)
03:00.5 InfiniBand: Mellanox Technologies MT27500 Family [ConnectX-3 Virtual Function] (rev b0)
```

Where:

- "03:00" represents the Physical Function
- "03:00.**X**" represents the Virtual Function connected to the Physical Function

4.2.11.2.2 Configuring SR-IOV for ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

> Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

> For SR-IOV configuration, please refer to SR-IOV Configuration on page 216.

Step 1. Install the MLNX_OFED driver for Linux that supports SR-IOV.

Step 2. Check if SR-IOV is enabled in the firmware.

```
mlxconfig -d /dev/mst/mt4113_pciconf0 q

  Device #1:
  ----------

  Device type:    Connect4
  PCI device:     /dev/mst/mt4115_pciconf0
  Configurations:         Current
     SRIOV_EN             1
     NUM_OF_VFS           8
     FPP_EN               1
```

> FPP_EN=1 is not supported in ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex and will fail.
> If needed, use mlxconfig to set the relevant fields:
> ```
> mlxconfig -d /dev/mst/mt4113_pciconf0 set SRIOV_EN=1 NUM_OF_VFS=16
> FPP_EN=1
> ```

> The supported number of VFs is 31 per PF.

Step 3. Either reboot or reset the firmware.

```
mlxfwreset / reboot
```

Step 4. Write to the sysfs file the number of Virtual Functions you need to create for the PF.

You can use one of the following equivalent files:

For Ethernet

• A standard Linux kernel generated file that is available in the new kernels.

```
echo [num_vfs] > /sys/class/net/<eth_interface>/device/sriov_numvfs
```

- A file generated by the mlx5_core driver with the same functionality as the kernel generated one. Used by old kernels that do not have the standard file.

```
echo [num_vfs] > /sys/class/net/<eth_interface>/device/mlx5_num_vfs
```

The following rules apply when writing to these files:

- If there are no VFs assigned, the number of VFs can be changed to any valid value (0 - max #VFs as set during FW burning)

- If there are VFs assigned to a VM, it is not possible to change the number of VFs

- If the administrator unloads the driver on the PF while there are no VFs assigned, the driver will unload and SRI-OV will be disabled

- If there are VFs assigned while the driver of the PF is unloaded, SR-IOV is not disabled. This means VFs will be visible on the VM. However, they will not be operational. This is applicable to OSes with kernels that use `pci_stub` and not `vfio`.

  - The VF driver will discover this situation and will close its resources
  - When the driver on the PF is reloaded, the VF becomes operational. The administrator of the VF will need to restart the driver in order to resume working with the VF.

Step 5.   Load the driver. To verify that the VFs were created. Run:

```
lspci | grep Mellanox
08:00.0 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4]
08:00.1 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4]
08:00.2 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4 Virtual
Function]
08:00.3 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4 Virtual
Function]
08:00.4 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4 Virtual
Function]
08:00.5 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4 Virtual
Function]
```

Step 6.   Configure the VFs.

After VFs are created, 3 sysfs entries per VF are available under `/sys/class/infini-band/mlx5_<PF INDEX>/device/sriov` (shown below for VFs 0 to 2):

```
+-- 0
|   +-- node
|   +-- policy
|   +-- port
+-- 1
|   +-- node
|   +-- policy
|   +-- port
+-- 2
    +-- node
    +-- policy
    +-- port
```

- Policy - The vport's policy. The policy can be one of:

  The user can set the port GUID by writing to the `/sys/class/infiniband/<PF>/device/sriov/<index>/port` file.

  - Down - the VPort PortState remains 'Down'

- Up - if the current VPort PortState is 'Down', it is modified to 'Initialize'. In all other states, it is unmodified. The result is that the SM may bring the VPort up.
- Follow - follows the PortState of the physical port. If the PortState of the physical port is 'Active', then the VPort implements the 'Up' policy. Otherwise, the VPort PortState is 'Down'.

Notes:

- The policy of all the vports is initialized to "Down" after the PF driver is restarted except for VPort0 for which the policy is modified to 'Follow' by the PF driver.
- To see the VFs configuration, you must unbind and bind them or reboot the VMs if the VFs were assigned.

Step 7. Make sure that the SM supports Virtualization.

The `/etc/opensm/opensm.conf` file should contain the following line:

```
virt_enabled 2
```

4.2.11.2.2.1 Note on VFs Initialization

Since the same `mlx5_core` driver supports both Physical and Virtual Functions, once the Virtual Functions are created, the driver of the PF will attempt to initialize them so they will be available to the OS owning the PF. If you want to assign a Virtual Function to a VM, you need to make sure the VF is not used by the PF driver. If a VF is used, you should first unbind it before assigning to a VM.

➢ *To unbind a device use the following command:*

1. Get the full PCI address of the device.

```
lspci -D
```

Example:

```
0000:09:00.2
```

2. Unbind the device.

```
echo 0000:09:00.2 > /sys/bus/pci/drivers/mlx5_core/unbind
```

3. Bind the unbound VF.

```
echo 0000:09:00.2 > /sys/bus/pci/drivers/mlx5_core/bind
```

4.2.11.2.2.2 PCI BDF Mapping of PFs and VFs

PCI addresses are sequential for both the PF and their VFs. Assuming the card's PCI slot is 05:00 and it has 2 ports, the PFs PCI address will be 05:00.0 and 05:00.1.

Given 3 VFs per PF, the VFs PCI addresses will be:

```
05:00.2-4 for VFs 0-2 of PF 0 (mlx5_0)
05:00.5-7 for VFs 0-2 of PF 1 (mlx5_1)
```

4.2.11.3 Uninstalling SR-IOV Driver

➢ *To uninstall SR-IOV driver, perform the following:*

Step 1. For Hypervisors, detach all the Virtual Functions (VF) from all the Virtual Machines (VM) or stop the Virtual Machines that use the Virtual Functions.

Please be aware, stopping the driver when there are VMs that use the VFs, will cause the machine to hang.

Step 2. Run the script below. Please be aware, uninstalling the driver deletes the entire driver's file, but does not unload the driver.

```
[root@swl022 ~]# /usr/sbin/ofed_uninstall.sh
This program will uninstall all OFED packages on your machine.
Do you want to continue?[y/N]:y
Running /usr/sbin/vendor_pre_uninstall.sh
Removing OFED Software installations

Running /bin/rpm -e --allmatches kernel-ib kernel-ib-devel libibverbs libibverbs-devel
libibverbs-devel-static libibverbs-utils libmlx4 libmlx4-devel libibcm libibcm-devel
libibumad libibumad-devel libibumad-static libibmad libibmad-devel libibmad-static
librdmacm librdmacm-utils librdmacm-devel ibacm opensm-libs opensm-devel perftest com-
pat-dapl compat-dapl-devel dapl dapl-devel dapl-devel-static dapl-utils srptools infini-
band-diags-guest ofed-scripts opensm-devel

warning: /etc/infiniband/openib.conf saved as /etc/infiniband/openib.conf.rpmsave
Running /tmp/2818-ofed_vendor_post_uninstall.sh
```

Step 3. Restart the server.

4.2.12 PFC Configuration Using LLDP DCBX

4.2.12.1 PFC Configuration on Hosts

4.2.12.1.1PFC Auto-Configuration Using LLDP Tool in the OS

Step 1. Start lldpad daemon on host.

```
lldpad -d Or
service lldpad start
```

Step 2. Send lldpad packets to the switch.

```
lldptool set-lldp -i <ethX>  adminStatus=rxtx  ;
lldptool -T -i <ethX>   -V  sysName enableTx=yes;
lldptool -T -i <ethX>   -V  portDesc enableTx=yes ;
lldptool -T -i <ethX>   -V  sysDesc enableTx=yes
lldptool -T -i <ethX>   -V sysCap enableTx=yess
lldptool -T -i <ethX>   -V mngAddr enableTx=yess
lldptool -T -i <ethX>   -V PFC enableTx=yes;
lldptool -T -I <ethX>   -V CEE-DCBX enableTx=yes;
```

Step 3. Set the PFC parameters.

* For the CEE protocol, use dcbtool:

```
dcbtool sc <ethX> pfc pfcup:<xxxxxxxx>
```

Example:

```
dcbtool sc eth6 pfc pfcup:01110001
```

where:

[pfcup:xxxxxxxx]	Enables/disables priority flow control. From left to right (priorities 0-7) - x can be equal to either 0 or 1. 1 indicates that the priority is configured to transmit priority pause.

* For IEEE protocol, use lldptool:

```
lldptool -T -i <ethX> -V PFC enabled=x,x,x,x,x,x,x,x
```

Example:

```
lldptool -T -i eth2 -V PFC enabled=1,2,4
```

where:

enabled	Displays or sets the priorities with PFC enabled. The set attribute takes a comma-separated list of priorities to enable, or the string `none` to disable all priorities.

4.2.12.1.2PFC Auto-Configuration Using LLDP in the Firmware (for mlx5 driver)

There are two ways to configure PFC and ETS on the server:

1. Local Configuration - Configuring each server manually.

2. Remote Configuration - Configuring PFC and ETS on the switch, after which the switch will pass the configuration to the server using LLDP DCBX TLVs.

    There are two ways to implement the remote configuration using ConnectX-4 and ConnectX-4 Lx adapters:

    a. Configuring the adapter firmware to enable DCBX.

    b. Configuring the host to enable DCBX.

For further information on how to auto-configure PFC using LLDP in the firmware, refer to the HowTo Auto-Config PFC and ETS on ConnectX-4 and ConnectX-4 Lx via LLDP DCBX Community post.

4.2.13 Data Plane Development Kit (DPDK)

DPDK is a set of libraries and optimized NIC drivers for fast packet processing in user space. DPDK provides a framework and common API for high speed networking applications.

Mellanox Poll Mode Driver (PMD) is designed for fast packet processing and low latency by providing kernel bypass for receive, send, and by avoiding the interrupt processing performance overhead.

To install and configure mlx4 and mlx5 DPDK Poll-Mode Driver (PMD) for Mellanox ConnectX®-3 Pro and ConnectX®-4, ConnectX®-4 Lx and ConnectX-5 Ex Ethernet adapters, please refer to Data Plane Development Kit (DPDK).

4.3 VMware Driver for ConnectX-3 and ConnectX-3 Pro

Applies to ConnectX-3 and ConnectX-3 Pro adapter cards.

The following procedure requires custom boot image downloading, mounting and booting from a USB device.

For VMware, download and install the latest Mellanox Ethernet Driver for VMware vSphere 5.5 and 6.0 from the VMware support site: http://www.vmware.com/support.

4.3.1 Installing and Running the Driver

To install the driver:

Please uninstall any previous Mellanox driver packages prior to installing the new version.

1. Log into the ESXi server with root permissions.
2. Install the driver.

```
#> esxcli software vib install -d <path>/<bundle_file>
```

3. Reboot the machine.
4. Verify the driver was installed successfully.

```
# esxcli software vib list | grep MEL
net-ib-core             2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-ib-ipoib            2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-ib-mad              2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-ib-sa               2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-mlx-compat          2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-mlx4-core           2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-mlx4-en             2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-mlx4-ib             2.4.0.0-1OEM.600.0.0.2494585    MEL    PartnerSupported  2016-01-31
net-mst                 4.3.0.4-1OEM.600.0.0.2494585    MEL    PartnerSupported  2015-12-10
```

After the installation process, all kernel modules are loaded automatically upon boot.

4.3.2 Removing Mellanox OFED Driver

> Please unload the driver before removing it.

➢ *To remove all drivers:*

1. Log into the ESXi server with root permissions.

2. List the existing OFED driver modules.

3. Remove each module.

```
#> esxcli software vib remove -n net-ib-ipoib
#> esxcli software vib remove -n net-mlx4-ib
#> esxcli software vib remove -n net-ib-sa
#> esxcli software vib remove -n net-ib-mad
#> esxcli software vib remove -n net-ib-core
#> esxcli software vib remove -n net-mlx4-en
#> esxcli software vib remove -n net-mlx4-core
#> esxcli software vib remove -n net-mlx-compat
```

> To remove the modules, the command must be run in the same order as shown in the example above.

4. Reboot the server.

4.3.3 Loading/Unloading Driver Kernel Modules

➢ *To unload the driver:*

```
#> /opt/mellanox/bin/openibd.sh stop
```

➢ *To load the driver:*

```
#> /opt/mellanox/bin/openibd.sh start
```

➢ *To restart the driver:*

```
#> /opt/mellanox/bin/openibd.sh restart
```

4.3.4 Firmware Programming

Firmware updates on ESX systems need to be done via iDRAC or Lifecycle Controller. Please see Updating Firmware using Dell iDRAC or Lifecycle Controller on page 186.

4.4   VMware Driver for ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex

Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

The following procedure requires custom boot image downloading, mounting and booting from a USB device.

For VMware, download and install the latest Mellanox Ethernet Driver for VMware vSphere 5.5 and 6.0 from the VMware support site: http://www.vmware.com/support.

4.4.1   Installing VMware

Please uninstall any previous Mellanox driver packages prior to installing the new version.

➢ *To install the driver:*

1. Log into the ESXi server with root permissions.
2. Install the driver.

```
#> esxcli software vib install -d <path>/<bundle_file>
```

Example:

```
#> esxcli software vib install -d <path>/<bundle_file>
```

3. Reboot the machine.
4. Verify the driver was installed successfully.

```
# esxcli software vib list | grep mlx
ESX 5.5:
nmlx5-core                    4.5.2.0-1OEM.550.0.0.1391871      MEL      PartnerSupported  2016-02-01
ESX 6.0:
nmlx5-core                    4.15.4.0-1OEM.600.0.0.2768847     MEL      PartnerSupported  2016-02-01
```

After the installation process, all kernel modules are loaded automatically upon boot.

4.4.2   Removing Previous Mellanox Driver

> Please unload the driver before removing it.

➢ *To remove all the drivers:*

1. Log into the ESXi server with root permissions.

2. List the existing ConnectX-4 /ConnectX-4 Lx/ ConnectX-5 Ex NATIVE ESX driver modules. (See Step 4 in page 91)

3. Remove each module.

```
#> esxcli software vib remove -n nmlx5-core
```

> To remove the modules, the command must be run in the same order as shown in the example above.

4. Reboot the server.

4.4.3   Loading/Unloading Driver Kernel Modules

➢ *To unload the driver:*

```
esxcfg-module -u nmlx5_core
```

➢ *To load the driver:*

```
/etc/init.d/sfcbd-watchdog stop
esxcfg-module nmlx5_core
/etc/init.d/sfcbd-watchdog start
kill -POLL $(cat /var/run/vmware/vmkdevmgr.pid)
```

➢ *To restart the driver:*

```
/etc/init.d/sfcbd-watchdog stop
esxcfg-module -u nmlx5_core
esxcfg-module nmlx5_core
/etc/init.d/sfcbd-watchdog start
kill -POLL $(cat /var/run/vmware/vmkdevmgr.pid)
```

4.4.4   Firmware Programming

Firmware updates on ESX systems need to be done via iDRAC or Lifecycle Controller. Please see Updating Firmware using Dell iDRAC or Lifecycle Controller on page 186.

4.5   Windows

WinOF supports ConnectX-3 and ConnectX-3 Pro adapter cards.

WinOF-2 supports ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

You do not need to download both if you are only using one type of card.

For Windows, download and install the latest Mellanox OFED for Windows (WinOF and WinOF-2) software package available at the Dell support site http://www.dell.com/support.

4.5.1   Installation Requirements

AMD EPYC based systems require WinOF-2 version 1.70 and higher.

4.5.1.1   Required Disk Space for Installation

- WinOF - 100 MB
- WinOF-2 - 120 MB

4.5.2   Software Requirements

- Microsoft Windows Server operating system

For the list of supported operating systems, please refer to the release notes file accompanying the Windows Driver Dell Update Package on the Dell support site.

4.5.2.1   Installer Privileges

- The installation requires administrator privileges on the target machine

4.5.3   Downloading Mellanox WinOF / WinOF-2

Step 1.   For Mellanox ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex Ethernet adapters, download the latest WinOF/WinOF-2 Update Package for Windows from Dell's support site http://www.dell.com/support.

Step 2.   Download the .exe image according to the operating system of your machine.

4.5.4   Installing Mellanox WinOF / WinOF-2

This section provides instructions for two types of installation procedures:

- • "Attended Installation"

An installation procedure that requires frequent user intervention.

- • "Unattended Installation"

An automated installation procedure that requires no user intervention.

> Both Attended and Unattended installations require administrator privileges.

4.5.4.1 Attended Installation

Double click the Dell Update Package and follow the GUI instructions to install Mellanox WinOF/WinOF-2.

4.5.4.2 Unattended Installation

From a CMD console, execute the Dell Update Package silently.

```
Network_Driver_NNNNN_WN_XX.XX.XX.EXE /s
```

> For a list of Dell Update Package command line options, execute the Dell Update Package with the option "/?" or "/h"
> ```
> Network_Driver_NNNNN_WN_XX.XX.XX.EXE /?
> ```

4.5.5   Uninstalling Mellanox WinOF / WinOF-2 Driver

➢ *To uninstall Mellanox_WinOF on a single node, perform one of the following options:*

Option 1. Click Start-> Control Panel-> Programs and Features. (NOTE: This requires elevated administrator privileges)

Option 2. Double click the Dell Update Package and follow the instructions of the install wizard.

Option 3. Click Start-> All Programs-> Mellanox Technologies-> MLNX_WinOF/MLNX-_WinOF-2-> Uninstall MLNX_WinOF/Uninstall MLNX_WinOF-2.

4.6 WinOF / WinOF-2 Features

4.6.1 Ethernet Network

Applies to ConnectX-3 and ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

4.6.1.1 Packet Burst Handling

This feature allows packet burst handling, while avoiding packet drops that may occur when a large amount of packets is sent in a short period of time.

1. By default, the feature is disabled, and the AsyncRecieveIndicate registry key is set to 0. To enable the feature, choose one of the following options:
   a. To enable packet burst buffering using threaded DPC (recommended), set the AsyncRecieveIndicate registry key to 1.
   b. To enable packet burst buffering using polling, set the AsyncRecieveIndicate to 2.

2. To control the number of reserved receive packets, set the RfdReservationFactor registry key:

Default	150
Recommended	10,000
Maximum	5,000,000

The memory consumption will increase in accordance with the "RfdReservationFactor" registry key value.

4.6.1.2 Assigning Port IP After Installation

By default, your machine is configured to obtain an automatic IP address via a DHCP server. In some cases, the DHCP server may require the MAC address of the network adapter installed in your machine.

➢ *To obtain the MAC address:*

Step 1. Open a CMD console-> Click Start-> Task Manager-> File-> Run new task-> and enter CMD.

Step 2. Display the MAC address as "Physical Address"

```
> ipconfig /all
```

Configuring a static IP is the same for Ethernet adapters.

➢ *To assign a static IP address to a network port after installation:*

Step 1. Open the Network Connections window. Locate Local Area Connections with Mellanox devices.
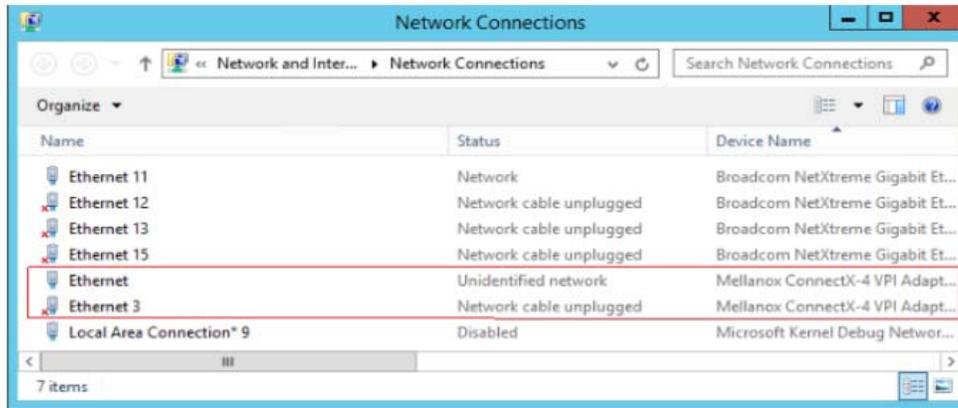
Step 2. Right-click a Mellanox Local Area Connection and left-click Properties.
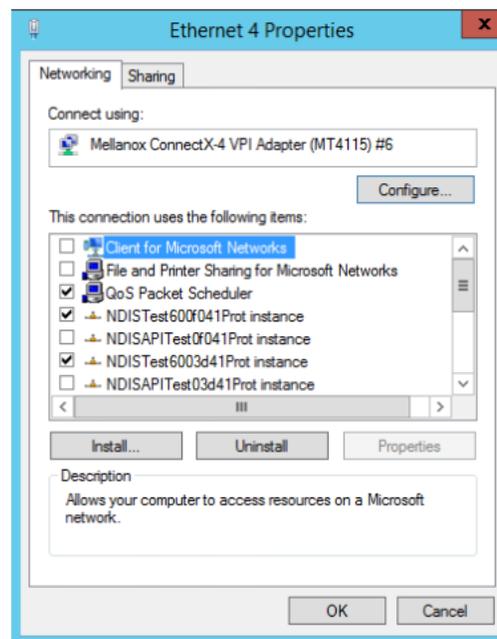
Step 3. Select Internet Protocol Version 4 (TCP/IPv4) from the scroll list and click Properties.

Step 4. Select the "Use the following IP address:" radio button and enter the desired IP information.
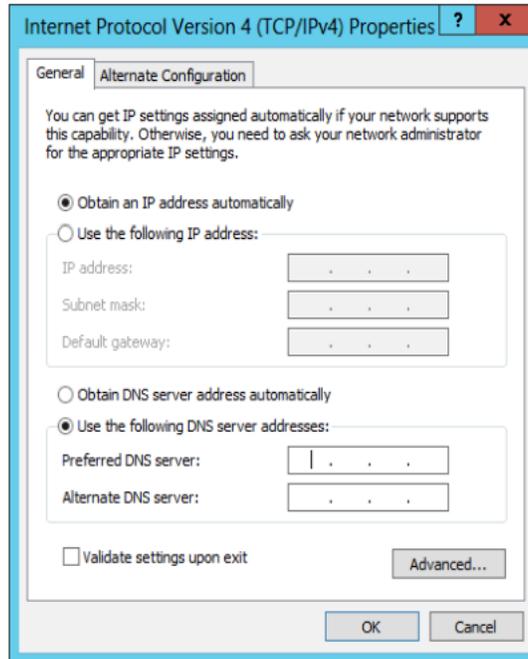
Step 5. Click OK.

Step 6. Close the Local Area Connection dialog.

Step 7. Verify the IP configuration by running 'ipconfig' from a CMD console.

```
> ipconfig
...
Ethernet adapter Local Area Connection 4:

   Connection-specific DNS Suffix  . :
   IP Address. . . . . . . . . . . : 11.4.12.63
   Subnet Mask . . . . . . . . . . : 255.255.0.0
   Default Gateway . . . . . . . . :
...
```

4.6.2 Configuring Quality of Service (QoS)

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

Prior to configuring Quality of Service, you must install Data Center Bridging using one of the following methods:

➢ **To Disable Flow Control Configuration**

Device Manager -> Network Adapters -> Mellanox Ethernet Adapters -> Properties -> Advanced Tab

Figure 7: Device Manager - Example
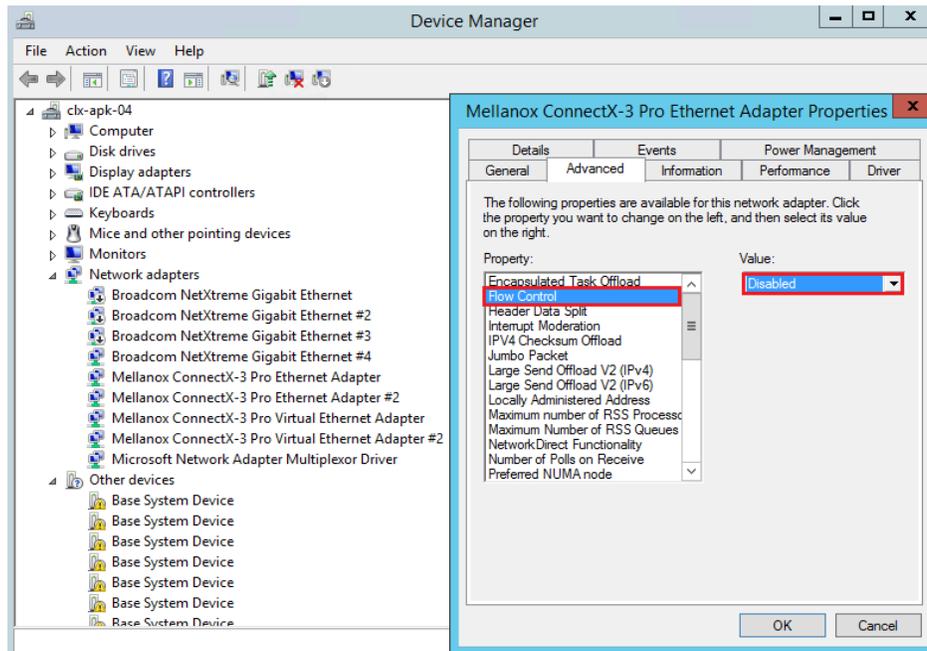
➢ **To install Data Center Bridging using the Server Manager:**

Step 1. Open the 'Server Manager'.

Step 2. Select 'Add Roles and Features'.

Step 3. Click Next.

Step 4. Select 'Features' on the left panel.

Step 5. Check the 'Data Center Bridging' checkbox.

Step 6. Click 'Install'.

➢ **To install Data Center Bridging using PowerShell:**

Step 1. Enable Data Center Bridging (DCB).

```
PS $ Install-WindowsFeature Data-Center-Bridging
```

➢ **To configure QoS on the host:**

The procedure below is not saved after you reboot your system. Hence, we recommend you create a script using the steps below and run it on the startup of the local machine. Please see the procedure below on how to add the script to the local machine startup scripts.

Step 1. Change the Windows PowerShell execution policy.

```
PS $ Set-ExecutionPolicy AllSigned
```

Step 2. Remove the entire previous QoS configuration.

```
PS $ Remove-NetQosTrafficClass
PS $ Remove-NetQosPolicy -Confirm:$False
```

Step 3. Set the DCBX Willing parameter to false as Mellanox drivers do not support this feature.

```
PS $ set-NetQosDcbxSetting -Willing 0
```

Step 4. Create a Quality of Service (QoS) policy and tag each type of traffic with the relevant priority.

In this example, TCP/UDP priority 1, SMB over TCP use priority 3.

```
PS $ New-NetQosPolicy "DEFAULT" -store Activestore -Default -PriorityValue8021Action 3
PS $ New-NetQosPolicy "TCP" -store Activestore -IPProtocolMatchCondition TCP -Priority-
Value8021Action 1
PS $ New-NetQosPolicy "UDP" -store Activestore -IPProtocolMatchCondition UDP -Priority-
Value8021Action 1
New-NetQosPolicy "SMB" -SMB -PriorityValue8021Action 3
```

Step 5. Create a QoS policy for SMB over SMB Direct traffic on Network Direct port 445.

```
PS $ New-NetQosPolicy "SMBDirect" -store Activestore -NetDirectPortMatchCondition 445 -
PriorityValue8021Action 3
```

Step 6. [Optional] If VLANs are used, mark the egress traffic with the relevant VlanID.
The NIC is referred as "Ethernet 4" in the examples below.

```
PS $ Set-NetAdapterAdvancedProperty -Name "Ethernet 4" -RegistryKeyword "VlanID" -Reg-
istryValue "55"
```

Step 7. [Optional] Configure the IP address for the NIC.

If DHCP is used, the IP address will be assigned automatically.

```
PS $ Set-NetIPInterface -InterfaceAlias "Ethernet 4" -DHCP Disabled
PS $ Remove-NetIPAddress -InterfaceAlias "Ethernet 4" -AddressFamily IPv4 -Con-
firm:$false
PS $ New-NetIPAddress -InterfaceAlias "Ethernet 4" -IPAddress 192.168.1.10 -Prefix-
Length 24 -Type Unicast
```

Step 8. [Optional] Set the DNS server (assuming its IP address is 192.168.1.2).

```
PS $ Set-DnsClientServerAddress -InterfaceAlias "Ethernet 4" -ServerAddresses
192.168.1.2
```

> After establishing the priorities of ND/NDK traffic, the priorities must have PFC enabled on them.

Step 9. Disable Priority Flow Control (PFC) for all other priorities except for 3.

```
PS $ Disable-NetQosFlowControl 0,1,2,4,5,6,7
```

Step 10. Enable QoS on the relevant interface.

```
PS $ Enable-NetAdapterQos -InterfaceAlias "Ethernet 4"
```

Step 11. Enable PFC on priority 3.

```
PS $ Enable-NetQosFlowControl -Priority 3
```

Step 12. Configure Priority 3 to use ETS. (ConnectX-3/ConnectX-3 Pro only)

```
PS $ New-NetQosTrafficClass -name "SMB class" -priority 3 -bandwidthPercentage 50 -
Algorithm ETS
```

➢ *To add the script to the local machine startup scripts:*

Step 1. From the PowerShell invoke:

```
gpedit.msc
```

Step 2. In the pop-up window, under the 'Computer Configuration' section, perform the following:

1. Select Windows Settings

2. Select Scripts (Startup/Shutdown)

3. Double click Startup to open the Startup Properties

4. Move to "PowerShell Scripts" tab
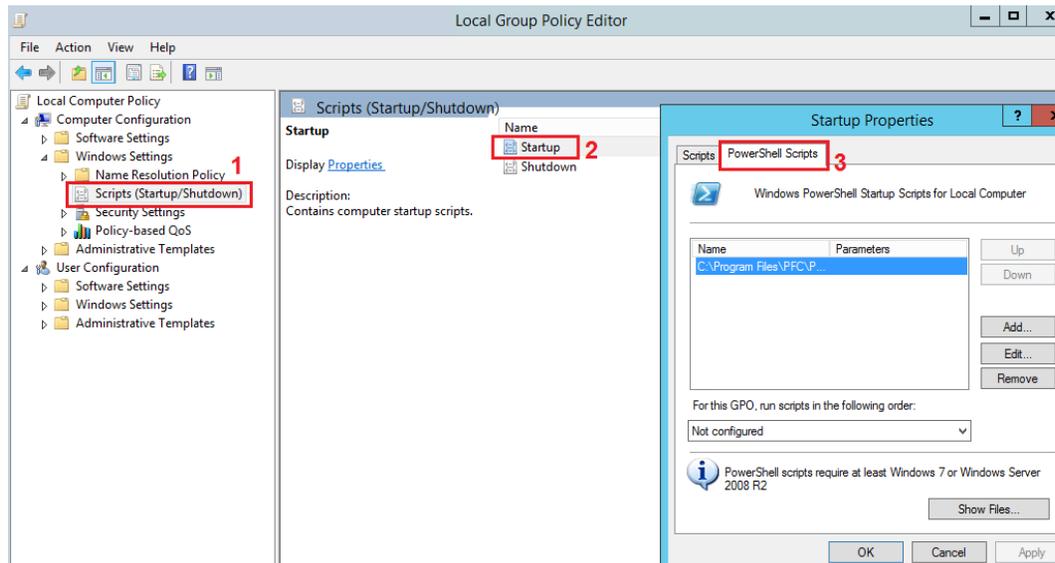
5. Click Add

The script should include only the following commands:

```
PS $ Remove-NetQosTrafficClass
PS $ Remove-NetQosPolicy -Confirm:$False
PS $ set-NetQosDcbxSetting -Willing 0
PS $ New-NetQosPolicy "SMB" -Policystore Activestore -NetDirectPortMatchCondition 445 -
PriorityValue8021Action 3
PS $ New-NetQosPolicy "DEFAULT" -Policystore Activestore -Default -PriorityValue8021Ac-
tion 3
PS $ New-NetQosPolicy "TCP" -Policystore Activestore -IPProtocolMatchCondition TCP -
PriorityValue8021Action 1
PS $ New-NetQosPolicy "UDP" -Policystore Activestore -IPProtocolMatchCondition UDP -
PriorityValue8021Action 1
```

```
PS $ Disable-NetQosFlowControl 0,1,2,4,5,6,7
PS $ Enable-NetAdapterQos -InterfaceAlias "port1"
PS $ Enable-NetAdapterQos -InterfaceAlias "port2"
PS $ Enable-NetQosFlowControl -Priority 3
PS $ New-NetQosTrafficClass -name "SMB class" -priority 3 -bandwidthPercentage 50 -
Algorithm ETS
```

6. Browse for the script's location.

7. Click OK

8. To confirm the settings applied after boot run:

```
PS $ get-netqospolicy -policystore activestore
```

4.6.2.1 Enhanced Transmission Selection

Enhanced Transmission Selection (ETS) provides a common management framework for assignment of bandwidth to frame priorities as described in the IEEE 802.1Qaz specification:

http://www.ieee802.org/1/files/public/docs2008/az-wadekar-ets-proposal-0608-v1.01.pdf

An example that you can use would be iSCSI with priority 4 with bandwidth allocation of 80% and PFC enabled:

```
PS $ New-NetQosPolicy -Name "iSCSI" -iSCSI -PriorityValue8021Action 4
PS $ New-NetQosTrafficClass "iscsi bw" -Algorithm ETS -Priority 4 -BandwidthPercentage
80
PS $ Enable-NetQosFlowControl -Priority 4
```

For further details on configuring ETS on Windows™ Server, please refer to:

https://docs.microsoft.com/en-us/powershell/module/dcbqos/?view=win10-ps

4.6.3   Differentiated Services Code Point (DSCP)

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

DSCP is a mechanism used for classifying network traffic on IP networks. It uses the 6-bit Differentiated Services Field (DS or DSCP field) in the IP header for packet classification purposes. Using Layer 3 classification enables you to maintain the same classification semantics beyond local network, across routers.

Every transmitted packet holds the information allowing network devices to map the packet to the appropriate 802.1Qbb CoS. For DSCP based PFC or ETS the packet is marked with a DSCP value in the Differentiated Services (DS) field of the IP header.

4.6.3.1 System Requirements

- Operating Systems: Windows Server 2012, Windows Server 2012 R2 and Windows Server 2016

• Firmware version: 12/14.20.1820 or higher

4.6.3.2 Setting the DSCP in the IP Header

Marking DSCP value in the IP header is done differently for IP packets constructed by the NIC (e.g. RDMA traffic) and for packets constructed by the IP stack (e.g. TCP traffic).

• For IP packets generated by the IP stack, the DSCP value is provided by the IP stack. The NIC does not validate the match between DSCP and Class of Service (CoS) values. CoS and DSCP values are expected to be set through standard tools, such as PowerShell command New-NetQosPolicy using PriorityValue8021Action and DSCPAction flags respectively.

• For IP packets generated by the NIC (RDMA), the DSCP value is generated according to the CoS value programmed for the interface. CoS value is set through standard tools, such as PowerShell command New-NetQosPolicy using PriorityValue8021Action flag. The NIC uses a mapping table between the CoS value and the DSCP value configured through the RroceDscpMarkPriorityFlow- Control[0-7] Registry keys

4.6.3.3 Configuring Quality of Service for TCP and RDMA Traffic

Step 1. Verify that DCB is installed and enabled (is not installed by default).

```
PS $ Install-WindowsFeature Data-Center-Bridging
```

Step 2. Import the PowerShell modules that are required to configure DCB.

```
PS $ import-module NetQos
PS $ import-module DcbQos
PS $ import-module NetAdapter
```

Step 3. Enable Network Adapter QoS.

```
PS $ Set-NetAdapterQos -Name "CX4_P1" -Enabled 1
```

Step 4. Enable Priority Flow Control (PFC) on the specific priority 3,5.

```
PS $ Enable-NetQosFlowControl 3,5
```

4.6.3.4 Configuring DSCP to Control PFC for TCP Traffic

• Create a QoS policy to tag All TCP/UDP traffic with CoS value 3 and DSCP value 9.

```
PS $ New-NetQosPolicy "DEFAULT" -Default -PriorityValue8021Action 3 -DSCPAction 9
```

DSCP can also be configured per protocol.

```
PS $ New-NetQosPolicy "TCP"  -IPProtocolMatchCondition TCP -PriorityValue8021Action 3 -
DSCPAction 16
PS $ New-NetQosPolicy "UDP"  -IPProtocolMatchCondition UDP -PriorityValue8021Action 3 -
DSCPAction 32
```

4.6.3.5 Configuring DSCP to Control ETS for TCP Traffic

- Create a QoS policy to tag All TCP/UDP traffic with CoS value 0 and DSCP value 8.

```
PS $ New-NetQosPolicy "DEFAULT" -Default -PriorityValue8021Action 0 -DSCPAction 8 -Pol-
icyStore activestore
```

- Configure DSCP with value 16 for TCP/IP connections with a range of ports.

```
PS $ New-NetQosPolicy "TCP1" -DSCPAction 16 -IPDstPortStartMatchCondition 31000 -IPDst-
PortEndMatchCondition 31999 -IPProtocol TCP -PriorityValue8021Action 0 -PolicyStore
activestore
```

- Configure DSCP with value 24 for TCP/IP connections with another range of ports.

```
PS $ New-NetQosPolicy "TCP2" -DSCPAction 24 -IPDstPortStartMatchCondition 21000 -IPDst-
PortEndMatchCondition 31999 -IPProtocol TCP -PriorityValue8021Action 0 -PolicyStore
activestore
```

- Configure two Traffic Classes with bandwidths of 16% and 80%.

```
PS $ New-NetQosTrafficClass -name "TCP1" -priority 3 -bandwidthPercentage 16 -Algorithm
ETS
PS $ New-NetQosTrafficClass -name "TCP2" -priority 5 -bandwidthPercentage 80 -Algorithm
ETS
```

4.6.3.6 Configuring DSCP to Control PFC for RDMA Traffic

- Create a QoS policy to tag the ND traffic for port 10000 with CoS value 3.

```
PS $ New-NetQosPolicy "ND10000" -NetDirectPortMatchCondition 10000 - PriorityVal-
ue8021Action 3
```

Related Commands:

- Get-NetAdapterQos - Gets the QoS properties of the network adapter
- Get-NetQosPolicy - Retrieves network QoS policies
- Get-NetQosFlowControl - Gets QoS status per priority

4.6.3.7 Receive Trust State

Received packets Quality of Service classification can be done according to the DSCP value, instead of PCP, using the RxTrustedState registry key. The mapping between wire DSCP values to the OS priority (PCP) is static, as follows:

Table 14 - DSCP to PCP Mapping

DSCP Value	Priority
0-7	0
8-15	1
16-23	2
24-31	3

Table 14 - DSCP to PCP Mapping

DSCP Value	Priority
32-39	4
40-47	5
48-55	6
56-63	7

When using this feature, it is expected that the transmit DSCP to Priority mapping (the Priority-ToDscpMappingTable _* registry key) will match the above table to create a consistent mapping on both directions.

4.6.3.8 Registry Settings

The following attributes must be set manually and will be added to the miniport registry.

Table 15 - DSCP Registry Keys Settings

Registry Key	Description
TxUntagPriorityTag	If 0x1, do not add 802.1Q tag to transmitted packets which are assigned 802.1p priority, but are not assigned a non-zero VLAN ID (i.e. priority-tagged). Default 0x0, for DSCP based PFC set to 0x1. Note: These packets will count on the original priority, even if the registry is on.
RxUntaggedMapToLossless	If 0x1, all untagged traffic is mapped to the lossless receive queue. Default 0x0, for DSCP based PFC set to 0x1.
PriorityToDscpMappingTable_<ID>	A value to mark DSCP for RoCE packets assigned to CoS=ID, when priority flow control is enabled. The valid values range is from 0 to 63, Default is ID value, e.g. PriorityToDscpMapping-Table_3 is 3. ID values range from 0 to 7.
DscpBasedEtsEnabled	If 0x1 - all DSCP based ETS feature is enabled, if 0x0 - disabled. Default 0x0.
DscpForGlobalFlowControl	Default DSCP value for flow control. Default 0x1a.
RxTrustedState	Default using host priority (PCP) is 1 Default using DSCP value is 2

For changes to take effect, please restart the network adapter after changing any of the above registry keys.

4.6.3.8.1 Default Settings

When DSCP configuration registry keys are missing in the miniport registry, the following defaults are assigned:

Table 16 - DSCP Default Registry Keys Settings

Registry Key	Default Value
TxUntagPriorityTag	0
RxUntaggedMapToLossles	0
PriorityToDscpMappingTable_0	0
PriorityToDscpMappingTable_1	1
PriorityToDscpMappingTable_2	2
PriorityToDscpMappingTable_3	3
PriorityToDscpMappingTable_4	4
PriorityToDscpMappingTable_5	5
PriorityToDscpMappingTable_6	6
PriorityToDscpMappingTable_7	7
DscpBasedEtsEnabled	eth:0
DscpForGlobalFlowControl	26

4.6.3.9 DSCP Sanity Testing

To verify that all QoS and DSCP settings are correct, you can capture the incoming and outgoing traffic by using the mlx5cmd sniffer tool. The tool allows you to see the DSCP value in the captured packets, as displayed in the figure below.
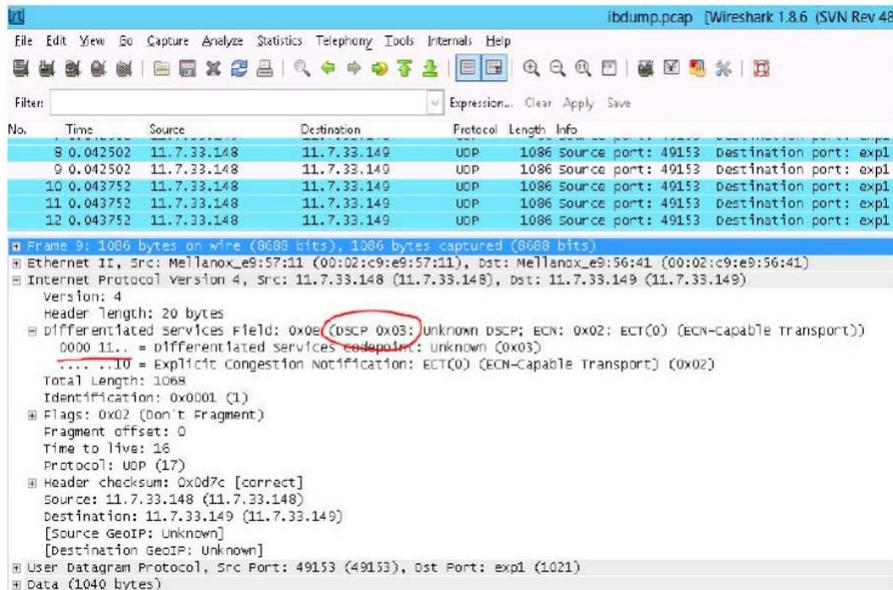
4.6.4 Configuring the Ethernet Driver

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

The following steps describe how to configure advanced features.

Step 1. Display the Device Manager.

Step 2. Right-click a Mellanox network adapter (under "Network adapters" list) and left-click Properties. Select the Advanced tab from the Properties sheet.

Step 3. Modify configuration parameters to suit your system.

Please note the following:

- For help on a specific parameter/option, check the help button at the bottom of the dialog.
- If you select one of the entries Offload Options, Performance Options, or Flow Control Options, you'll need to click the Properties button to modify parameters via a pop-up dialog.

4.6.5 Receive Segment Coalescing (RSC)

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

RSC allows reduction of CPU utilization when dealing with large TCP message size. It allows the drive to indicate to the Operating System once, per-message and not per-MTU that Packet Offload can be disabled for IPv4 or IPv6 traffic in the Advanced tab of the driver proprieties.

4.6.6 Receive Side Scaling (RSS)

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

RSS settings can be set per individual adapters as well as globally.

➢ *To do so, set the registry keys listed below:*

.

Table 17 - Registry Keys Setting

Sub-key	Description
HKLM\SYSTEM\CurrentControlSet\Control\Class\{4d36e972-e325-11ce-bfc1-08002be10318}\<nn>*MaxRSSProcessors	**Maximum number of CPUs allotted.** Sets the desired maximum number of processors for each interface. The number can be different for each interface. Note: Restart the network adapter after you change this registry key.

Table 17 - Registry Keys Setting

Sub-key	Description
HKLM\SYSTEM\CurrentControlSet\Control\Class\{4d36e972-e325-11ce-bfc1-08002be10318}\<nn>*RssBaseProcNumber	**Base CPU number.** Sets the desired base CPU number for each interface. The number can be different for each interface. This allows partitioning of CPUs across network adapters. Note: Restart the network adapter when you change this registry key.
HKLM\SYSTEM\CurrentControlSet\Control\Class\{4d36e972-e325-11ce-bfc1-08002be10318}\<nn>*NumaNodeID	NUMA node affinitization
HKLM\SYSTEM\CurrentControlSet\Control\Class\{4d36e972-e325-11ce-bfc1-08002be10318}\<nn>*RssBaseProcGroup	Sets the RSS base processor group for systems with more than 64 processors.

4.6.7   Wake on LAN (WoL)

Applies to the following adapter cards:
- Mellanox ConnectX-3 10 GbE Mezzanine card
- Mellanox ConnectX-3 Pro 10 GbE Mezzanine card
- Mellanox ConnectX-4 Lx 25 GbE Rack NDC

Wake on LAN is a technology that allows a network admin to remotely power on a system or to wake it up from sleep mode by a network message.

For Wake on LAN configuration, please refer to .

4.6.8   Data Center Bridging Exchange (DCBX)

Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

In WinOF-2 version 1.50 and newer, DCBX mode is set to "Firmware Controlled" by default. In order to allow DCBX control exchange via third party software, DCBX mode needs to be set to "Host in Charge". This setting can be changed in the Driver Advanced Properties or with the PowerShell Command:
```
Set-NetAdapterAdvancedProperty -Name "Adapter Name" -DisplayName "DcbxMode" -DisplayValue "Host in Charge."
```

WinOF-2 supports inserting priority tagging for RDMA traffic only when set by a local administrator. WinOF-2 does not support applying peer application settings.

Data Center Bridging Exchange (DCBX) protocol is an LLDP based protocol which manages and negotiates host and switch configuration. The WinOF-2 driver supports the following:

- PFC - Priority Flow Control

- ETS - Enhance Transmission Selection

- Application priority

The protocol is widely used to assure lossless path when running multiple protocols at the same time. It is functional as part of configuring QoS mentioned in Section 4.6.2, "Configuring Quality of Service (QoS)", on page 97. Users should make sure the willing bit on the host is enabled, using PowerShell if needed.:

```
set-NetQosSetting -Willing 1
```

This is required to allow negotiating and accepting peer configurations. Willing bit is set to 1 by default by the operating system.

The new settings can be queried by calling the following command in PowerShell

```
Get-NetAdapterQos
```

Note: The below configuration was received from the switch in the below example.

The output would look like the following:

```
PS C:\Users\Administrator> get-netadapterqos

Name    : Ethernet 9
Enabled : True

Name    : Ethernet 10
Enabled : True

Name                        : Ethernet 7
Enabled                     : True
Capabilities                :                       Hardware      Current
                                                     --------      -------
                              MacSecBypass         : NotSupported NotSupported
                              DcbxSupport          : IEEE         IEEE
                              NumTCs(Max/ETS/PFC)  : 8/8/8        8/8/8

OperationalTrafficClasses   : TC TSA    Bandwidth Priorities
                              -- ---    --------- ----------
                              0 ETS     25%       0-1
                              1 ETS     25%       2-3
                              2 ETS     25%       4-5
                              3 ETS     25%       6-7

OperationalFlowControl      : Priorities 0-4 Enabled
OperationalClassifications  : Not Available
RemoteTrafficClasses        : TC TSA    Bandwidth Priorities
                              -- ---    --------- ----------
                              0 ETS     25%       0-1
                              1 ETS     25%       2-3
                              2 ETS     25%       4-5
                              3 ETS     25%       6-7

RemoteFlowControl           : Priorities 0-4 Enabled
RemoteClassifications       : Not Available
```

In a scenario where both peers are set to Willing, the adapter with a lower MAC address takes the settings of the peer.

is disabled in the driver by default and in some firmware versions as well.

➢ *To use DCBX:*

1. Download the MFT Package from www.mellanox.com.
2. Install the package.
3. Query and enable in the firmware.

    a. Install WinMFT package and go to `\Program Files\Mellanox\WinMFT`

    b. Get the list of devices, run "mst status".

   c. Verify the DCBX is enabled or disabled, run "`mlxconfig.exe -d mt4117_pciconf0 query`".

    d. If disabled, run the following commands for dual-port cards.

```
mlxconfig -d mt4117_pciconf0 set LLDP_NB_RX_MODE_P1=2
mlxconfig -d mt4117_pciconf0 set LLDP_NB_TX_MODE_P1=2
mlxconfig -d mt4117_pciconf0 set LLDP_NB__P1=1
mlxconfig -d mt4117_pciconf0 set LLDP_NB_RX_MODE_P2=2
mlxconfig -d mt4117_pciconf0 set LLDP_NB_TX_MODE_P2=2
mlxconfig -d mt4117_pciconf0 set LLDP_NB__P2=1
```

4. Add the "`DcbxMode`" registry key, set the value to "2" and reload the adapter.
The registry key should be added to `HKEY_LOCAL_MACHINE\SYSTEM\CurrentControl-`

```
Set\Control\Class\{4d36e972-e325-11ce-bfc1-08002be10318}\<IndexValue>
```
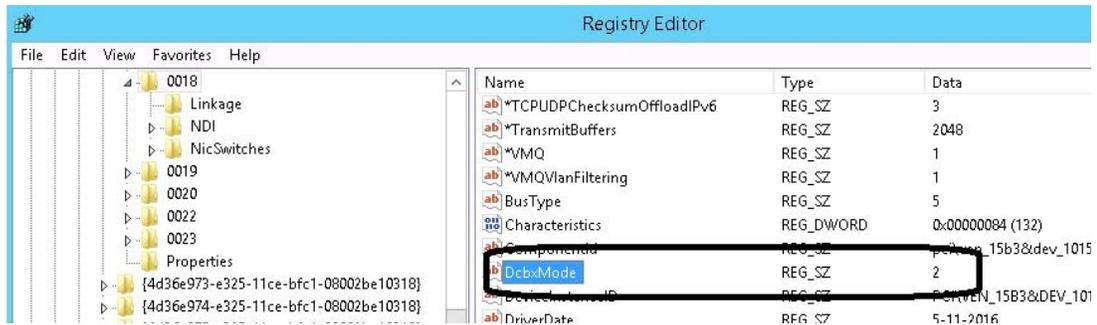

4.6.9   Receive Path Activity Monitoring

> Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

In the event where the device or the Operating System unexpectedly becomes unresponsive for a long period of time, the Flow Control mechanism may send pause frames, which will cause congestion spreading to the entire network.

To prevent this scenario, the device monitors its status continuously, attempting to detect when the receive pipeline is stalled. When the device detects a stall for a period longer than a pre-configured timeout, the Flow Control mechanisms (Global Pause and PFC) are automatically disabled.

If the PFC is in use, and one or more priorities are stalled, the PFC will be disabled on all priorities. When the device detects that the stall has ceased, the flow control mechanism will resume with its previously configured behavior.

Two registry parameters control the mechanism's behavior: the DeviceRxStallTime-out key controls the time threshold for disabling the flow control, and the DeviceRxStallWatermark key controls a diagnostics counter that can be used for early detection of stalled receive. WinOF-2 provides two counters to monitor the activity of this feature: "Minor Stall Watermark Reached" and "Critical Stall Watermark Reached".

4.6.10   Head of Queue Lifetime Limit

> Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

This feature enables the system to drop the packets that have been awaiting transmission for a long period of time, preventing the system from hanging. The implementation of the feature complies with the Head of Queue Lifetime Limit (HLL) definition.

The HLL has three registry keys for configuration:

TCHeadOfQueueLifeTimeLimit, TCStallCount and TCHeadOfQueueLifeTimeLimitEnable

4.6.11 Threaded DPC

> Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

A threaded DPC is a DPC that the system executes at IRQL = PASSIVE_LEVEL. An ordinary DPC preempts the execution of all threads, and cannot be preempted by a thread or by another DPC. If the system has a large number of ordinary DPCs queued, or if one of those DPCs runs for a long period of time, every thread will remain paused for an arbitrarily long period of time. Thus, each ordinary DPC increases the system latency, which can damage the performance of time-sensitive applications, such as audio or video playback.

Conversely, a threaded DPC can be preempted by an ordinary DPC, but not by other threads. Therefore, the user should use threaded DPCs rather than ordinary DPCs, unless a particular DPC must not be preempted, even by another DPC.

4.6.11.1 Registry Configuration

4.6.11.1.1 Mlx4_bus Registry Parameters

To enable or disable this feature in the driver, set the below registry key.

Location:

HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\mlx4_bus\Parameters

Table 18 - Threaded DPC Registry Keys

Key Name	Key Type	Values	Notes
ThreadDpcEnable	DWORD	• 0 = Disabled • 1 = Enabled	If the registry key *doesn't* exist, driver will set TheadDpc as enabled for *Azure* packages

4.6.12 RDMA over Converged Ethernet

> Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards.

4.6.12.1 IP Routable (RoCEv2)

RoCE has two addressing modes: MAC based GIDs, and IP address based GIDs. In RoCE IP based, if the IP address changes while the system is running, the GID for the port will automatically be updated with the new IP address, using either IPv4 or IPv6.

RoCE IP based allows RoCE traffic between Windows and Linux systems, which use IP based GIDs by default.

A straightforward extension of the RoCE protocol enables traffic to operate in layer 3 environments. This capability is obtained via a simple modification of the RoCE packet format. Instead of the GRH used in RoCE, routable RoCE packets carry an IP header which allows traversal of IP L3 Routers and a UDP header that serves as a stateless encapsulation layer for the RDMA Transport Protocol Packets over IP.

Figure 8: RoCE and RoCE v2 Frame Format Differences
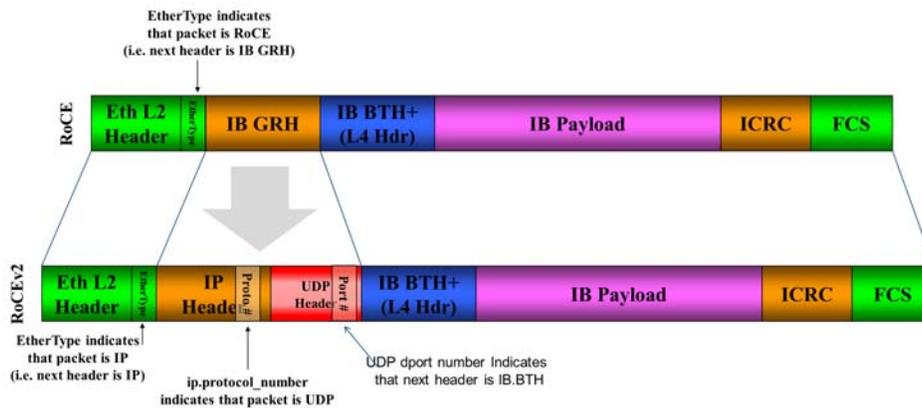
The proposed RoCEv2 packets use a well-known UDP destination port value that unequivocally distinguishes the datagram. Similar to other protocols that use UDP encapsulation, the UDP source port field is used to carry an opaque flow-identifier that allows network devices to implement packet forwarding optimizations (e.g. ECMP) while staying agnostic to the specifics of the protocol header format.

The UDP source port is calculated as follows: `UDP.SrcPort = (SrcPort XOR DstPort) OR 0xC000`, where SrcPort and DstPort are the ports used to establish the connection.
For example, in a Network Direct application, when connecting to a remote peer, the destination IP address and the destination port must be provided as they are used in the calculation above. The source port provision is optional.

Furthermore, since this change exclusively affects the packet format on the wire, and due to the fact that with RDMA semantics packets are generated and consumed below the AP applications can seamlessly operate over any form of RDMA service (including the routable version of RoCE as shown in Figure 8,"RoCE and RoCE v2 Frame Format Differences" on page 112), in a completely transparent way[1].

---

1. Standard RDMA APIs are IP based already for all existing RDMA technologies

Figure 9: RoCE and RoCEv2 Protocol Stack
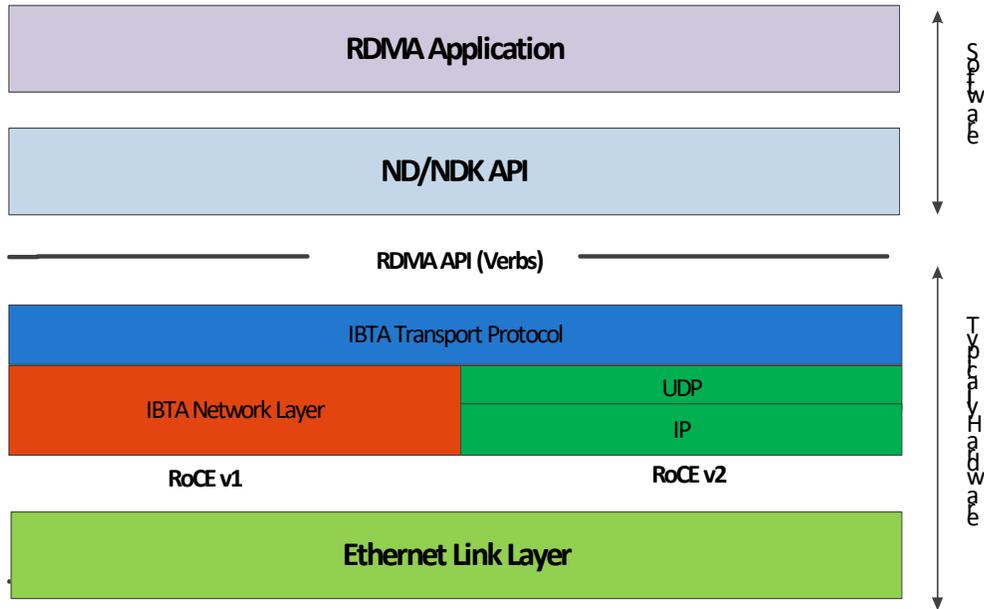
> The fabric must use the same protocol stack in order for nodes to communicate.

> For ConnectX-3 ConnectX-3 Pro adapter cards:
> The default RoCE mode in Windows is MAC based.
> The default RoCE mode in Linux is IP based.
> In order to communicate between Windows and Linux over RoCE, please change the RoCE mode in Windows to IP based.

> For to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards. adapter cards:
> In earlier versions, the default value of RoCE mode was RoCE v1. Starting from v1.30, the default value of RoCE mode will be RoCEv2.
> Upgrading from earlier versions to version 1.30 or above will save the old default value (RoCEv1).

4.6.12.2 RoCE Configuration

In order to function reliably, RoCE requires a form of flow control. While it is possible to use global flow control, this is normally undesirable, for performance reasons.

The normal and optimal way to use RoCE is to use Priority Flow Control (PFC). To use PFC, it must be enabled on all endpoints and switches in the flow path.

The following section presents instructions to configure PFC on Mellanox ConnectX™ cards. There are multiple configuration steps required, all of which may be performed via PowerShell. Therefore, although we present each step individually, you may ultimately choose to write a

PowerShell script to do them all in one step. Note that administrator privileges are required for these steps.

4.6.12.2.1 System Requirements

The following are the driver's prerequisites in order to set or configure RoCE:

- RoCE: ConnectX®-3 and ConnectX®-3 Pro firmware version 2.30.3000 or higher.

- RoCEv2: ConnectX®-3 Pro firmware version 2.31.5050 or higher.

- RoCE and RoCEv2 are supported on all ConnectX-4 ConnectX-4 Lx and ConnectX-5 Ex firmware versions.

- Operating Systems:

  Windows Server 2008 R2, Windows Server 2012, Windows Server 2012 R2, and Windows Server 2016.

- Set NIC to use Ethernet protocol:

  Display the Device Manager and expand "System Devices".

4.6.12.2.2 Configuring Windows Host

Configuring Windows host requires configuring QoS. To configure QoS, please follow the procedure described in Section 4.6.2, "Configuring Quality of Service (QoS)", on page 97

> Since PFC is responsible for flow controlling at the granularity of traffic priority, it is necessary to assign different priorities to different types of network traffic.
> As per RoCE configuration, all ND/NDK traffic is assigned to one or more chosen priorities, where PFC is enabled on those priorities.

4.6.12.2.2.1Global Pause (Flow Control)

➢ *To use Global Pause (Flow Control) mode, disable QoS and Priority:*

```
PS $ Disable-NetQosFlowControl
PS $ Disable-NetAdapterQos <interface name>
```

➢ *To confirm flow control is enabled in adapter parameters:*

Device Manager -> Network Adapters -> Mellanox ConnectX-4/ConnectX-4 Lx/ConnectX-5 Ex Ethernet Adapter -> Properties -> Advanced tab
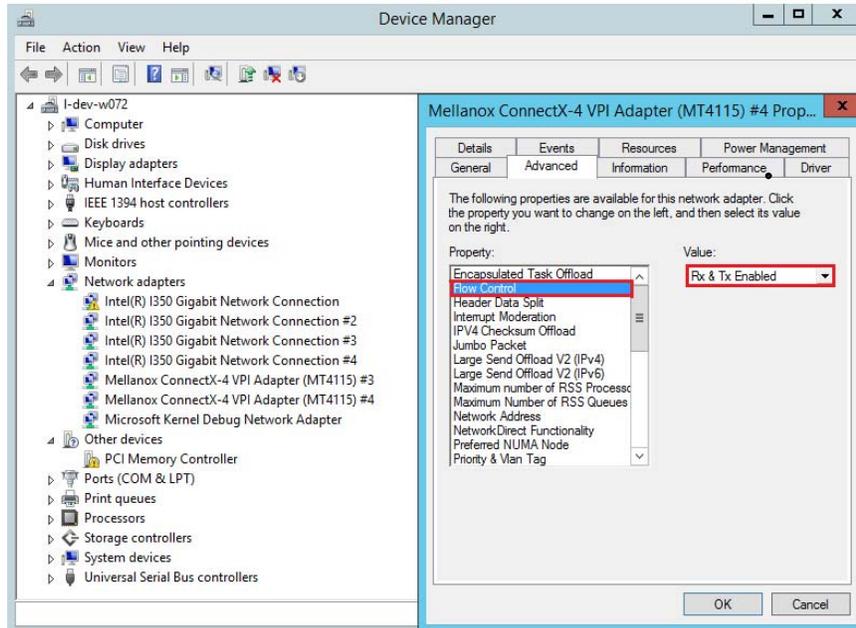
4.6.12.2.3Using Global Pause (Flow Control)

> Applies to ConnectX-3 and ConnectX-3 Pro adapter cards.

➢ *To enable Global Pause on ports that face the hosts, perform the following:*

```
(config)# interface et10
(config-if-Et10)# flowcontrol receive on
(config-if-Et10)# flowcontrol send on
```

4.6.12.3 Configuring Router (PFC only)

The router uses L3's DSCP value to mark the egress traffic of L2 PCP. The required mapping, maps the three most significant bits of the DSCP into the PCP. This is the default behavior, and no additional configuration is required.

4.6.12.3.1Copying Port Control Protocol (PCP) between Subnets

The captured PCP option from the Ethernet header of the incoming packet can be used to set the PCP bits on the outgoing Ethernet header.

4.6.13 Teaming and VLAN

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex

Windows Server 2012 and above supports Teaming as part of the operating system. Please refer to Microsoft guide "NIC Teaming in Windows Server 2012" following the link below:

http://www.microsoft.com/en-us/download/confirmation.aspx?id=40319

For other earlier operating systems, please refer to the sections below. Note that the Microsoft teaming mechanism is only available on Windows Server distributions.

4.6.13.1 Configuring a Network Interface to Work with VLAN in Windows Server 2012 and Above

In this procedure you DO NOT create a VLAN, rather use an existing VLAN ID.

➢ *To configure a port to work with VLAN using the Device Manager.*

Step 1. Open the Device Manager.

Step 2. Go to the Network adapters.

Step 3. Go to the properties of Mellanox ConnectX®-4 Ethernet Adapter card.

Step 4. Go to the Advanced tab.

Step 5. Choose the VLAN ID in the Property window.

Step 6. Set its value in the Value window.
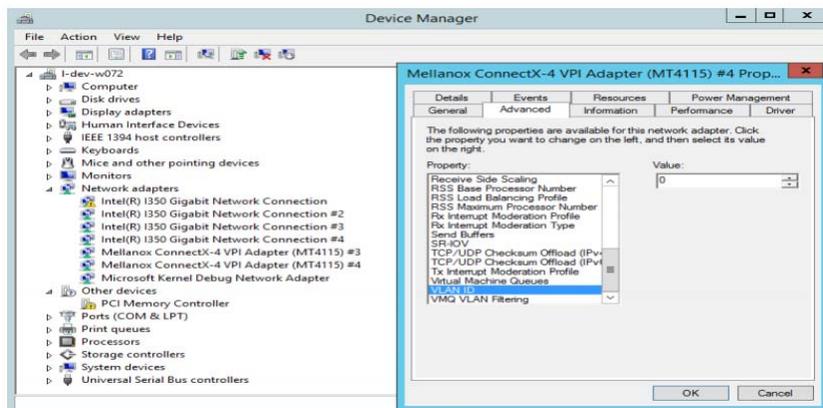
4.6.14 Deploying SMB Direct

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex

The Server Message Block (SMB) protocol is a network file sharing protocol implemented in Microsoft Windows. The set of message packets that defines a particular version of the protocol is called a dialect.

The Microsoft SMB protocol is a client-server implementation and consists of a set of data packets, each containing a request sent by the client or a response sent by the server.

SMB protocol is used on top of the TCP/IP protocol or other network protocols. Using the SMB protocol allows applications to access files or other resources on a remote server, to read, create, and update them. In addition, it enables communication with any server program that is set up to receive an SMB client request.

4.6.14.1 System Requirements

The following are hardware and software prerequisites:

- Two or more machines running Windows Server 2012 and above
- One or more Mellanox ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex network adapters for each server
- Mellanox ConnectX-3/ConnectX-3 Pro Adapter with firmware v2.30.8000 or higher
- Mellanox ConnectX-4 Adapter with firmware v2.17.20.52 or higher
- Mellanox ConnectX-4 Lx Adapter with firmware v14.14.2320 or higher
- Mellanox ConnectX-5 Ex Adapter with firmware v16.23.10.20 or higher

4.6.14.2 SMB Configuration Verification - ConnectX-3 and ConnectX-3 Pro

Applies to ConnectX-3 and ConnectX-3 Pro adapter cards.

4.6.14.2.1 Verifying SMB Configuration

Use the following PowerShell cmdlets to verify SMB Multichannel is enabled, confirm the adapters are recognized by SMB and that their RDMA capability is properly identified.

- On the SMB client, run the following PowerShell cmdlets:

```
Get-SmbClientConfiguration | Select EnableMultichannel
Get-SmbClientNetworkInterface
```

- On the SMB server, run the following PowerShell cmdlets:

```
Get-SmbServerConfiguration | Select EnableMultichannel
Get-SmbServerNetworkInterface
netstat.exe -xan | ? {$_ -match "445"}
```

4.6.14.3 Verifying SMB Connection

> Applies to ConnectX-3 and ConnectX-3 Pro adapter cards.

➢ *To verify the SMB connection on the SMB client:*

Step 1. Copy the large file to create a new session with the SMB Server.

Step 2. Open a PowerShell window while the copy is ongoing.

Step 3. Verify the SMB Direct is working properly and that the correct SMB dialect is used.

```
Get-SmbConnection
Get-SmbMultichannelConnection
netstat.exe -xan | ? {$_ -match "445"}
```

> If there is no activity while running the commands above, you might get an empty list due to session expiration and no current connections.

4.6.14.4 Verifying SMB Events that Confirm RDMA Connection

> Applies to ConnectX-3 and ConnectX-3 Pro adapter cards.

➢ *To confirm RDMA connection, verify the SMB events:*

Step 1. Open a PowerShell window on the SMB client.

Step 2. Run the following cmdlets.
NOTE: Any RDMA-related connection errors will be displayed as well.

```
PS $ Get-WinEvent -LogName Microsoft-Windows-SMBClient/Operational | ? Message -match "RDMA"
```

4.6.14.5 SMB Configuration Verification - ConnectX-4 and ConnectX-4 Lx

Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex

4.6.14.5.1 Verifying Network Adapter Configuration

Use the following PowerShell cmdlets to verify Network Direct is globally enabled and that you have NICs with the RDMA capability.

- Run on both the SMB server and the SMB client.

```
PS $ Get-NetOffloadGlobalSetting | Select NetworkDirect
PS $ Get-NetAdapterRDMA
PS $ Get-NetAdapterHardwareInfo
```

4.6.14.5.2 Verifying SMB Configuration

Use the following PowerShell cmdlets to verify SMB Multichannel is enabled, confirm the adapters are recognized by SMB and that their RDMA capability is properly identified.

- On the SMB client, run the following PowerShell cmdlets:

```
PS $ Get-SmbClientConfiguration | Select EnableMultichannel
PS $ Get-SmbClientNetworkInterface
```

- On the SMB server, run the following PowerShell cmdlets[1]:

```
PS $ Get-SmbServerConfiguration | Select EnableMultichannel
PS $ Get-SmbServerNetworkInterface
PS $ netstat.exe -xan | ? {$_ -match "445"}
```

---

1. The NETSTAT command confirms if the File Server is listening on the RDMA interfaces.

4.6.15 Network Virtualization using Generic Routing Encapsulation (NVGRE)

> Available in ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex.

> Network Virtualization using Generic Routing Encapsulation (NVGRE) offload is currently supported in Windows Server 2012 R2 with the latest updates for Microsoft.

4.6.15.1 System Requirements

- Operating Systems: Windows Server 2012 R2 and above
- Mellanox ConnectX®-3 Pro Adapter with firmware v2.30.8000 or higher
- Mellanox ConnectX®-4 Adapter with firmware v12.17.20.52 or higher
- Mellanox ConnectX®-4 Lx Adapter with firmware v14.17.20.52 or higher
- Mellanox ConnectX®-5 Ex Adapter with firmware v16.23.10.20 or higher

4.6.15.2 Using NVGRE

Network Virtualization using Generic Routing Encapsulation (NVGRE) is a network virtualization technology that attempts to alleviate the scalability problems associated with large cloud computing deployments. It uses Generic Routing Encapsulation (GRE) to tunnel layer 2 packets across an IP fabric, and uses 24 bits of the GRE key as a logical network discriminator (which is called a tenant network ID).

Configuring the Hyper-V Network Virtualization requires two types of IP addresses:

- **Provider Addresses (PA)** - unique IP addresses assigned to each Hyper-V host that are routable across the physical network infrastructure. Each Hyper-V host requires at least one PA to be assigned.

- **Customer Addresses (CA)** - unique IP addresses assigned to each Virtual Machine that participate on a virtualized network. Using NVGRE, multiple CAs for VMs running on a Hyper-V host can be tunneled using a single PA on that Hyper-V host. CAs must be unique across all VMs on the same virtual network, but they do not need to be unique across virtual networks with different Virtual Subnet ID.

The VM generates a packet with the addresses of the sender and the recipient within the CA space. Then Hyper-V host encapsulates the packet with the addresses of the sender and the recipient in PA space.

PA addresses are determined by using virtualization table. Hyper-V host retrieves the received packet, identifies recipient and forwards the original packet with the CA addresses to the desired VM.

NVGRE can be implemented across an existing physical IP network without requiring changes to physical network switch architecture. Since NVGRE tunnels terminate at each Hyper-V host, the hosts handle all encapsulation and de-encapsulation of the network traffic. Firewalls that block GRE tunnels between sites have to be configured to support forwarding GRE (IP Protocol 47) tunnel traffic.

Figure 10: NVGRE Packet Structure

Outer MAC	Outer IP Header (PA)	GRE Header (Includes 24 Bit TNI)	Inner MAC	Inner IP Header (CA)	TCP Header	TCP user data

4.6.15.3 Enabling/Disabling NVGRE Offloading

To leverage NVGRE to virtualize heavy network IO workloads, the Mellanox ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex network NIC provides hardware support for GRE offload within the network NICs by default.

➢ *To enable/disable NVGRE offloading:*

Step 1. Open the Device Manager.

Step 2. Go to the Network adapters.

Step 3. Right click 'Properties' on Mellanox Ethernet Adapter card.

Step 4. Go to Advanced tab.

Step 5. Choose the '`Encapsulate Task Offload`' option.

Step 6. Set one of the following values:

- Enable - GRE offloading is Enabled by default
- Disabled - When disabled the Hyper-V host will still be able to transfer NVGRE traffic, but TCP and inner IP checksums will be calculated by software which significantly reduces performance.

4.6.15.3.1 Configuring the NVGRE using PowerShell

Hyper-V Network Virtualization policies can be centrally configured using PowerShell 3.0 and PowerShell Remoting.

Step 1. **[Windows Server 2012 Only]** Enable the Windows Network Virtualization binding on the physical NIC of each Hyper-V Host (Host 1 and Host 2).

```
PS $ Enable-NetAdapterBinding <EthInterfaceName>(a)-ComponentID ms_netwnv
```

<EthInterfaceName> - Physical NIC name

Step 2. Create a vSwitch.

```
PS $ New-VMSwitch <vSwitchName> -NetAdapterName <EthInterfaceName>-AllowManagementOS $true
```

Step 3. Shut down the VMs.

```
PS $ Stop-VM -Name <VM Name> -Force -Confirm
```

Step 4. Configure the Virtual Subnet ID on the Hyper-V Network Switch Ports for each Virtual Machine on each Hyper-V Host (Host 1 and Host 2).

```
PS $ Add-VMNetworkAdapter -VMName <VMName> -SwitchName <vSwitchName> -StaticMacAddress
<StaticMAC Address>
```

Step 5. Configure a Subnet Locator and Route records on all Hyper-V Hosts (same command on all Hyper-V hosts). Add customer route on all Hyper-V hosts (same command on all Hyper V host.

```
PS $ New-NetVirtualizationLookupRecord -CustomerAddress <VMInterfaceIPAddress 1/n> -
ProviderAddress <HypervisorInterfaceIPAddress1> -VirtualSubnetID <virtualsubnetID> -
MACAddress <VMmacaddress1>ª -Rule "TranslationMethodEncap"

PS $ New-NetVirtualizationLookupRecord -CustomerAddress <VMInterfaceIPAddress 2/n> -
ProviderAddress <HypervisorInterfaceIPAddress2> -VirtualSubnetID <virtualsubnetID> -
MACAddress <VMmacaddress2>ª -Rule "TranslationMethodEncap"
```

    a. This is the VM's MAC address associated with the vSwitch connected to the Mellanox device.

Step 6. Add customer route on all Hyper-V hosts (same command on all Hyper-V hosts).

```
PS $ New-NetVirtualizationCustomerRoute -RoutingDomainID "{11111111-2222-3333-4444-
000000005001}" -VirtualSubnetID <virtualsubnetID> -DestinationPrefix <VMInterfaceIPAd-
dress/Mask> -NextHop "0.0.0.0" -Metric 255
```

Step 7. Configure the Provider Address and Route records on each Hyper-V Host using an appropriate interface name and IP address.

```
PS $ $NIC = Get-NetAdapter <EthInterfaceName>
PS $ New-NetVirtualizationProviderAddress -InterfaceIndex $NIC.InterfaceIndex -Provid-
erAddress <HypervisorInterfaceIPAddress> -PrefixLength 24
```

```
PS $ New-NetVirtualizationProviderRoute -InterfaceIndex $NIC.InterfaceIndex -Destina-
tionPrefix "0.0.0.0/0" -NextHop <HypervisorInterfaceIPAddress>
```

Step 8. Configure the Virtual Subnet ID on the Hyper-V Network Switch Ports for each Virtual Machine on each Hyper-V Host (Host 1 and Host 2).

```
PS $ Get-VMNetworkAdapter -VMName <VMName> | where {$_.MacAddress -eq <VMmacaddress1>}
| Set-VMNetworkAdapter -VirtualSubnetID <virtualsubnetID>
```

Please repeat steps 5 to 8 on each Hyper-V after rebooting the Hypervisor.

4.6.15.4 Verifying the Encapsulation of the Traffic

Once the configuration using PowerShell is completed, verifying that packets are indeed encapsulated as configured is possible through any packet capturing utility. If configured correctly, an encapsulated packet should appear as a packet consisting of the following headers:

Outer ETH Header, Outer IP, GRE Header, Inner ETH Header, Original Ethernet Payload.

4.6.15.5 Removing NVGRE configuration

Step 1. Set VSID back to 0 (on each Hyper-V for each Virtual Machine where VSID was set)

```
PS $ Get-VMNetworkAdapter <VMName>(a) | where {$_.MacAddress -eq <VMMacAddress>(b)} |
Set-VMNetworkAdapter -VirtualSubnetID 0
```

- VMName - the name of Virtual machine
- VMMacAddress - the MAC address of VM's network interface associated with vSwitch that was connected to Mellanox device.

Step 2. Remove all lookup records (same command on all Hyper-V hosts).

```
PS $ Remove-NetVirtualizationLookupRecord
```

Step 3. Remove customer route (same command on all Hyper-V hosts).

```
PS $ Remove-NetVirtualizationCustomerRoute
```

Step 4. Remove Provider address (same command on all Hyper-V hosts).

```
PS $ Remove-NetVirtualizationProviderAddress
```

Step 5. Remove provider routed for a Hyper-V host.

```
PS $ Remove-NetVirtualizationProviderRoute
```

Step 6. For HyperV running Windows Server 2012 only disable network adapter binding to ms_-netwnv service

```
PS $ Disable-NetAdapterBinding <EthInterfaceName>(a) -ComponentID ms_netwnv
<EthInterfaceName> - Physical NIC name
```

4.6.16 Performance Tuning and Counters

Applies to ConnectX-4 and ConnectX-4 Lx adapter cards.

For further information on WinOF-2 performance, please refer to the Performance Tuning Guide for Mellanox Network Adapters.

This section describes how to modify Windows registry parameters in order to improve performance.

> Please note that modifying the registry incorrectly might lead to serious problems, including the loss of data, system hang, and you may need to reinstall Windows. As such it is recommended to back up the registry on your system before implementing recommendations included in this section. If the modifications you apply lead to serious problems, you will be able to restore the original registry state. For more details about backing up and restoring the registry, please visit www.microsoft.com.

4.6.16.1 General Performance Optimization and Tuning

To achieve the best performance for Windows, you may need to modify some of the Windows registries.

4.6.16.1.1 Registry Tuning

The registry entries that may be added/changed by this "General Tuning" procedure are:

Under HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters:

• Disable TCP selective acks option for better cpu utilization:

```
SackOpts, type REG_DWORD, value set to 0.
```

Under HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\AFD\Parameters:

• Enable fast datagram sending for UDP traffic:

```
FastSendDatagramThreshold, type REG_DWORD, value set to 64K.
```

Under HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Ndis\Parameters:

• Set RSS parameters:

```
RssBaseCpu, type REG_DWORD, value set to 1.
```

4.6.16.1.2 Enable RSS

Enabling Receive Side Scaling (RSS) is performed by means of the following command:

```
"netsh int tcp set global rss = enabled"
```

4.6.16.1.3 Improving Live Migration

In order to improve live migration over SMB direct performance, please set the following registry key to 0 and reboot the machine:

```
HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\LanmanServer\Parameters\RequireSecuritySig-
nature
```

4.6.16.2 Application Specific Optimization and Tuning

4.6.16.2.1 Ethernet Performance Tuning

The user can configure the Ethernet adapter by setting some registry keys. The registry keys may affect Ethernet performance.

➢ *To improve performance, activate the performance tuning tool as follows:*

Step 1. Start the "Device Manager" (open a command line window and enter: devmgmt.msc).

Step 2. Open "Network Adapters".

Step 3. Right click the relevant Ethernet adapter and select Properties.

Step 4. Select the "Advanced" tab

Step 5. Modify performance parameters (properties) as desired.

4.6.16.2.1.1 Performance Known Issues

- On Intel I/OAT supported systems, it is highly recommended to install and enable the latest I/OAT driver (download from www.intel.com).

- With I/OAT enabled, sending 256-byte messages or larger will activate I/OAT. This will cause a significant latency increase due to I/OAT algorithms. On the other hand, throughput will increase significantly when using I/OAT.

4.6.16.3 Ethernet Bandwidth Improvements

➢ *To improve Ethernet Bandwidth:*

Step 1. Check you are running on the closest NUMA.

Step a. In the PowerShell run: `Get-NetAdapterRss -Name "adapter name"`

Step b. Validate that the `IndirectionTable` CPUs are located at the closest NUMA.
As illustrated in the figure above, the CPUs are 0:0 - 0:7, CPU 0 -7 and the distance from the NUMA is 0, 0:0/0 - 0:7/0, unlike CPU 14-27/32767.

Step c.   If the CPUs are not close to the NUMEA, change the "`RSS Base Processor Number`" and "`RSS Max Processor Number`" settings under the Advance tab to point to the closest CPUs.

> For high performance, it is recommended to work with at least 8 processors.

Step 2.   Check the Ethernet bandwidth, run ntttcp.exe.

- **Server side:** ntttcp -r -m 32,*,server_ip
- **Client side:** ntttcp -s -m 32,*,server_ip

4.6.16.4 Tunable Performance Parameters

The following is a list of key parameters for performance tuning.

- **Jumbo Packet**

  The maximum available size of the transfer unit, also known as the Maximum Transmission Unit (MTU). The MTU of a network can have a substantial impact on performance. A 4K MTU size improves performance for short messages, since it allows the OS to coalesce many small messages into a large one.

  - Valid MTU values range for an Ethernet driver is between 614 and 9614.

    > All devices on the same physical network, or on the same logical network, must have the same MTU.

- **Receive Buffers**

  The number of receive buffers (default 512).

- **Send Buffers**

  The number of sent buffers (default 2048).

- **Performance Options**

  Configures parameters that can improve adapter performance.

  - Interrupt Moderation

    Moderates or delays the interrupts' generation. Hence, optimizes network throughput and CPU utilization (default Enabled).

    - When the interrupt moderation is enabled, the system accumulates interrupts and sends a single interrupt rather than a series of interrupts. An interrupt is generated after receiving 5 packets or after 10ms from the first packet received. It improves performance and reduces CPU load however, it increases latency.

    - When the interrupt moderation is disabled, the system generates an interrupt each time a packet is received or sent. In this mode, the CPU utilization data rates increase, as the system handles a larger number of interrupts. However, the latency decreases as the packet is handled faster.

  - Receive Side Scaling (RSS Mode)

    Improves incoming packet processing performance. RSS enables the adapter port to utilize the multiple CPUs in a multi-core system for receiving incoming packets and steering them to the designated destination. RSS can significantly improve the number of transactions, the number of connections per second, and the network throughput.

    This parameter can be set to one of the following values:

    - Enabled (default): Set RSS Mode

• Disabled: The hardware is configured once to use the Toeplitz hash function, and the indirection table is never changed.

> IOAT is not used while in RSS mode.

• Receive Completion Method

Sets the completion methods of the received packets, and can affect network throughput and CPU utilization.

- • **Polling Method**

  Increases the CPU utilization as the system polls the received rings for the incoming packets. However, it may increase the network performance as the incoming packet is handled faster.

- • **Adaptive (Default Settings)**

  A combination of the interrupt and polling methods dynamically, depending on traffic type and network usage. Choosing a different setting may improve network and/or system performance in certain configurations.

• Rx Interrupt Moderation Type

Sets the rate at which the controller moderates or delays the generation of interrupts making it possible to optimize network throughput and CPU utilization. The default setting (Adaptive) adjusts the interrupt rates dynamically depending on the traffic type and network usage. Choosing a different setting may improve network and system performance in certain configurations.

• Send Completion Method

Sets the completion methods of the Send packets and it may affect network throughput and CPU utilization.

• **Offload Options**

Allows you to specify which TCP/IP offload settings are handled by the adapter rather than the operating system.

Enabling offloading services increases transmission performance as the offload tasks are performed by the adapter hardware rather than the operating system. Thus, freeing CPU resources to work on other tasks.

• IPv4 Checksums Offload

Enables the adapter to compute IPv4 checksum upon transmit and/or receive instead of the CPU (default Enabled).

• TCP/UDP Checksum Offload for IPv4 packets

Enables the adapter to compute TCP/UDP checksum over IPv4 packets upon transmit and/or receive instead of the CPU (default Enabled).

• TCP/UDP Checksum Offload for IPv6 packets

Enables the adapter to compute TCP/UDP checksum over IPv6 packets upon transmit and/or receive instead of the CPU (default Enabled).

• Large Send Offload (LSO)

Allows the TCP stack to build a TCP message up to 64KB long and sends it in one call down the stack. The adapter then re-segments the message into multiple TCP packets for transmission on the wire with each pack sized according to the MTU. This option offloads a large amount of kernel processing time from the host CPU to the adapter.

4.6.16.5 Adapter Proprietary Performance Counters

Proprietary Performance Counters are used to provide information on Operating System, application, service or the drivers' performance. Counters can be used for different system debugging purposes, help to determine system bottlenecks and fine-tune system and application performance. The Operating System, network, and devices provide counter data that the application can consume to provide users with a graphical view of the system's performance quality. WinOF-2 counters hold the standard Windows CounterSet API that includes:

• Network Interface

• RDMA activity

• SMB Direct Connection

4.6.16.5.1 Supported Standard Performance Counters

4.6.16.5.1.1 Proprietary Mellanox WinOF-2 Port Traffic Counters

Proprietary Mellanox WinOF-2 port traffic counters set consists of global traffic statistics which gather information from ConnectX®-4 and ConnectX®-4 Lx network adapters, and includes traffic statistics, and various types of error and indications from both the Physical Function and Virtual Function.

Table 19 - Mellanox WinOF-2 Port Traffic Counters

Mellanox Adapter Traffic Counters	Description
Bytes IN	
Bytes Received	Shows the number of bytes received by the adapter. The counted bytes include framing characters.
Bytes Received/Sec	Shows the rate at which bytes are received by the adapter. The counted bytes include framing characters.
Packets Received	Shows the number of packets received by the network interface.
Packets Received/Sec	Shows the rate at which packets are received by the network interface.
Bytes/ Packets OUT	
Bytes Sent	Shows the number of bytes sent by the adapter. The counted bytes include framing characters.
Bytes Sent/Sec	Shows the rate at which bytes are sent by the adapter. The counted bytes include framing characters.
Packets Sent	Shows the number of packets sent by the network interface.
Packets Sent/Sec	Shows the rate at which packets are sent by the network interface.
Bytes' TOTAL	

Table 19 - Mellanox WinOF-2 Port Traffic Counters

Mellanox Adapter Traffic Counters	Description
Bytes Total	Shows the total of bytes handled by the adapter. The counted bytes include framing characters.
Bytes Total/Sec	Shows the total rate of bytes that are sent and received by the adapter. The counted bytes include framing characters.
Packets Total	Shows the total of packets handled by the network interface.
Packets Total/Sec	Shows the rate at which packets are sent and received by the network interface.
Control Packets	The total number of successfully received control frames.[a]
ERRORS, DROP, AND MISC. INDICATIONS	
Packets Outbound Errors[b]	Shows the number of outbound packets that could not be transmitted because of errors found in the physical layer.[a]
Packets Outbound Discarded[a]	Shows the number of outbound packets to be discarded in the physical layer, even though no errors had been detected to prevent transmission. One possible reason for discarding packets could be to free up some buffer space.
Packets Received Errors[a]	Shows the number of inbound packets that contained errors in the physical layer, preventing them from being deliverable.
Packets Received with Frame Length Error	Shows the number of inbound packets that contained error where the frame has length error. Packets received with frame length error are a subset of packets received errors.[a]
Packets Received with Symbol Error	Shows the number of inbound packets that contained symbol error or an invalid block. Packets received with symbol error are a subset of packets received errors.
Packets Received with Bad CRC Error	Shows the number of inbound packets that failed the CRC check. Packets received with bad CRC error are a subset of packets received errors.
Packets Received Discarded	Shows the number of inbound packets that were chosen to be discarded in the physical layer, even though no errors had been detected to prevent their being deliverable. One possible reason for discarding such a packet could be a buffer overflow.
Receive Segment Coalescing (RSC)	
RSC Aborts	Number of RSC abort events. That is, the number of exceptions other than the IP datagram length being exceeded. This includes the cases where a packet is not coalesced because of insufficient hardware resources.[a]
RSC Coalesced Events	Number of RSC coalesced events. That is, the total number of packets that were formed from coalescing packets.[a]
RSC Coalesced Octets	Number of RSC coalesced bytes.[a]
RSC Coalesced Packets	Number of RSC coalesced packets.[a]
RSC Average Packet Size	RSC Average Packet Size is the average size in bytes of received packets across all TCP connections.[a]

a. This counter is relevant only for ETH ports.
b. Those error/discard counters are related to layer-2 issues, such as CRC, length, and type errors. There is a possibility of an error/discard in the higher interface level. For example, a packet can be discarded for the lack of a receive buffer. To see the sum of all error/discard packets, read the Windows Network-Interface Counters. Note that for IPoIB, the Mellanox counters are for IB layer-2 issues only, and Windows Network-Interface counters are for interface level issues.

4.6.16.5.1.2Mellanox WinOF-2 Virtual Function (VF) Port Traffic Counters

Mellanox WinOF-2 VF Port Traffic set consists of counters that measure the rates at which bytes and packets are sent and received over a virtual port network connection that is bound to a virtual PCI function. It includes counters that monitor connection errors.

This set is available only on hypervisors and not on virtual network adapters.

Table 20 - Mellanox WinOF-2 Virtual Function (VF) Port Traffic Counters

Mellanox WinOF-2 VF Port Traffic Counters	Description
Bytes/Packets IN	
Bytes Received/Sec	Shows the rate at which bytes are received over each network VPort. The counted bytes include framing characters.
Bytes Received Unicast/Sec	Shows the rate at which subnet-unicast bytes are delivered to a higher-layer protocol.
Bytes Received Broadcast/Sec	Shows the rate at which subnet-broadcast bytes are delivered to a higher-layer protocol.
Bytes Received Multicast/Sec	Shows the rate at which subnet-multicast bytes are delivered to a higher-layer protocol.
Packets Received Unicast/Sec	Shows the rate at which subnet-unicast packets are delivered to a higher-layer protocol.
Packets Received Broadcast/Sec	Shows the rate at which subnet-broadcast packets are delivered to a higher-layer protocol.
Packets Received Multicast/Sec	Shows the rate at which subnet-multicast packets are delivered to a higher-layer protocol.
Bytes/Packets IN	
Bytes Sent/Sec	Shows the rate at which bytes are sent over each network VPort. The counted bytes include framing characters.
Bytes Sent Unicast/Sec	Shows the rate at which bytes are requested to be transmitted to subnet-unicast addresses by higher-level protocols. The rate includes the bytes that were discarded or not sent.
Bytes Sent Broadcast/Sec	Shows the rate at which bytes are requested to be transmitted to subnet-broadcast addresses by higher-level protocols. The rate includes the bytes that were discarded or not sent.
Bytes Sent Multicast/Sec	Shows the rate at which bytes are requested to be transmitted to subnet-multicast addresses by higher-level protocols. The rate includes the bytes that were discarded or not sent.
Packets Sent Unicast/Sec	Shows the rate at which packets are requested to be transmitted to subnet-unicast addresses by higher-level protocols. The rate includes the packets that were discarded or not sent.
Packets Sent Broadcast/Sec	Shows the rate at which packets are requested to be transmitted to subnet-broadcast addresses by higher-level protocols. The rate includes the packets that were discarded or not sent.
Packets Sent Multicast/Sec	Shows the rate at which packets are requested to be transmitted to subnet-multicast addresses by higher-level protocols. The rate includes the packets that were discarded or not sent.

Table 20 - Mellanox WinOF-2 Virtual Function (VF) Port Traffic Counters

Mellanox WinOF-2 VF Port Traffic Counters	Description
ERRORS, DISCARDED	
Packets Outbound Discarded	Shows the number of outbound packets to be discarded even though no errors had been detected to prevent transmission. One possible reason for discarding a packet could be to free up buffer space.
Packets Outbound Errors	Shows the number of outbound packets that could not be transmitted because of errors.
Packets Received Discarded	Shows the number of inbound packets that were chosen to be discarded even though no errors had been detected to prevent their being deliverable to a higher-layer protocol. One possible reason for discarding such a packet could be to free up buffer space.
Packets Received Errors	Shows the number of inbound packets that contained errors preventing them from being deliverable to a higher-layer protocol.

4.6.16.5.1.3 Proprietary Mellanox WinOF-2 Port QoS Counters

Proprietary Mellanox WinOF-2 Port QoS counters set consists of flow statistics per (VLAN) priority. Each QoS policy is associated with a priority. The counter presents the priority's traffic, pause statistic.

Table 21 - Mellanox WinOF-2 Port QoS Counters

Mellanox Qos Counters	Description
Bytes/Packets IN	
Bytes Received	The number of bytes received that are covered by this priority. The counted bytes include framing characters (modulo 2^{64}).
Bytes Received/Sec	The number of bytes received per second that are covered by this priority. The counted bytes include framing characters.
Packets Received	The number of packets received that are covered by this priority (modulo 2^{64}).
Packets Received/Sec	The number of packets received per second that are covered by this priority.
Bytes/Packets OUT	
Bytes Sent	The number of bytes sent that are covered by this priority. The counted bytes include framing characters (modulo 2^{64}).
Bytes Sent/Sec	The number of bytes sent per second that are covered by this priority. The counted bytes include framing characters.
Packets Sent	The number of packets sent that are covered by this priority (modulo 2^{64}).
Packets Sent/Sec	The number of packets sent per second that are covered by this priority.
Bytes and Packets Total	

Table 21 - Mellanox WinOF-2 Port QoS Counters

Mellanox Qos Counters	Description
Bytes Total	The total number of bytes that are covered by this priority. The counted bytes include framing characters (modulo 2^64).
Bytes Total/Sec	The total number of bytes per second that are covered by this priority. The counted bytes include framing characters.
Packets Total	The total number of packets that are covered by this priority (modulo 2^64).
Packets Total/Sec	The total number of packets per second that are covered by this priority.
PAUSE INDICATION	
Sent Pause Frames	The total number of pause frames sent from this priority to the far-end port. The untagged instance indicates the number of global pause frames that were sent.
Sent Pause Duration	The total duration of packets transmission being paused on this priority in microseconds.
Received Pause Frames	The number of pause frames that were received to this priority from the far-end port. The untagged instance indicates the number of global pause frames that were received.
Received Pause Duration	The total duration that far-end port was requested to pause for the transmission of packets in microseconds.
Sent Discard Frames	The number of packets discarded by the transmitter. **Note**: this counter is per TC and not per priority.

4.6.16.5.1.4RDMA Activity Counters

RDMA Activity counter set consists of NDK performance counters. These performance counters allow you to track Network Direct Kernel (RDMA) activity, including traffic rates, errors, and control plane activity.

Table 22 - RDMA Activity Counters

RDMA Activity Counters	Description
RDMA Accepted Connections	The number of inbound RDMA connections established.
RDMA Active Connections	The number of active RDMA connections.
RDMA Completion Queue Errors	This counter is not supported, and always is set to zero.
RDMA Connection Errors	The number of established connections with an error before a consumer disconnected the connection.
RDMA Failed Connection Attempts	The number of inbound and outbound RDMA connection attempts that failed.
RDMA Inbound Bytes/sec	The number of bytes for all incoming RDMA traffic. This includes additional layer two protocol overhead.
RDMA Inbound Frames/sec	The number, in frames, of layer two frames that carry incoming RDMA traffic.
RDMA Initiated Connections	The number of outbound connections established.

Table 22 - RDMA Activity Counters

RDMA Activity Counters	Description
RDMA Outbound Bytes/sec	The number of bytes for all outgoing RDMA traffic. This includes additional layer two protocol overhead.
RDMA Outbound Frames/sec	The number, in frames, of layer two frames that carry outgoing RDMA traffic.

4.6.16.5.1.5 Mellanox WinOF-2 Congestion Control Counters

Mellanox WinOF-2 Congestion Control counters set consists of counters that measure the DCQCN statistics over the network adapter.

Table 23 - Congestion Control Counters

Congestion Control Counters	Description
Notification Point	
Notification Point – CNPs Sent Successfully	Number of congestion notification packets (CNPs) successfully sent by the notification point.
Notification Point – RoCEv2 DCQCN Marked Packets	Number of RoCEv2 packets that were marked as congestion encountered.
Reaction Point	
Reaction Point – Current Number of Flows	Current number of Rate Limited Flows due to RoCEv2 Congestion Control.
Reaction Point – Ignored CNP Packets	Number of ignored congestion notification packets (CNPs).
Reaction Point – Successfully Handled CNP Packets	Number of congestion notification packets (CNPs) received and handled successfully.

4.6.16.5.1.6 Mellanox WinOF-2 Diagnostics Counters

Mellanox WinOF-2 diagnostics counters set consists of the following counters:

Table 24 - WinOF-2 Diagnostics Counters

Mellanox WinOF-2 Diagnostics Counters	Description
Reset Requests	Number of resets requested by NDIS.
Link State Change Events	Number of link status updates received from the hardware.
Queued Send Packets	Number of send packets pending transmission due to hardware queues overflow.
Send Completions in Passive/Sec	Number of send completion events handled in passive mode per second.
Receive Completions in Passive/Sec	Number of receive completion events handled in passive mode per second.
Copied Send Packets	Number of send packets that were copied in slow path.
Correct Checksum Packets In Slow Path	Number of receive packets that required the driver to perform the checksum calculation and resulted in success.
Bad Checksum Packets In Slow Path	Number of receive packets that required the driver to perform checksum calculation and resulted in failure.

Table 24 - WinOF-2 Diagnostics Counters

Mellanox WinOF-2 Diagnostics Counters	Description
Undetermined Checksum Packets In Slow Path	Number of receive packets with undetermined checksum result.
Watch Dog Expired/Sec	Number of watch dogs expired per second.
Requester time out received	Number of time out received when the local machine generates outbound traffic.
Requester out of order sequence NAK	Number of Out of Sequence NAK received when the local machine generates outbound traffic, i.e. the number of times the local machine received NAKs indicating OOS on the receiving side.
Requester RNR NAK	Number of RNR (Receiver Not Ready) NAKs received when the local machine generates outbound traffic.
Responder RNR NAK	Number of RNR (Receiver Not Ready) NAKs sent when the local machine receives inbound traffic.
Responder out of order sequence received	Number of Out of Sequence packets received when the local machine receives inbound traffic, i.e. the number of times the local machine received messages that are not consecutive.
Responder duplicate request received	Number of duplicate requests received when the local machine receives inbound traffic.
Requester RNR NAK retries exceeded errors	Number of RNR (Receiver Not Ready) NAKs retries exceeded errors when the local machine generates outbound traffic.
Responder Local Length Errors	Number of times the responder detected local length errors
Requester Local Length Errors	Number of times the requester detected local length errors
Responder Local QP Operation Errors	Number of times the responder detected local QP operation errors
Local Operation Errors	Number of local operation errors
Responder Local Protection Errors	Number of times the responder detected memory protection error in its local memory subsystem
Requester Local Protection Errors	Number of times the requester detected a memory protection error in its local memory subsystem
Responder CQEs with Error	Number of times the responder flow reported a completion with error
Requester CQEs with Error	Number of times the requester flow reported a completion with error
Responder CQEs Flushed with Error	Number of times the responder flow completed a work request as flushed with error
Requester CQEs Flushed with Error	Number of times the requester completed a work request as flushed with error
Requester Memory Window Binding Errors	Number of times the requester detected memory window binding error
Requester Bad Response	Number of times an unexpected transport layer opcode was returned by the responder
Requester Remote Invalid Request Errors	Number of times the requester detected remote invalid request error

Table 24 - WinOF-2 Diagnostics Counters

Mellanox WinOF-2 Diagnostics Counters	Description
Responder Remote Invalid Request Errors	Number of times the responder detected remote invalid request error
Requester Remote Access Errors	Number of times the requester detected remote access error
Responder Remote Access Errors	Number of times the responder detected remote access error
Requester Remote Operation Errors	Number of times the requester detected remote operation error
Requester Retry Exceeded Errors	Number of times the requester detected transport retries exceed error
CQ Overflow	Counts the QPs attached to a CQ with overflow condition
Received RDMA Write requests	Number of RDMA write requests received
Received RDMA Read requests	Number of RDMA read requests received
Implied NAK Sequence Errors	Number of times the Requester detected an ACK with a PSN larger than the expected PSN for an RDMA READ or ATOMIC response. The QP retry limit was not exceeded

4.6.16.5.1.7 Mellanox WinOF-2 Device Diagnostic Counters

Mellanox WinOF-2 device diagnostic counters set consists of the following counters:.

Table 25 - Device Diagnostics Counters

Mellanox WinOF-2 Device Diagnostic Counters	Description
L0 MTT miss	The number of access to L0 MTT that were missed
L0 MTT miss/Sec	The rate of access to L0 MTT that were missed
L0 MTT hit	The number of access to L0 MTT that were hit
L0 MTT hit/Sec	The rate of access to L0 MTT that were hit
L1 MTT miss	The number of access to L1 MTT that were missed
L1 MTT miss/Sec	The rate of access to L1 MTT that were missed
L1 MTT hit	The number of access to L1 MTT that were hit
L1 MTT hit/Sec	The rate of access to L1 MTT that were hit
L0 MPT miss	The number of access to L0 MKey that were missed
L0 MPT miss/Sec	The rate of access to L0 MKey that were missed
L0 MPT hit	The number of access to L0 MKey that were hit
L0 MPT hit/Sec	The rate of access to L0 MKey that were hit
L1 MPT miss	The number of access to L1 MKey that were missed
L1 MPT miss/Sec	The rate of access to L1 MKey that were missed
L1 MPT hit	The number of access to L1 MKey that were hit
L1 MPT hit/Sec	The rate of access to L1 MKey that were hit
RXS no slow path credis	No room in RXS for slow path packets
RXS no fast path credis	No room in RXS for fast path packets
RXT no slow path credis	No room in RXT for slow path packets

Table 25 - Device Diagnostics Counters

Mellanox WinOF-2 Device Diagnostic Counters	Description
RXT no fast path credis	No room in RXT for fast path packets
Slow path packets slice load	Number of slow path packets loaded to HCA as slices from the network
Fast path packets slice load	Number of fast path packets loaded to HCA as slices from the network
Steering pipe 0 processing time	Number of clocks that steering pipe 0 worked
Steering pipe 1 processing time	Number of clocks that steering pipe 1 worked
WQE address translation back-pressure	No credits between RXW and TPT
Receive WQE cache miss	Number of packets that got miss in RWqe buffer L0 cache
Receive WQE cache hit	Number of packets that got hit in RWqe buffer L0 cache
Slow packets miss in LDB L1 cache	Number of slow packet that got miss in LDB L1 cache
Slow packets hit in LDB L1 cache	Number of slow packet that got hit in LDB L1 cache
Fast packets miss in LDB L1 cache	Number of fast packet that got miss in LDB L1 cache
Fast packets hit in LDB L1 cache	Number of fast packet that got hit in LDB L1 cache
Packets miss in LDB L2 cache	Number of packet that got miss in LDB L2 cache
Packets hit in LDB L2 cache	Number of packet that got hit in LDB L2 cache
Slow packets miss in REQSL L1	Number of slow packet that got miss in REQSL L1 fast cache
Slow packets hit in REQSL L1	Number of slow packet that got hit in REQSL L1 fast cache
Fast packets miss in REQSL L1	Number of fast packet that got miss in REQSL L1 fast cache
Fast packets hit in REQSL L1	Number of fast packet that got hit in REQSL L1 fast cache
Packets miss in REQSL L2	Number of packet that got miss in REQSL L2 fast cache
Packets hit in REQSL L2	Number of packet that got hit in REQSL L2 fast cache
No PXT credits time	Number of clocks in which there were no PXT credits
EQ slices busy time	Number of clocks where all EQ slices were busy
CQ slices busy time	Number of clocks where all CQ slices were busy
MSIX slices busy time	Number of clocks where all MSIX slices were busy
QP done due to VL limited	Number of QP done scheduling due to VL limited (e.g. lack of VL credits)
QP done due to desched	Number of QP done scheduling due to desched (Tx full burst size)
QP done due to work done	Number of QP done scheduling due to work done (Tx all QP data)
QP done due to limited	Number of QP done scheduling due to limited rate (e.g. max read)
QP done due to E2E creadits	Number of QP done scheduling due to e2e credits (other peer credits)
Packets sent by SXW to SXP	Number of packets that were authorized to send by SXW (to SXP)
Steering hit	Number of steering lookups that were hit
Steering miss	Number of steering lookups that were miss
Steering processing time	Number of clocks that steering pipe worked

Table 25 - Device Diagnostics Counters

Mellanox WinOF-2 Device Diagnostic Counters	Description
No send credits for scheduling time	The number of clocks that were no credits for scheduling (Tx)
No slow path send credits for scheduling time	The number of clocks that were no credits for scheduling (Tx) for slow path
TPT indirect memory key access	The number of indirect mkey accesses

4.6.16.5.1.8Mellanox WinOF-2 PCI Device Diagnostic Counters

Mellanox WinOF-2 PCI device diagnostic counters set consists of the following counters:

Table 26 - PCI Device Diagnostic Counters

Mellanox WinOF-2 PCI Device Diagnostic Counters	Description
PCI back-pressure cycles	The number of clocks where BP was received from the PCI, while trying to send a packet to the host.
PCI back-pressure cycles/Sec	The rate of clocks where BP was received from the PCI, while trying to send a packet to the host.
PCI write back-pressure cycles	The number of clocks where there was lack of posted outbound credits from the PCI, while trying to send a packet to the host.
PCI write back-pressure cycles/Sec	The rate of clocks where there was lack of posted outbound credits from the PCI, while trying to send a packet to the host.
PCI read back-pressure cycles	The number of clocks where there was lack of non-posted outbound credits from the PCI, while trying to send a packet to the host.
PCI read back-pressure cycles/Sec	The rate of clocks where there was lack of non-posted outbound credits from the PCI, while trying to send a packet to the host.
PCI read stuck no receive buffer	The number of clocks where there was lack in global byte credits for non-posted outbound from the PCI, while trying to send a packet to the host.
Available PCI BW	The number of 128 bytes that are available by the host.
Used PCI BW	The number of 128 bytes that were received from the host.
RX PCI errors	The number of physical layer PCIe signal integrity errors. The number of transitions to recovery due to Framing errors and CRC (dlp and tlp). If the counter is advancing, try to change the PCIe slot in use. **Note:** Only a continues increment of the counter value is considered an error.
TX PCI errors	The number of physical layer PCIe signal integrity errors. The number of transition to recovery initiated by the other side (moving to Recovery due to getting TS/EIEOS). If the counter is advancing, try to change the PCIe slot in use. **Note:** transitions to recovery can happen during initial machine boot. The counter should not increment after boot. **Note:** Only a continues increment of the counter value is considered an error.

Table 26 - PCI Device Diagnostic Counters

Mellanox WinOF-2 PCI Device Diagnostic Counters	Description
TX PCI non-fatal errors	The number of PCI transport layer Non-Fatal error msg sent. If the counter is advancing, try to change the PCIe slot in use.
TX PCI fatal errors	The number of PCIe transport layer fatal error msg sent. If the counter is advancing, try to change the PCIe slot in use.

4.6.16.5.1.9 Mellanox WinOF-2 Hardware Counters

Mellanox WinOF-2 hardware counters set provides monitoring for hardware RSS behavior. These counters are accumulative and collect packets per type (IPv4 or IPv6 only, IPv4/6 TCP or UDP), for tunneled and non-tunneled traffic separately, and when the hardware RSS is functional or dysfunctional.

The counters are activated upon first addition into perfmon, and are stopped upon removal.

Setting "RssCountersActivatedAtStartup" registry key to 1 in the NIC properties will cause the RSS counters to collect data from the startup of the device.

All RSS counters are provided under the counter set "Mellanox Adapter RSS Counters"

Each Ethernet adapter provides multiple instances:

- Instance per vPort per CPU in HwRSS mode is formatted: <NetworkAdapter> + vPort_<id> CPU_<cpu>
- Instance per network adapter per CPU in native RSS per CPU is formatted: <Network-Adapter> CPU_<cpu> .

Table 27 - RSS Diagnostic Counters

Mellanox WinOF-2 RSS Diagnostic Counters	Description
Rss IPv4 Only	Shows the number of received packets that have RSS hash calculated on IPv4 header only
Rss IPv4/TCP	Shows the number of received packets that have RSS hash calculated on IPv4 and TCP headers
Rss IPv4/UDP	Shows the number of received packets that have RSS hash calculated on IPv4 and UDP headers
Rss IPv6 Only	Shows the number of received packets that have RSS hash calculated on IPv6 header only
Rss IPv6/TCP	Shows the number of received packets that have RSS hash calculated on IPv6 and TCP headers
Rss IPv6/UDP	Shows the number of received packets that have RSS hash calculated on IPv6 and UDP headers
Encapsulated Rss IPv4 Only	Shows the number of received encapsulated packets that have RSS hash calculated on IPv4 header only
Encapsulated Rss IPv4/TCP	Shows the number of received encapsulated packets that have RSS hash calculated on IPv4 and TCP headers

Table 27 - RSS Diagnostic Counters

Mellanox WinOF-2 RSS Diagnostic Counters	Description
Encapsulated Rss IPv4/UDP	Shows the number of received encapsulated packets that have RSS hash calculated on IPv4 and UDP headers
Encapsulated Rss IPv6 Only	Shows the number of received encapsulated packets that have RSS hash calculated on IPv6 header only
Encapsulated Rss IPv6/TCP	Shows the number of received encapsulated packets that have RSS hash calculated on IPv6 and TCP headers
Encapsulated Rss IPv6/UDP	Shows the number of received encapsulated packets that have RSS hash calculated on IPv6 and UDP headers
NonRss IPv4 Only	Shows the number of IPv4 packets that have no RSS hash calculated by the hardware
NonRss IPv4/TCP	Shows the number of IPv4 TCP packets that have no RSS hash calculated by the hardware
NonRss IPv4/UDP	Shows the number of IPv4 UDP packets that have no RSS hash calculated by the hardware
NonRss IPv6 Only	Shows the number of IPv6 packets that have no RSS hash calculated by the hardware
NonRss IPv6/TCP	Shows the number of IPv6 TCP packets that have no RSS hash calculated by the hardware
NonRss IPv6/UDP	Shows the number of IPv6 UDP packets that have no RSS hash calculated by the hardware
Encapsulated NonRss IPv4 Only	Shows the number of encapsulated IPv4 packets that have no RSS hash calculated by the hardware
Encapsulated NonRss IPv4/TCP	Shows the number of encapsulated IPv4 TCP packets that have no RSS hash calculated by the hardware
Encapsulated NonRss IPv4/UDP	Shows the number of encapsulated IPv4 UDP packets that have no RSS hash calculated by the hardware
Encapsulated NonRss IPv6 Only	Shows the number of encapsulated IPv6 packets that have no RSS hash calculated by the hardware
Encapsulated NonRss IPv6/TCP	Shows the number of encapsulated IPv6 TCP packets that have no RSS hash calculated by the hardware
Encapsulated NonRss IPv6/UDP	Shows the number of encapsulated IPv6 UDP packets that have no RSS hash calculated by the hardware
Rss Misc	Shows the number of received packets that have RSS hash calculated with unknown RSS hash type
Encapsulated Rss Misc	Shows the number of received encapsulated packets that have RSS hash calculated with unknown RSS hash type
NonRss Misc	Shows the number of packets that have no RSS hash calculated by the hardware for no apparent reason
Encapsulated NonRss Misc	Shows the number of encapsulated packets that have no RSS hash calculated by the hardware for no apparent reason

4.6.17 Single Root IO Virtualization (SR-IOV)

> Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-Ex adapter cards.

4.6.17.1 System Requirements

➢ *To set up an SR-IOV environment, the following is required:*

• A server and BIOS with SR-IOV support. BIOS settings might need to be updated to enable virtualization support and SR-IOV support.

• Hypervisor OS - Windows Server 2012 R2

• Virtual Machine (VM) OS

  • The VM OS can be either Windows Server 2012 and above.

• Mellanox ConnectX-3/ ConnectX-3 Pro adapter card with SR-IOV capability

  • Mellanox WinOF 4.61 driver or higher version (ConnectX-3/ConnectX-3 Pro only)

  • Firmware version - 2.30.8000 or higher.

• Mellanox ConnectX®-4 Lx adapter cards

  • Mellanox WinOF-2 1.50 or higher

  • Firmware version - 14.17.20.52 or higher.

• Mellanox ConnectX®-4 adapter cards

  • Mellanox WinOF-2 1.50 or higher

  • Firmware version - 12.17.20.52

• Mellanox ConnectX®-5 Ex adapter cards

  • Mellanox WinOF-2 1.50 or higher

  • Firmware version - 16.23.10.20

4.6.17.1.1 Feature Limitations (ConnectX-4 and ConnectX-4 Lx Only)

RDMA (i.e RoCE) capability is not available in SR-IOV mode

4.6.18 Configuring SR-IOV Host Machines

This section provides the necessary steps required for configuring host machine.

➢ *Enabling SR-IOV in BIOS:*

Depending on your system, perform the steps below to set up your BIOS. The figures used in this section are for illustration purposes only. For further information, please refer to the appropriate BIOS User Manual.

➢ *To enable SR-IOV in BIOS:*

Step 1. Make sure the machine's BIOS supports SR-IOV. Please consult BIOS vendor website for SR-IOV supported BIOS versions list. Update the BIOS version if necessary.

Step 2. Follow BIOS vendor guidelines to enable SR-IOV according to BIOS User Manual.

For example,

a. Enable SR-IOV.

b. Enable " Virtualization Technologhy" Support
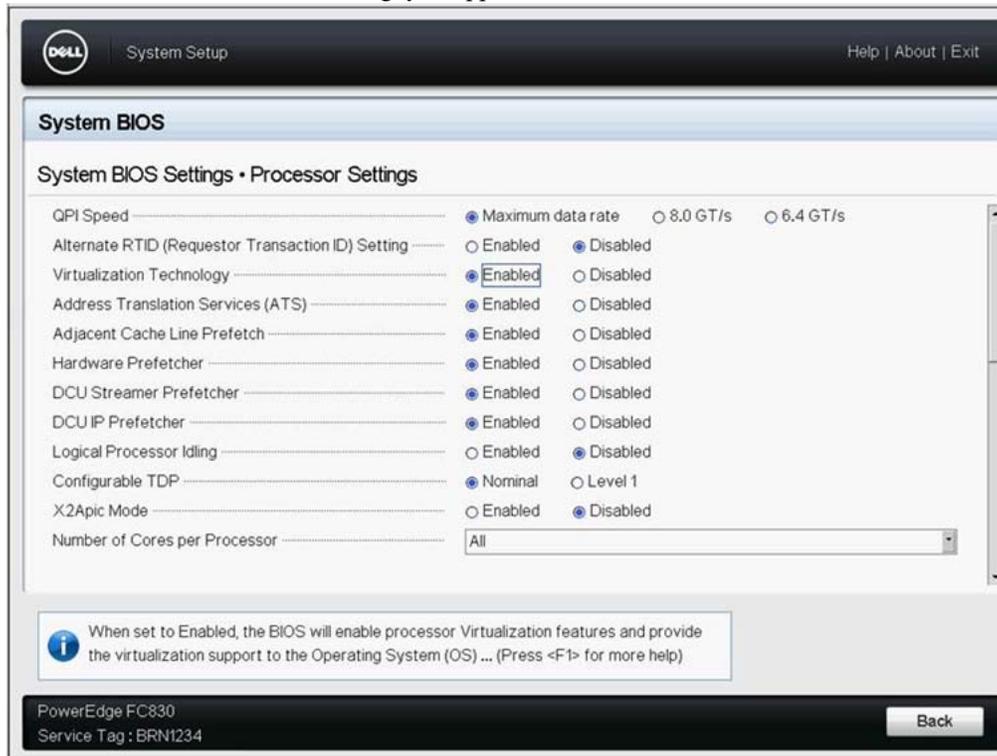
For further details, please refer to the vendor's website.

4.6.18.1 Installing Hypervisor Operating System

➢ *To install Hypervisor Operating System:*

Step 1. Install Windows Server 2012 R2.

Step 2. Install Hyper-V role:

- Go to: Server Manager -> Manage -> Add Roles and Features and set the following:

- Installation Type -> Role-based or Feature-based Installation.

- Server Selection -> Select a server from the server pool.

- Server Roles -> Hyper-V (see figures below).
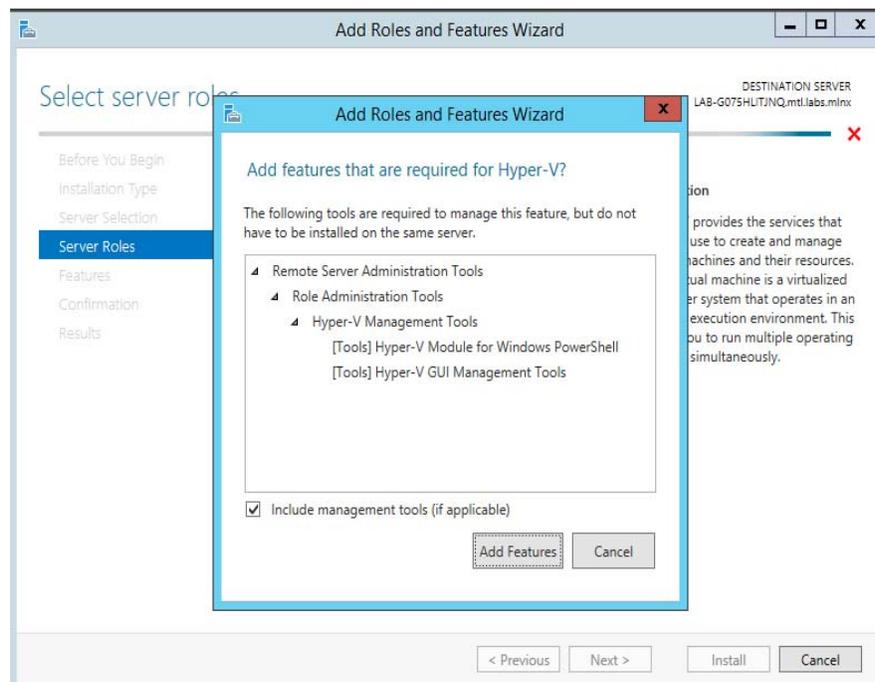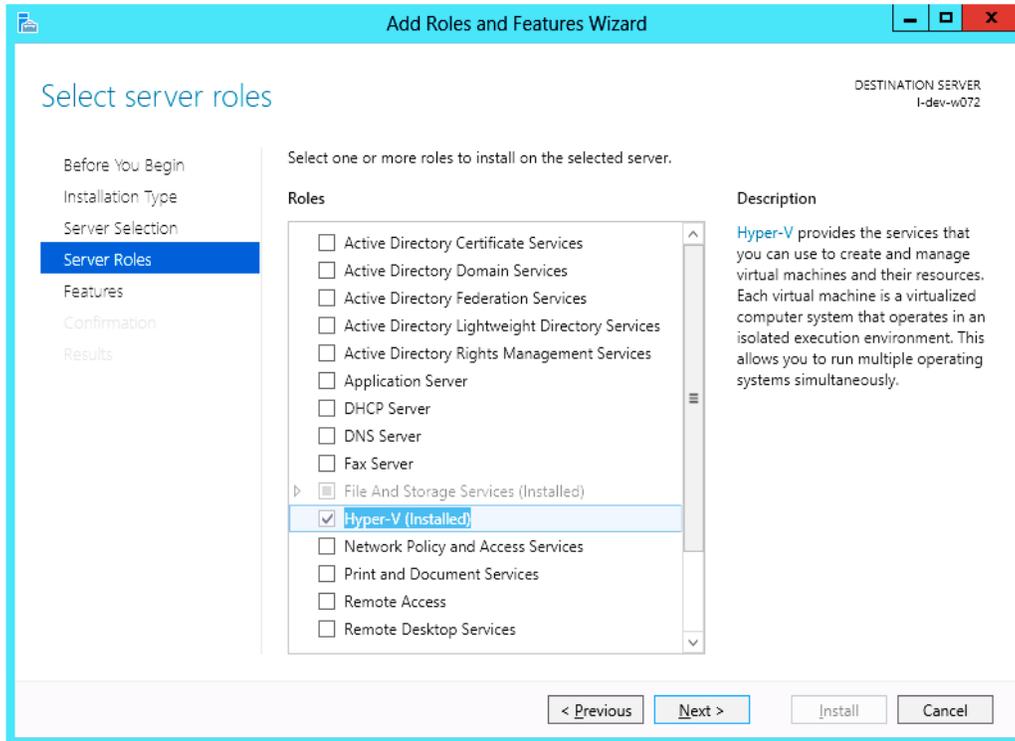
Step 3. Install Hyper-V Management Tools.

Go to: Features -> Remote Server Administration Tools -> Role Administration Tools -> Hyper-V Administration Tool.
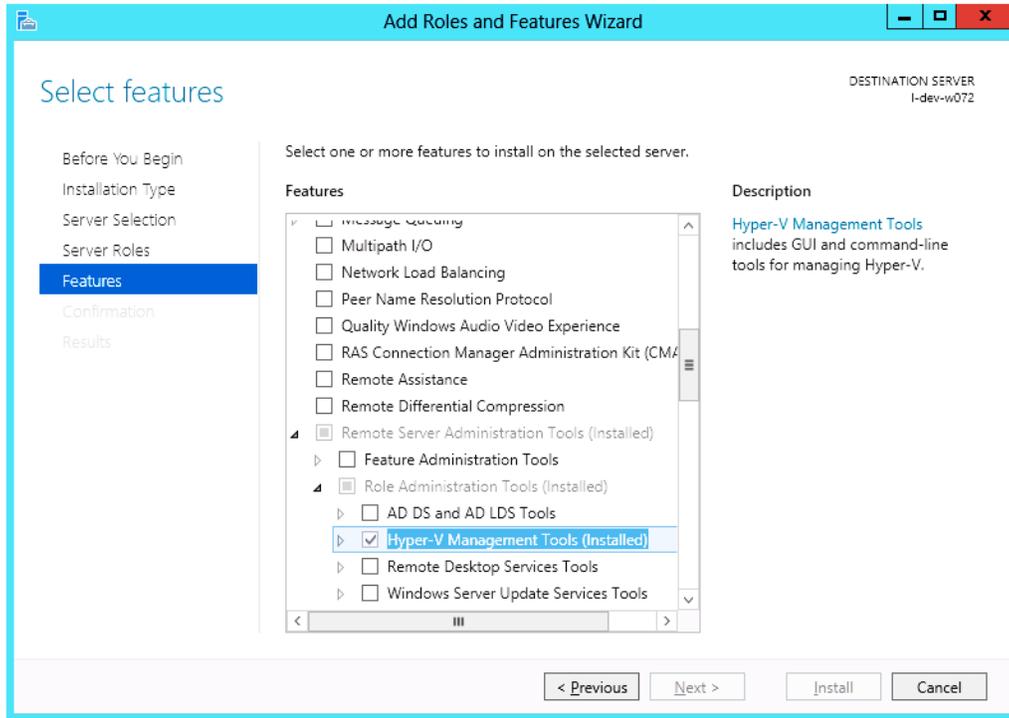
Step 4. Confirm the Installation.
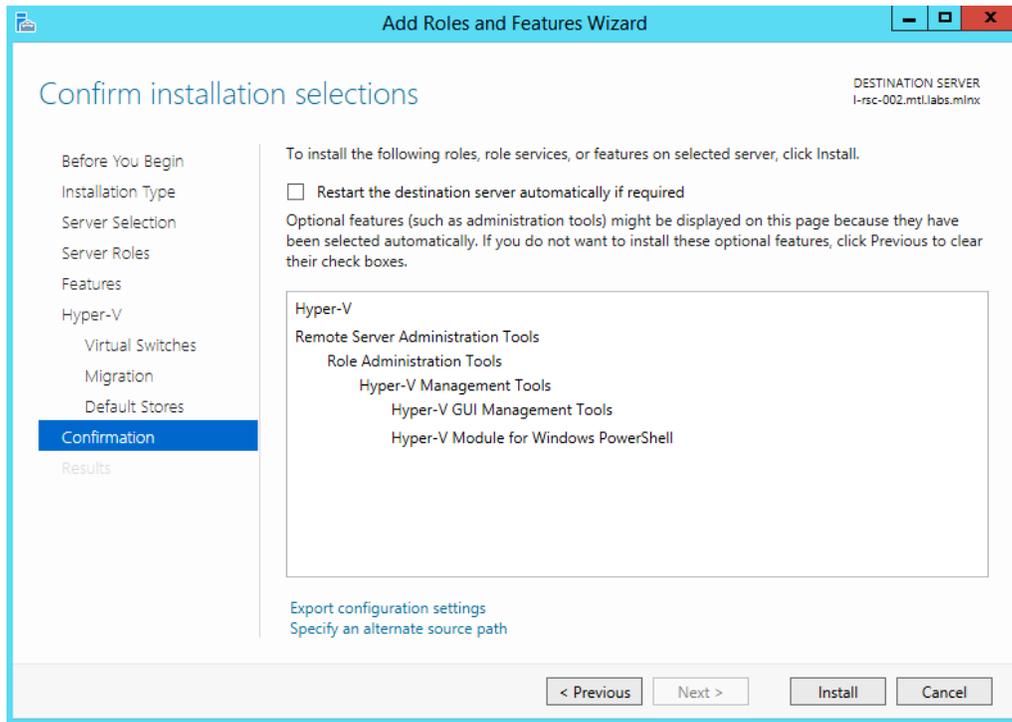
Step 5. Click Install.
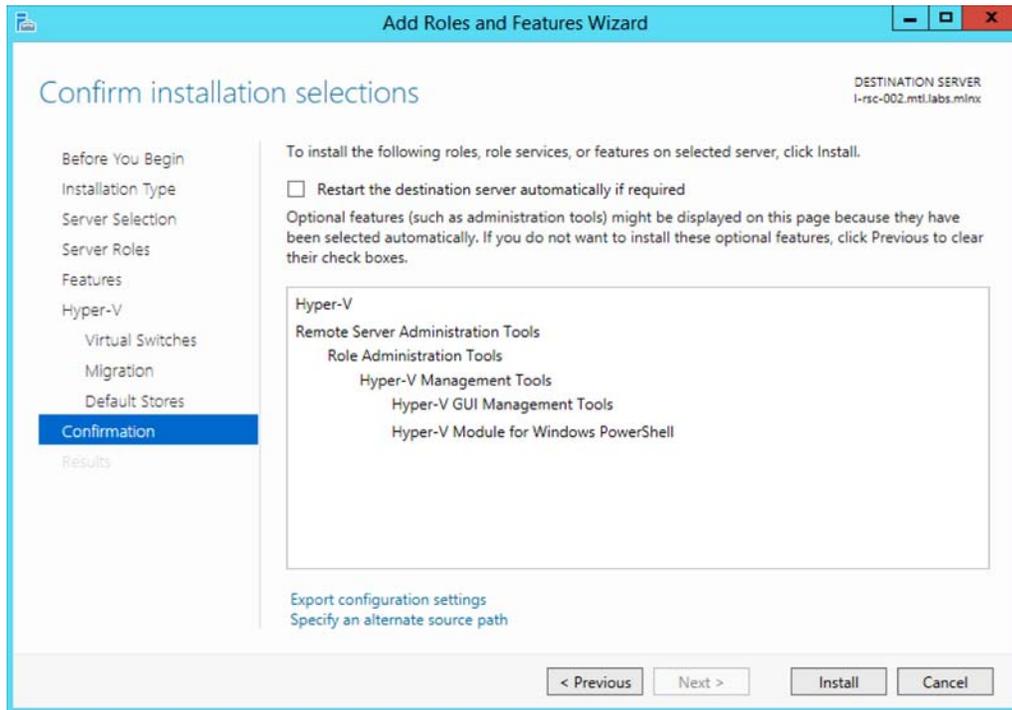
Step 6. Reboot the system.

4.6.18.2 Verifying SR-IOV Support Within the Host Operating System

➢ *To verify the system is properly configured for SR-IOV:*

Step 1. Go to: Start-> Windows Powershell.

Step 2. Run the following PowerShell commands.

```
PS $(Get -VMhost).IovSupport
PS $(Get -VMhost).IovSupportReasons
```

If SR-IOV is supported by the OS, the output in the PowerShell is as in Figure 11.

Figure 11: Operating System Supports SR-IOV
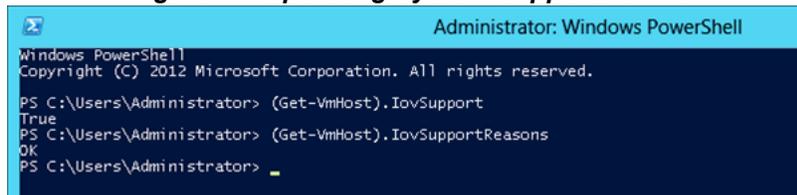
> If BIOS was updated according to BIOS vendor instructions and you see the message displayed in Figure 12, update the registry configuration as described in the (Get-VmHost).IovSupportReasons message.

Figure 12: SR-IOV Support

Step 3. Reboot.

Step 4. Verify the system is configured correctly for SR-IOV as described in Steps 1 and 2.

4.6.18.3 Creating a Virtual Machine

➢ *To create a virtual machine:*

Step 1.   Go to: Server Manager -> Tools -> Hyper-V Manager.

Step 2.   Go to: New -> Virtual Machine and set the following:

Name: <name>

Startup memory: 4096 MB

Connection: Not Connected

Figure 13: Hyper-V Manager
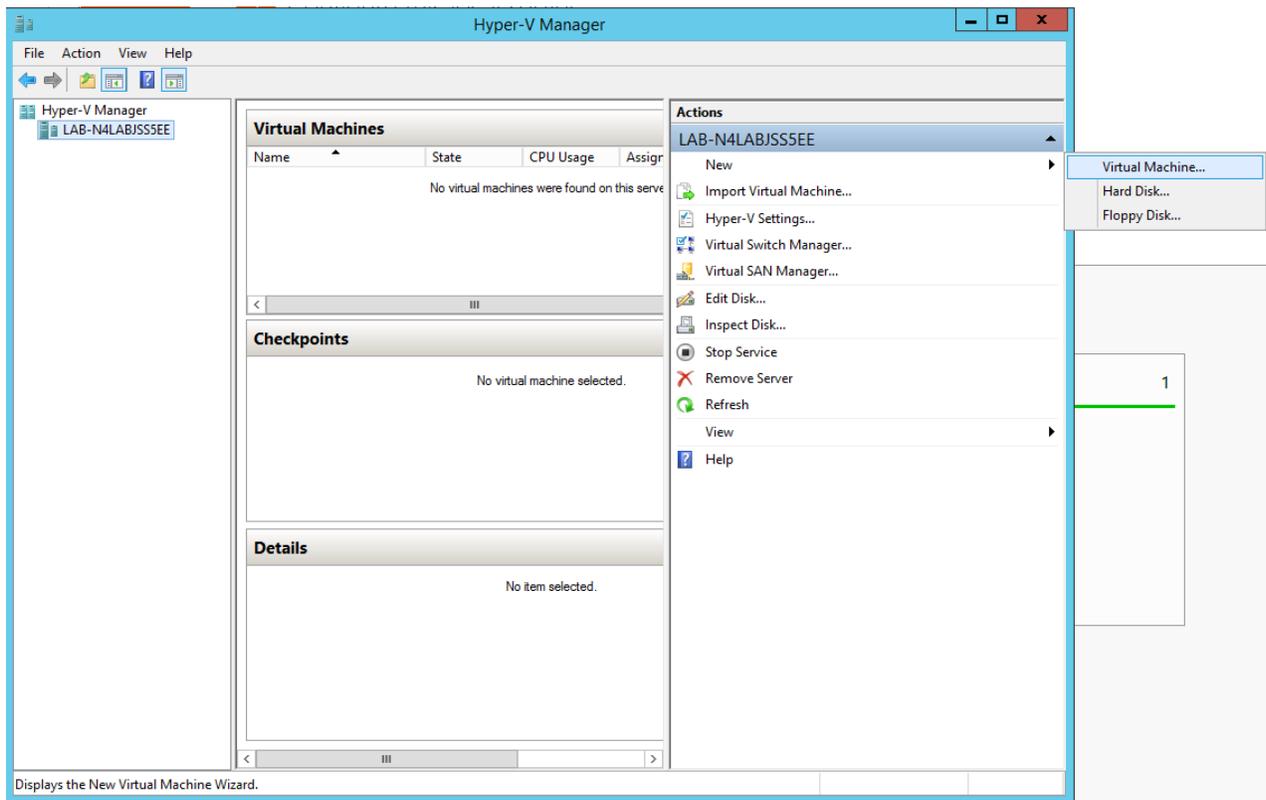
Step 3.   Connect the virtual hard disk in the New Virtual Machine Wizard.

Step 4.   Go to: Connect Hard Disk -> Use an existing virtual hard disk.

Step 5.   Select the location of the vhd file.

Figure 14: Connect Virtual Hard Disk
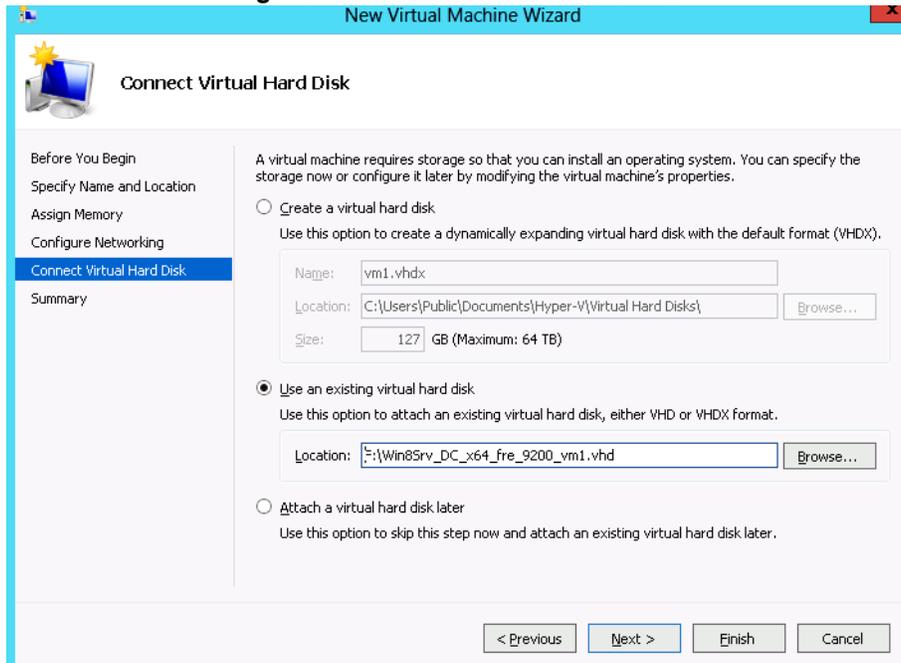
4.6.18.4 Enabling SR-IOV in Mellanox WinOF Package

Applies to ConnectX-3 and ConnectX-3 Pro only.

For SR-IOV Configuration, please refer to SR-IOV Configuration on page 216.

➢ *To enable SR-IOV in Mellanox WinOF Package*

Step 1. Install Mellanox WinOF package that supports SR-IOV.

Step 2. Query SR-IOV configuration with Powershell.

```
PS $ Get-MlnxPCIDeviceSriovSetting
```

Example:

```
Caption          : MLNX_PCIDeviceSriovSettingData 'Mellanox ConnectX-3 PRO EN (MT04103)
Network Adapter'
Description      : Mellanox ConnectX-3 PRO EN (MT04103) Network Adapter
ElementName      : HCA 0
InstanceID       : PCI\VEN_15B3&DEV_1007&SUBSYS_22F5103C&REV_00\24BE05FFFFB9E2E000
Name             : HCA 0
Source           : 3
SystemName       : LAB-N4LABJSS5EE
SriovEnable      : False
SriovPort1NumVFs : 16
SriovPort2NumVFs : 0
SriovPortMode    : 0
PSComputerName   :
```

Step 3. Enable SR-IOV through Powershell on both ports.[1]

```
PS $ Set-MlnxPCIDeviceSriovSetting -Name "HCA 0" -SriovEnable $true -SriovPortMode 2
-SriovPort1NumVFs 8 -SriovPort2NumVFs 8
```

Example:

```
Confirm
Are you sure you want to perform this action?
Performing the operation "SetValue" on target "MLNX_PCIDeviceSriovSettingData: MLNX_P-
CIDeviceSriovSettingData 'Mellanox
 ConnectX-3 PRO VPI (MT04103) Network Adapter' (InstanceID =
"PCI\VEN_15B3&DEV_1007&SUBSYS_22F5103C&R...)".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is "Y"):
y
```

---

1. **SriovPortMode 2** - Enables SR-IOV on both ports.
   SriovPort1NumVFs 8 & **SriovPort2NumVFs 8** - Enable 8 Virtual Functions for each port when working in manual mode. By default, there are assigned 16 virtual functions on the first port.

Mellanox device is a dual-port single-PCI function. Virtual Functions' pool belongs to both ports. To define how the pool is divided between the two ports use the Powershell "`SriovPort1NumVFs`" command.

Table 28 - SR-IOV Mode Configuration Parameters

Parameter Name	Values	Description
SriovEnable	0 = RoCE (default) 1 = SR-IOV	Configures the RDMA or SR-IOV mode. The default WinOF configuration mode is RoCE. To switch to SR-IOV, set the `SriovEnable` registry key value to 1. By default in SR-IOV mode, all VF pool belongs to Port 1. To change the VF pool distribution, change the `PortMode` to manual and choose how many VFs to assign to each port. **Note**: RDMA is not supported in SR-IOV mode.
SriovPortMode	0 = auto_port1 (default) 1 = auto_port2 2 = manual	Configures the number of VFs to be enabled by the bus driver to each port. **Note**: In auto_portX mode, port X will have the number of VFs according to the committed value in the device and the other port will have no SR-IOV and it will support native Ethernet (i.e. no RoCE). Setting this parameter to "Manual" will configure the number of VFs for each port according to the registry key MaxVFPortX. **Note**: The number of VFs can be configured both on a Mellanox bus driver level and Network Interface level (i.e using Set-NetAdapterSriov Powershell cmdlet). The number of VFs actually available to the Network Interface is the minimum value between mellanox bus driver configuration and Network Interface configuration. For example, if 8 VFs support was set in firmware, SriovPortMode is auto_port1, and Network Interface was allowed 32 VFs using SetNetAdapterSriov Powershell cmdlet, the actual number of VFs available to Network Interface will be 8.
MaxVFPort1 MaxVFPort2	16=(default)	MaxVFPort\<i\> specifies the maximum number of VFs that are allowed per port. This is the number of VFs the bus driver will open when working in manual mode. Note: If the total number of VFs requested is larger than the number of VFs set in firmware, each port X(1\2) will have the number of VFs according to the following formula: (SriovPortXNumVFs / (SriovPort1NumVFs+SriovPort2NumVFs))*number of VFs set in firmware.

Step 4. Verify the new values were set correctly.

```
PS $ Get-MlnxPCIDeviceSriovSetting
```

Example:

```
Caption            : MLNX_PCIDeviceSriovSettingData 'Mellanox ConnectX-3 PRO EN (MT04103)
Network Adapter'
Description        : Mellanox ConnectX-3 PRO EN (MT04103) Network Adapter
ElementName        : HCA 0
InstanceID         : PCI\VEN_15B3&DEV_1007&SUBSYS_22F5103C&REV_00\24BE05FFFFB9E2E000
Name               : HCA 0
Source             : 3
SystemName         : LAB-N4LABJSS5EE
SriovEnable        : True
SriovPort1NumVFs   : 8
SriovPort2NumVFs   : 8
SriovPortMode      : 2
PSComputerName     :
```

Step 5. Check in the System Event Log that SR-IOV is enabled.

    Step a. Open the View Event Logs/Event Viewer.
Go to: Start -> Control Panel -> System and Security -> Administrative Tools ->
View Event Logs/Event Viewer

    Step b. Open the System logs.
Event Viewer (Local) -> Windows Logs -> System

Figure 15: System Event Log
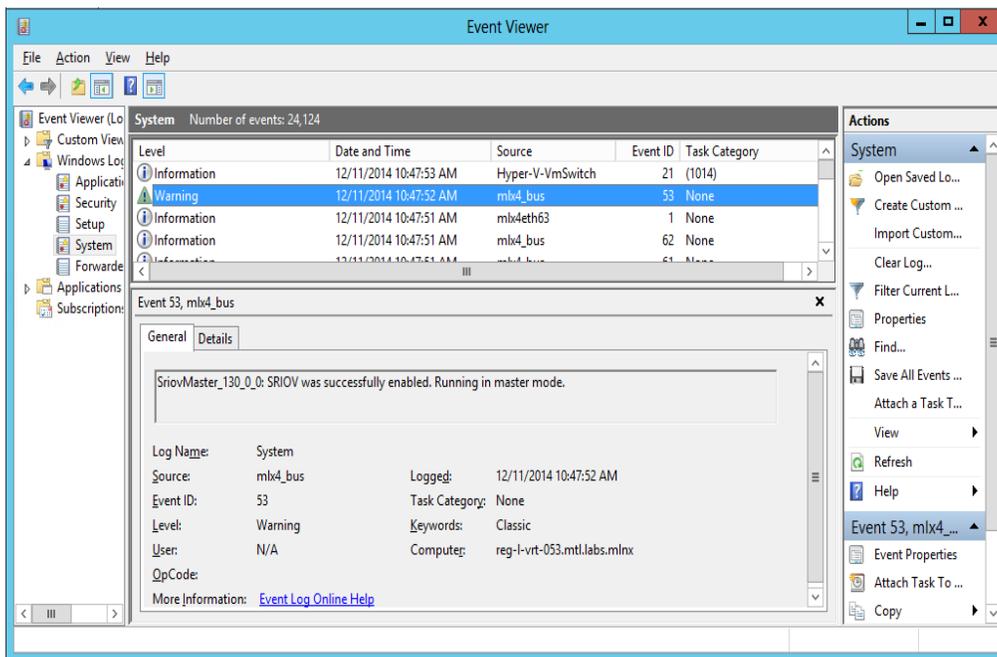
4.6.18.5 Enabling SR-IOV in Firmware - ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex

> Applies to ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex adapter cards only.

> For SR-IOV Configuration, please refer to SR-IOV Configuration on page 216.

➢ **To enable SR-IOV using mlxconfig:**

mlxconfig is part of MFT tools used to simplify firmware configuration. The tool is available with MFT tools 3.6.0 or higher.

Step 1. Download MFT for Windows.

www.mellanox.com > Products > Software > Firmware Tools

Step 2. Get the device ID (look for the "`_pciconf`" string in the output).

```
> mst status
```

Example:

```
MST devices:
------------
mt4115_pciconf0
```

Step 3. Check the current SR-IOV configuration.

```
> mlxconfig -d mt4115_pciconf0 q
```

Example:

```
Device #1:
----------

Device type:    ConnectX4 Lx
PCI device:     mt4115_pciconf0

Configurations:         Current
        SRIOV_EN        N/A
        NUM_OF_VFS      N/A
        WOL_MAGIC_EN_P2 N/A
        LINK_TYPE_P1    N/A
        LINK_TYPE_P2    N/A
```

Step 4. Enable SR-IOV with 16 VFs.

```
> mlxconfig -d mt4115_pciconf0 s SRIOV_EN=1 NUM_OF_VFS=16
```

> ⚠️ **Warning**: Care should be taken in increasing the number of VFs. More VFs can lead to exceeding the BIOS limit of MMIO available address space.

Example:

```
Device #1:
----------

Device type:    ConnectX4 Lx
PCI device:     mt4115_pciconf0

Configurations:          Current New
        SRIOV_EN         N/A     1
        NUM_OF_VFS       N/A     16
        WOL_MAGIC_EN_P2  N/A     N/A
        LINK_TYPE_P1     N/A     N/A
        LINK_TYPE_P2     N/A     N/A


 Apply new Configuration? ? (y/n) [n] : y
Applying... Done!
-I- Please reboot machine to load new configurations.
```

4.6.18.6 Networking - ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex

➢ *To configure Virtual Machine networking:*

Step 1.   Create SR-IOV-enabled Virtual Switch over Mellanox Ethernet Adapter. Go to: Start -> Server Manager -> Tools -> Hyper-V Manager. In the Hyper-V Manager: Actions -> Virtual SwitchManager -> External-> Create Virtual Switch.

Step 2.   Set the following:

Name:

External network:

Enable single-root I/O virtualization (SR-IOV)

Figure 16: Virtual Switch with SR-IOV
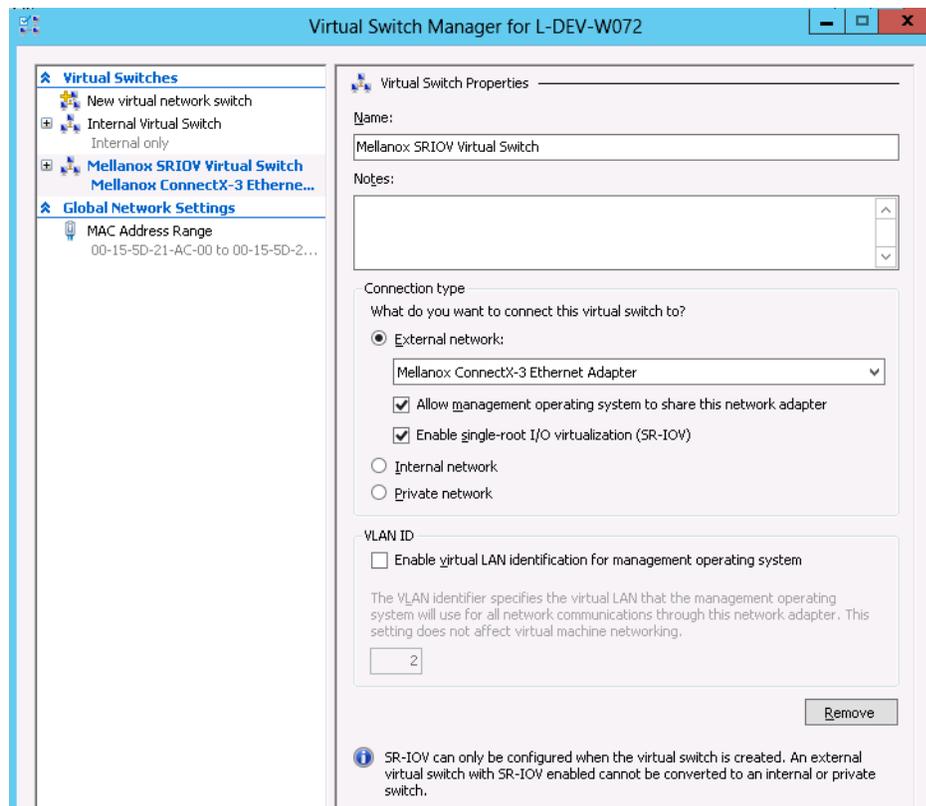
Step 3.   Click **Apply**.

Step 4.   Click **OK**.

Step 5.   Add a VMNIC connected to a Mellanox vSwitch in the VM hardware settings:

- Under Actions, go to Settings -> Add New Hardware-> Network Adapter-> OK.

- In "Virtual Switch" dropdown box, choose Mellanox SR-IOV Virtual Switch.

Figure 17: Adding a VMNIC to a Mellanox V-switch
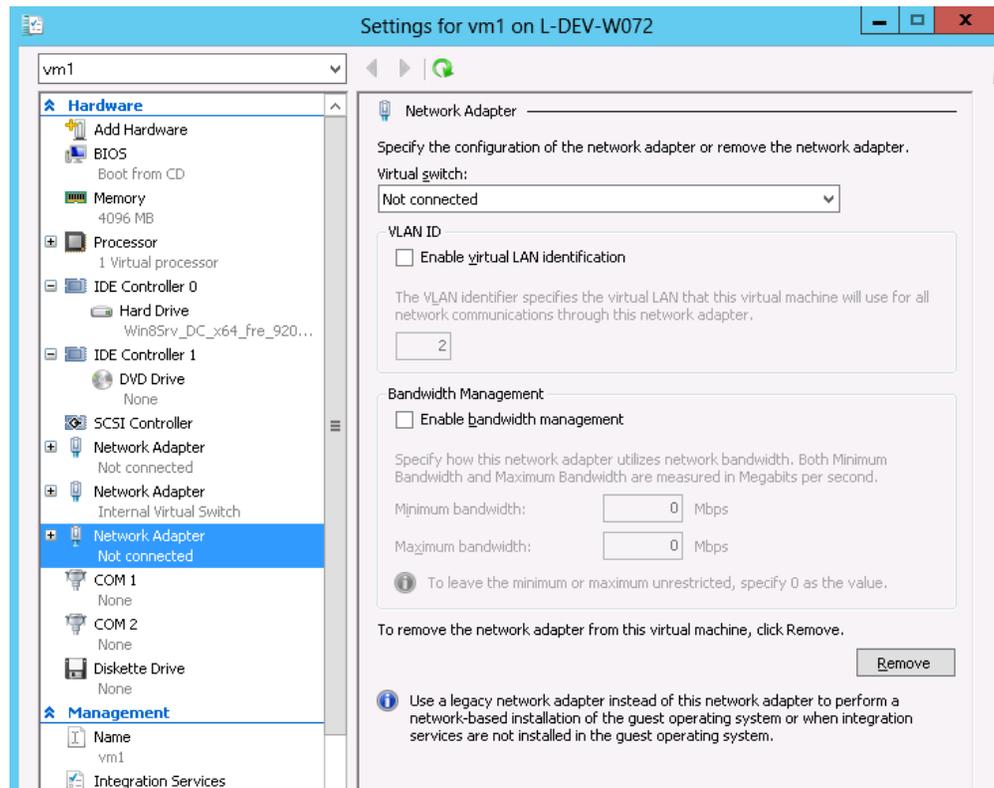
Step 6.   Enable the SR-IOV for Mellanox VMNIC:

- Open VM settings Wizard.

- Open the Network Adapter and choose Hardware Acceleration.

- Tick the "Enable SR-IOV" option.

- Click OK.

Figure 18: Enable SR-IOV on VMNIC
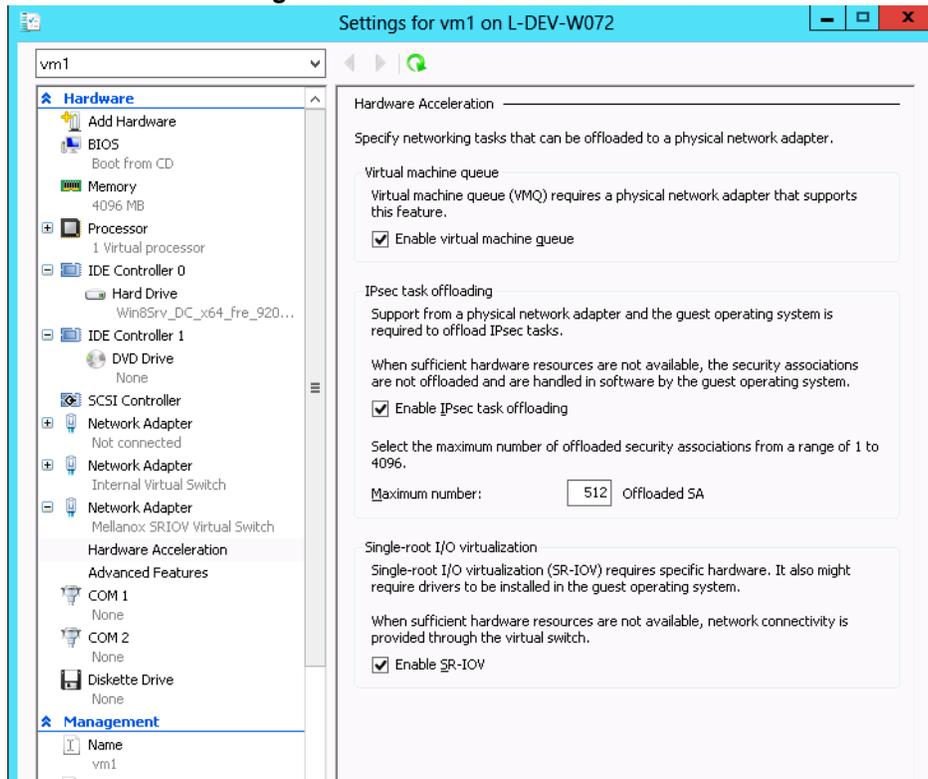
Step 7. Start and connect to the Virtual Machine:

- Select the newly created Virtual Machine and go to: Actions panel-> Connect.

- In the virtual machine window go to: Actions-> Start

Step 8. Copy the WinOF-2 driver package to the VM using Mellanox VMNIC IP address.

Step 9. Install WinOF-2 driver package on the VM.

Step 10. Reboot the VM at the end of installation.

Step 11. Verify that Mellanox Virtual Function appears in the device manager.

Figure 19: Virtual Function in the VM
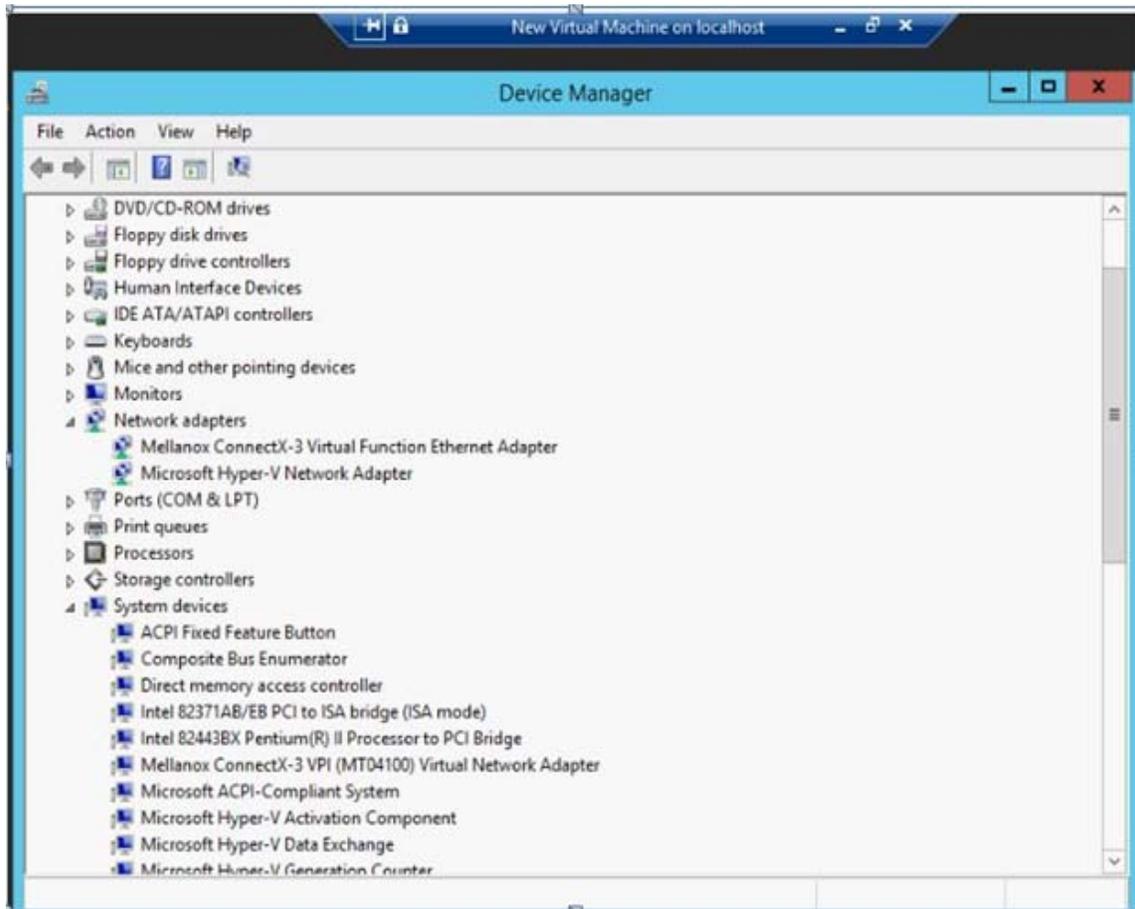
> To achieve best performance on SR-IOV VF, please run the following powershell commands on the host:
>
> For 10GbE:
>
> PS $ Set-VMNetworkAdapter -Name "Network Adapter" - VMName vm1 - IovQueue-PairsRequested4
>
> For 40GbE:
>
> PS $ Set-VMNetworkAdapter -Name "Network Adapter" - VMName vm1 - IovQueue-PairsRequested8

4.6.19 Virtualization - ConnectX-3 and ConnectX-3 Pro

4.6.19.1 Virtual Machine Multiple Queue (VMMQ)

> Applies to ConnectX-3 and CnnectX-3 Pro adapter cards only.

Virtual Machine Multiple Queues (VMMQ), formerly known as Hardware vRSS, is a NIC off-load technology that provides scalability for processing network traffic of a VPort in the host (root partition) of a virtualized node. In essence, VMMQ extends the native RSS feature to the VPorts that are associated with the physical function (PF) of a NIC including the default VPort.

VMMQ is available for the VPorts exposed in the host (root partition) regardless of whether the NIC is operating in SR-IOV or VMQ mode. VMMQ is a feature available in Windows Server 2016.

4.6.19.1.1 System Requirements

- Operating System(s): Windows Server 2016
- Available only for Ethernet (no IPOIB)

4.6.19.2 Network Direct Kernel Provider Interface

> Applies to ConnectX-3 and ConnectX-3 Pro adapter cards only.

As of version 5.25, WinOF supports NDIS Network Direct Kernel Provider Interface version 2. The Network Direct Kernel Provider Interface (NDKPI) is an extension to NDIS that allows IHVs to provide kernel-mode Remote Direct Memory Access (RDMA) support in a network adapter.

4.6.19.2.1 System Requirement

- Operating System: Windows Server 2012 R2 (Without NDK from/to a VM) and Windows 2016
- Firmware Version:2.40.50.48 or higher

4.6.20 PacketDirect Provider Interface

> Applies to ConnectX-3 and ConnectX-3 Pro adapter cards only.

As of v5.25, WinOF supports NDIS PacketDirect Provider Interface. PacketDirect extends NDIS with an accelerated I/O model, which can increase the number of packets processed per second by an order of magnitude and significantly decrease jitter when compared to the traditional NDIS I/O path.

> PacketDirect is supported only on Ethernet ports.

4.6.21 System Requirements

- Hypervisor OS: Windows Server 2016
- Virtual Machine (VM) OS: Windows Server 2012 and above
- Mellanox ConnectX-3 and ConnectX-3 Pro
- Mellanox WinOF 5.25 or higher
- Firmware version: 2.40.50.48 or higher

4.6.22 Using PacketDirect for VM

> ➢ *To allow a VM to send/receive traffic in PacketDirect mode:*

Step 1. Enable PacketDirect:

- On the Ethernet adapter.

```
PS $ Enable-NetAdapterPacketDirect -Name <EthInterfaceName>
```

- In the Device Manager.
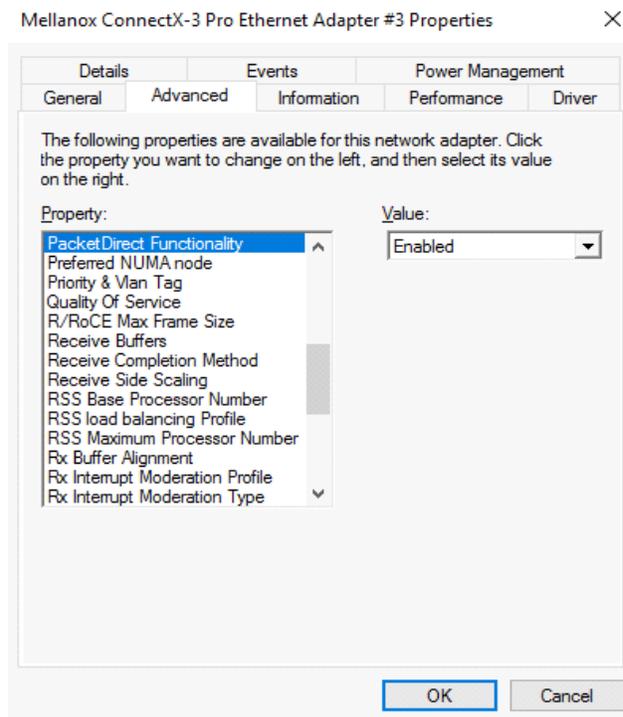
Step 2. Create a vSwitch with PacketDirect enabled.

```
PS $ New-VMSwitch <vSwitchName> -NetAdapterName <EthInterfaceName> -EnablePacketDirect
$true -AllowManagementOS $true
```

Step 3. Enable VFP extension:

- On the vSwitch.

```
PS $ Enable-VMSwitchExtension -VmSwitchName <vSwitchName> -Name "Windows Azure VFP
Switch Extension"
```

• In the Hyper-V Manager: Action->Virtual Switch Manager...
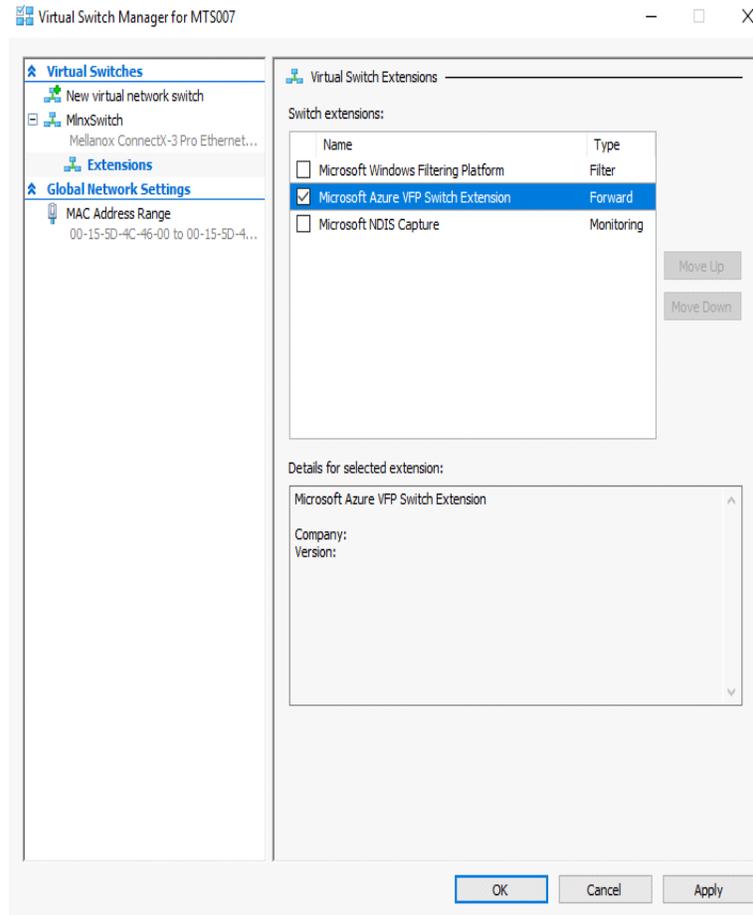
Step 4. Shut down the VM.

```
PS $ Stop-VM -Name <VMName> -Force -Confirm
```

Step 5. Add a virtual network adapter for the VM.

```
PS $ Add-VMNetworkAdapter -VMName <VMName> -SwitchName <vSwitchName> -StaticMacAddress
<StaticMAC Address>
```

Step 6. Start the VM.

```
PS $ Start-VM -Name <VMName>
```

Since VFP is enabled, without any forwarding rules, it will block all traffic going through the VM.

Follow the following steps to unblock the traffic

Step a.   Find the port name for the VM.

```
CMD > vfpctrl /list-vmswitch-port
…...
Port name              : E431C413-D31F-40EB-AD96-0B2D45FE34AA
Port Friendly name     :
Switch name            : 8B288106-9DB6-4720-B144-6CC32D53E0EC
Switch Friendly name   : MlnxSwitch
PortId                 : 3
VMQ Usage              : 0
SR-IOV Usage           : 0
Port type              : Synthetic
 Port is Initialized.
 MAC Learning is Disabled.
NIC name               : bd65960d-4215-4a4f-bddc-962a5d0e2fa0--e7199a49-6cca-
4d3c-a4cd-22907592527e
 NIC Friendly name   : testnic
 MTU                 : 1500
 MAC address         : 00-15-5D-4C-46-00
 VM name             : vm
......
Command list-vmswitch-port succeeded!
```

Step 7.   Disable the port to allow traffic.

```
CMD > vfpctrl /disable-port /port <PortName>
Command disable-port succeeded!
```

The port should be disabled after each reboot of the VM to allow traffic.

5   Remote Boot

Applies to ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex.

5.1   iSCSI Boot

5.1.1   Setting Up iSCSI Boot to RH6.x

5.1.1.1   Configure iSCSI Parameters in HII

a.   On boot press F2 to enter "System setup".

b.   Select "Device Settings".

c.   Select Device/Interface you wish to iSCSI boot from.

d.   Select "NIC Configuration".

e.   Set "Legacy Boot Protocol" to "iSCSI".
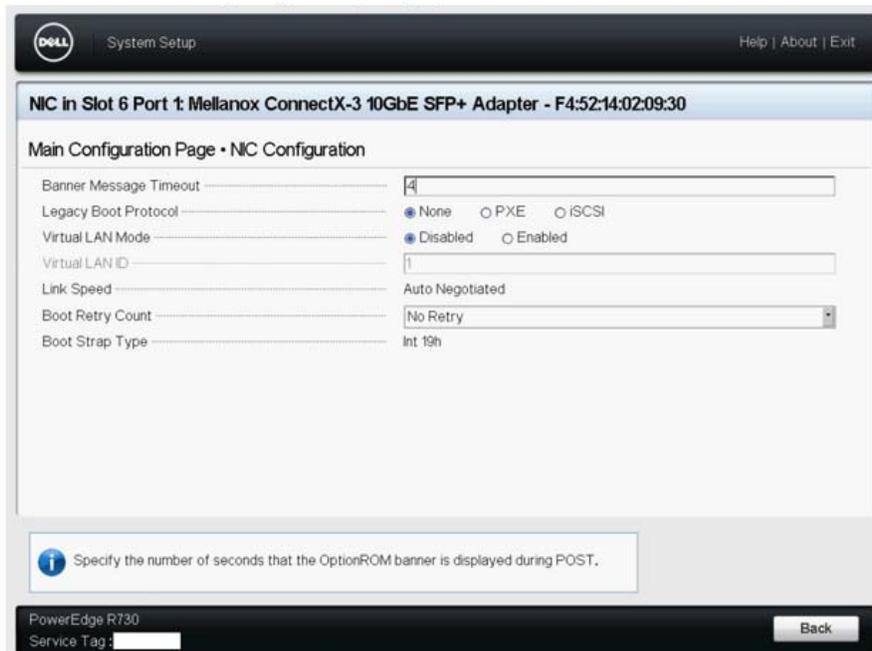
"Boot Option Rom" Enable/Disable has been removed. Attribute is only configurable by "Legacy Boot Protocol"

f.   Go "Back" to the "Main Configuration Page".

g. Select "iSCSI Configuration".

h. Select "iSCSI Initiator Parameters".

i. Enable "TCP/IP Parameters via DHCP" (if you want the initiator to obtain an IP address from the DHCP server.

j. Select "iSCSI Initiator Parameters.

k. Configure the iSCSI Initiator parameters.

l. Go "Back" to the "iSCSI Configuration"

m. Select "iSCSI First Target Parameters".

n. Configure iSCSI Target Paramters & set "Connect" to "Enabled".

o. Exit out of the Configuration pages and save changes.

5.1.1.2 Configure Boot Order of the System

a. Reboot the system and press F2 to enter "System setup".

b. Select "System BIOS".

c. Select "Boot Settings".

d. Confirm "Boot Mode" is set to "BIOS".

e. Select "Boot Option Settings".

f. Set the boot order to the following:

• First, boot to the MLNX adapter/port configured for iSCSI Boot.

• Second, boot to the OS installation media.
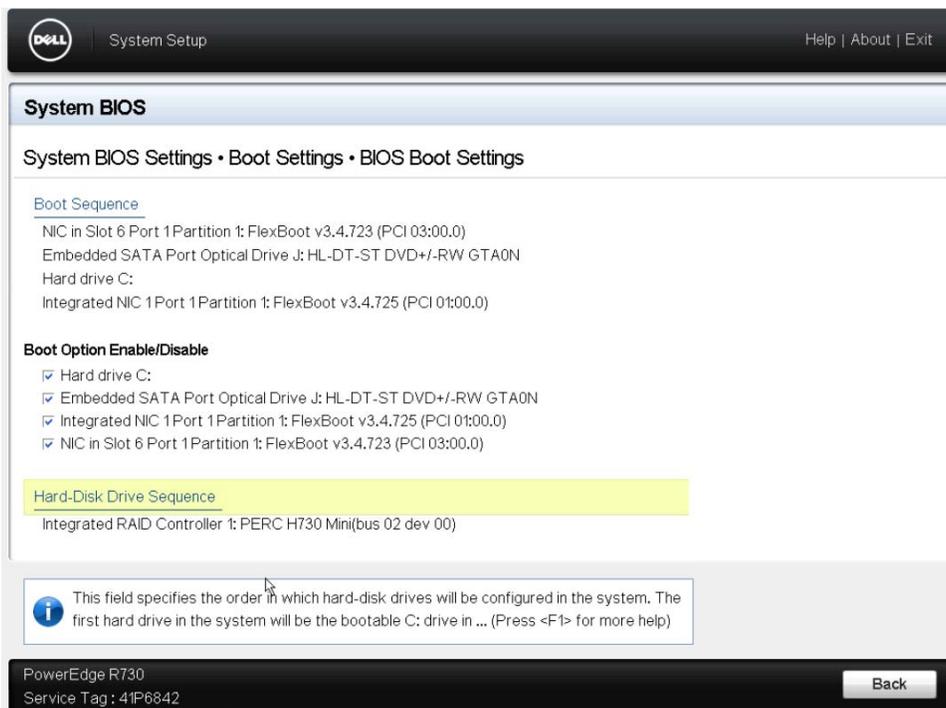
g. Exit out of the Configuration pages and save changes.

h. Reboot the system.

i. The system will now connect to the iSCSI target and detect no bootable image, preserve the iSCSI Connection, then it will boot to the OS installation media.

```
FlexBoot initialising devices...
Initialising completed.

FlexBoot v3.4.723
FlexBoot http://mellanox.com
Features: DNS HTTP iSCSI TFTP VLAN ELF MBOOT PXE bzImage COMBOOT PXEXT
net0: f4:52:14:0a:ba:e0
Using ConnectX-3 on PCI03:00.0 (open)
  [Link:down, TX:0 TXE:0 RX:0 RXE:0]
  [Link status: Unknown (http://ipxe.org/1a086101)]
Waiting for link-up on net0....... ok
Configuring (net0 f4:52:14:0a:ba:e0)....... ok
net0: 172.24.44.115/255.255.255.0 gw 172.24.44.254
net0: fe80::f652:14ff:fe0a:bae0/64
Root path: iscsi:192.168.104.228::3260:0:iqn.2001-05.com.equallogic:0-8a0906-a1f
28c60c-10d30c17d1d57041-demo
Registered SAN device 0x80
Booting from SAN device 0x80
Boot from SAN device 0x80 failed: Exec format error (http://ipxe.org/2e852001)
Preserving SAN device 0x80
No more ports. Exiting FlexBoot...

Booting from HL-DT-ST DVD+/-RW GTA0N
_
```

5.1.1.3 OS Installation Instructions

a. Select "Install or upgrade an existing system".

b. Skip testing the media.

c. Select Language.

d. Select "Specialized Storage Devices".

e. Select tab "Other SAN Devices," and note that the Installer is automatically connected to the target.

f. Select your iSCSI target and press "Next".

g. Select "Yes, discard any data" in the Storage Device Warning popup.

h. Advance to the Installation Type page and select "Use all Spaces" and press "Next".

i. Write changes to disk.

j. Select the OS package you wish to install.

The default installation of Red Hat Enterprise Linux is a basic server install. You can optionally select a different set of software now.

○ Basic Server
○ Database Server
○ Web Server
○ Identity Management Server
○ Virtualization Host
● Desktop
○ Software Development Workstation
○ Minimal
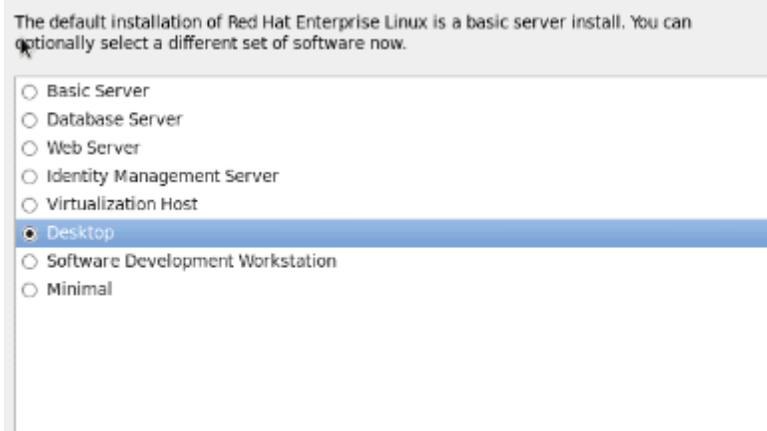
k. At the bottom of the page select "Customize now".

Please select any additional repositories that you want to use for software installation.

☐ Hight Availability
☐ Load Balancer
☑ Red Hat Enterprise Linux

＋ Add additional software repositories   📝 Modify repository

You can further customize the software selection now, or after install via the software management application.

○ Customize later   ● Customize now
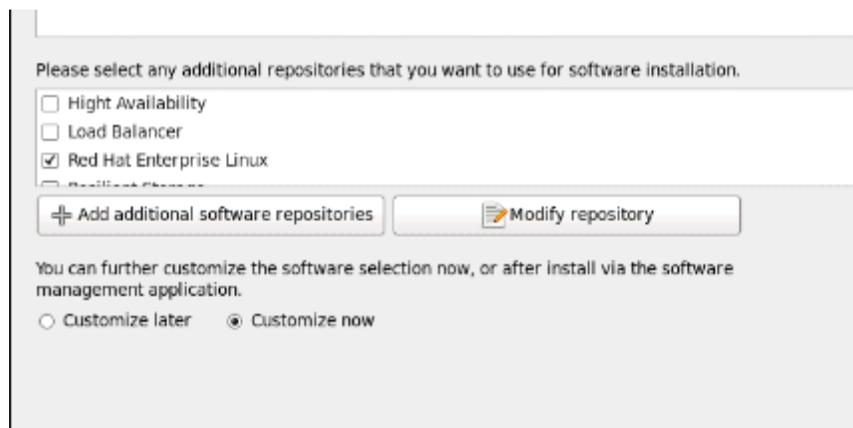
l. Press next.

m. Under "Base System" select "Infiniband Support." This will install the Mellanox inbox drivers.
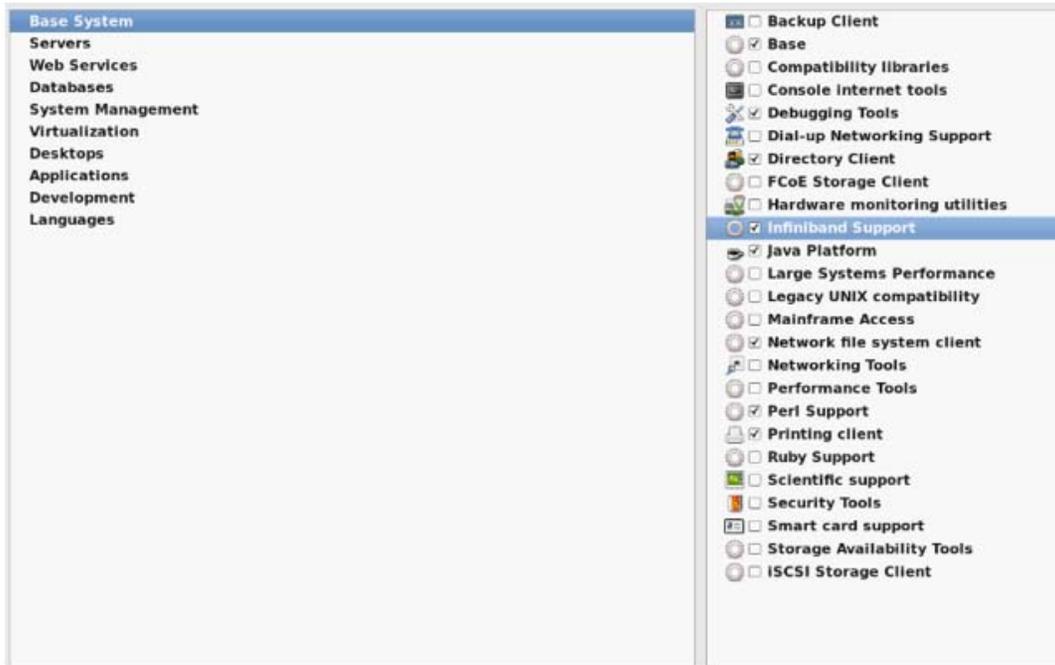
n. Select any other RPMs you wish to install.

o. Press "Next" and the OS installation will begin.

p. Once the OS installation is complete, reboot the system.

q. Finalize your OS configuration.

r. The OS installation process is complete.

5.1.2   Booting Windows from an iSCSI Target

5.1.2.1   Configuring the WDS, DHCP and iSCSI Servers

5.1.2.1.1 Configuring the WDS Server

➢ *To configure the WDS server:*

1. Install the WDS server.

2. Extract the Mellanox drivers to a local directory using the '-a' parameter.

    For boot over Ethernet, when using adapter cards with older firmware version than 2.30.8000, you will need to extract the PXE package, otherwise use Mellanox WinOF VPI package.

    Example:

    ```
    Mellanox.msi.exe -a
    ```

3. Add the Mellanox driver to boot.wim[1].

    ```
    dism /Mount-Wim /WimFile:boot.wim /index:2 /MountDir:mnt
    dism /Image:mnt /Add-Driver /Driver:drivers /recurse
    dism /Unmount-Wim /MountDir:mnt /commit
    ```

4. Add the Mellanox driver to install.wim[1].

```
dism /Mount-Wim /WimFile:install.wim /index:4 /MountDir:mnt
dism /Image:mnt /Add-Driver /Driver:drivers /recurse
dism /Unmount-Wim /MountDir:mnt /commit
```

5. Add the new boot and install images to WDS.

For additional details on WDS, please refer to:

http://technet.microsoft.com/en-us/library/jj648426.aspx

5.1.2.1.2 Configuring iSCSI Target

➢ *To configure iSCSI Target:*

1. Install iSCSI Target (e.g StartWind).

2. Add to the iSCSI target initiators the IP addresses of the iSCSI clients.

5.1.2.1.3 Configuring the DHCP Server

➢ *To configure the DHCP server:*

1. Install a DHCP server.

2. Add to IPv4 a new scope.

3. Add iSCSI boot client identifier (MAC/GUID) to the DHCP reservation.

4. Add to the reserved IP address the following options: if DHCP and WDS are deployed on the same server:

Table 29 - Reserved IP Address Options

Option	Name	Value
017	Root Path	iscsi:11.4.12.65::::iqn:2011-01:iscsiboot Assuming the iSCSI target IP is: **11.4.12.65** and the Target Name: **iqn:2011-01:iscsi-boot**
060	PXEClient	PXEClient
066	Boot Server Host Name	WDS server IP address
067	Boot File Name	boot\x86\wdsnbp.com

When DHCP and WDS are NOT deployed on the same server, DHCP options (60, 66, 67) should be empty, and the WDS option 60 must be configured.

---

1. Use 'index:2' for Windows setup and 'index:1' for WinPE.
1. When adding the Mellanox driver to install.wim, verify the appropriate index for the OS distribution is used. To check the OS run 'imagex /info install.win'.

5.1.2.2 Configuring the Client Machine

➤ *To configuring your client:*

1. Verify the Mellanox adapter card is updated with the correct Mellanox FlexBoot version..

> Please refer to the firmware release notes to see if a particular firmware supports iSCSI boot or PXE capability.

For boot over Ethernet, when using adapter cards with older firmware version than 2.30.8000, you need to burn the adapter card with Ethernet FlexBoot, otherwise use the VPI FlexBoot.

2. Verify the Mellanox adapter card is burned with the correct firmware version.

3. Set the "Mellanox Adapter Card" as the first boot device in the BIOS settings boot order.

5.1.2.3 Installing iSCSI

1. Reboot your iSCSI client.

2. Press F12 when asked to proceed to iSCSI boot.

3. Choose the relevant boot image from the list of all available boot images presented.

4. Choose the Operating System to install.
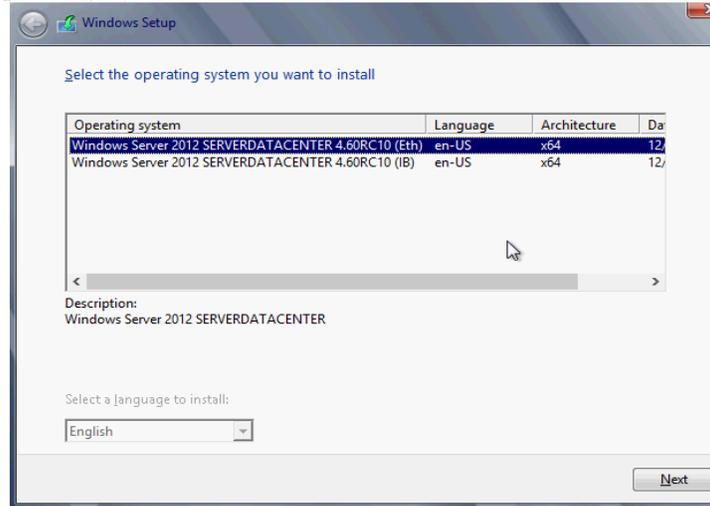
5. Run the Windows Setup Wizard.

6. Choose iSCSI target drive to install Windows and follow the instructions presented by the installation Wizard.
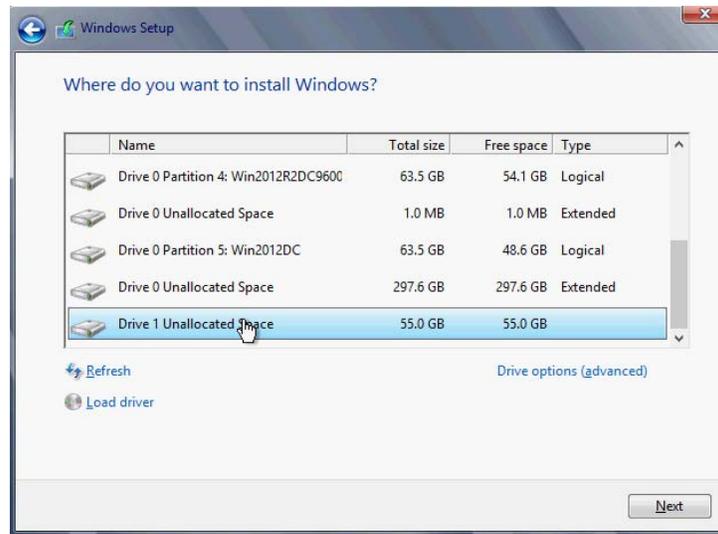
Installation process will start once completing all the required steps in the Wizard, the Client will reboot and will boot from the iSCSI target.

5.1.3 SLES11 SP3

5.1.3.1 Configuring the iSCSI Target Machine

➢ *To configure the iSCSI target:*

Step 1. Download the IET target software from.

http://sourceforge.net/projects/iscsitarget/files/iscsitarget/1.4.20.2/

Step 2. Install iSCSI target and additional required software on target server.

```
[root@sqa030 ~]# yum install kernel-devel openssl-devel gcc rpm-build
[root@sqa030 tmp]# tar xzvf iscsitarget-1.4.20.2.tar.gz
[root@sqa030 tmp]# cd iscsitarget-1.4.20.2/
[root@sqa030 iscsitarget-1.4.20.2]# make && make install
```

Step 3. Create the IQN in the ietd configuration file.

```
Target iqn.2013-10.qalab.com:sqa030.prt9
Lun 0 Path=/dev/cciss/c0d0p9,Type=fileio,IOMode=wb
MaxConnections          1        # Number of connections/session. We only support 1
InitialR2T              Yes      # Wait first for R2T
ImmediateData           Yes      # Data can accompany command
MaxRecvDataSegmentLength 8192    # Max data per PDU to receive
MaxXmitDataSegmentLength 8192    # Max data per PDU to transmit
MaxBurstLength          262144   # Max data per sequence (R2T)
FirstBurstLength        65536    # Max unsolicited data sequence
DefaultTime2Wait        2        # Secs wait for ini to log out    Not used
DefaultTime2Retain      20       # Secs keep cmnds after log out Not used
MaxOutstandingR2T       1        # Max outstanding R2Ts per cmnd
DataPDUInOrder          Yes      # Data in PDUs is ordered. We only support ordered
DataSequenceInOrder     Yes      # PDUs in sequence are ordered. We only support
                                       ordered
ErrorRecoveryLevel      0        # We only support level 0
HeaderDigest            NONE     # PDU header checksum algo list.  None or CRC32C
                                 # If only one is set then the initiator must agree
                                 # to it or the connection will fail
DataDigest              NONE     # PDU data checksum algo list     Same as above
MaxSessions             0        # Maximum number of sessions to this target 0 =
                                       unlimited
NOPInterval             0        # Send a NOP-In ping each after that many seconds
                                     if the
                                 # conn is otherwise idle  0 = off
NOPTimeout              0        # Wait that many seconds for a response on a
                                       NOP-In ping
                                 # If 0 or > NOPInterval, NOPInterval is used!
                                 # Various target parameters
Wthreads                8        # Number of IO threads
QueuedCommands          32       # Number of queued commands
```

> ⚠️ The local Hard Disk partition assigned to the LUN (/dev/cciss/c0d0p9 in the example above) must not contain any valuable data, as this data will be destroyed by the installation process taking place later in this procedure

Step 4. Edit the /etc/sysconfig/iscsi-target file as follow.

```
OPTIONS="-c /etc/iet/ietd.conf --address=12.7.6.30"
```

Step 5. Start the iSCSI target service.

```
[root@sqa030 ~]# /etc/init.d/iscsi-target start
```

Step 6. Perform a sanity check by connecting to the iSCSI target from a remote PC on the 10GE network link.

5.1.3.2 Configuring the DHCP Server

Edit a host-declaration for your PXE client in the DHCP configuration file, serving it with pxelinux.0, and restart your DHCP.

Here is an example of such host declaration inside DHCP config file:

```
host qadell011 {
        filename "pxelinux.0" ;
        next-server 12.7.6.30;
        option host-name "qadell011";
        fixed-address 12.7.6.11  ;
        hardware ethernet 00:02:C9:E5:D8:E0 ;
}
```

5.1.3.3 Installing SLES11 SP3 on a Remote Storage over iSCSI

Step 1. In the DHCP server configuration, the PXELINUX (pxelinux.0) and a SLES11 SP3 distribution media will be provided for network installation.

The clients' HDD was removed beforehand; therefore the installer will ask to locate a HDD. The built-in iSCSI discovery will be used to connect to the iSCSI target LUN partition.

Step 2. Reboot the client and invoke PXE boot with the Mellanox boot agent.
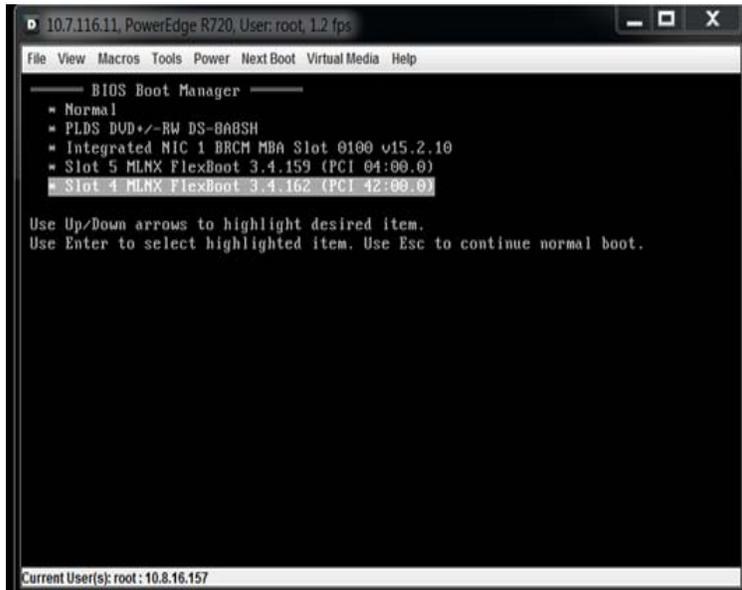
Step 3. Select the "Install SLES11.3" boot option from the menu (see pxelinux.cfg example above).

After about 30 seconds, the SLES installer will issue the notification below due to the PXELINUX boot label we used above.
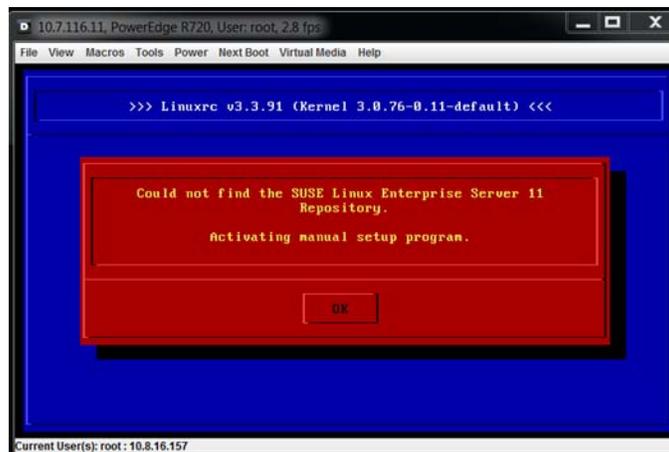
Step 4. Click OK.

Step 5. Click on the **Configure iSCSI Disks** button.

Step 6. Choose **Connected Targets** tab.

Step 7. Click **Add**.
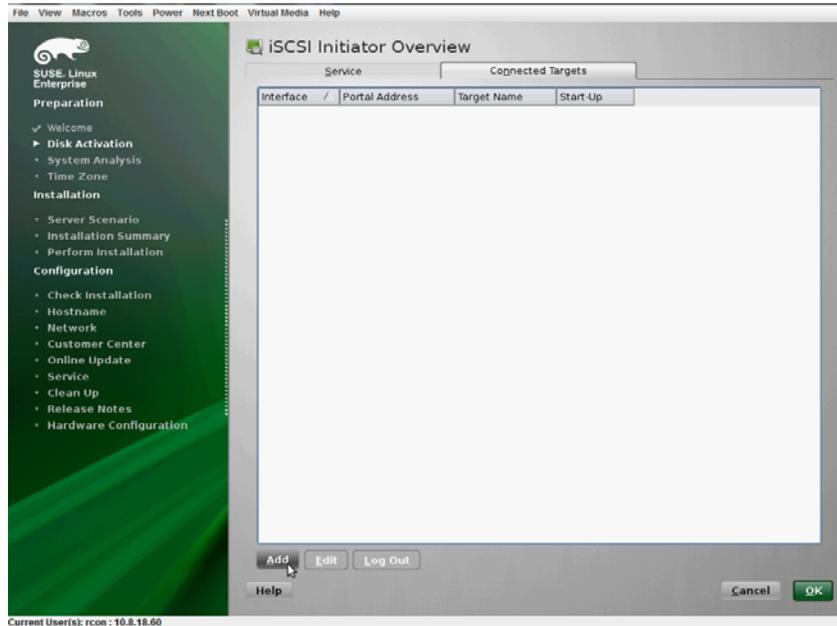
Step 8. Enter the IP address of the iSCSI storage target.

Step 9. Click **Next**.
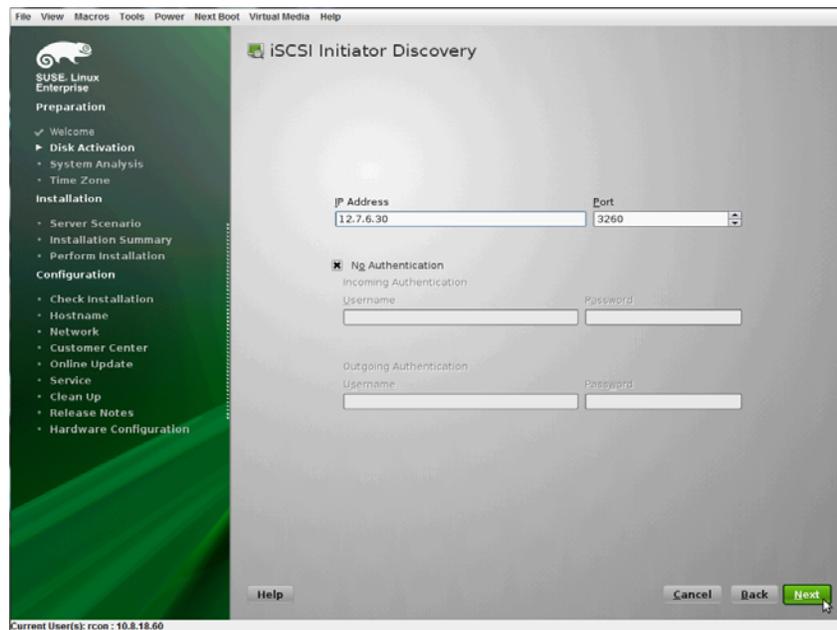
Step 10. Select the relevant target from the table (In our example, only one target exist so only one was discovered).

Step 11. Click **Connect**.
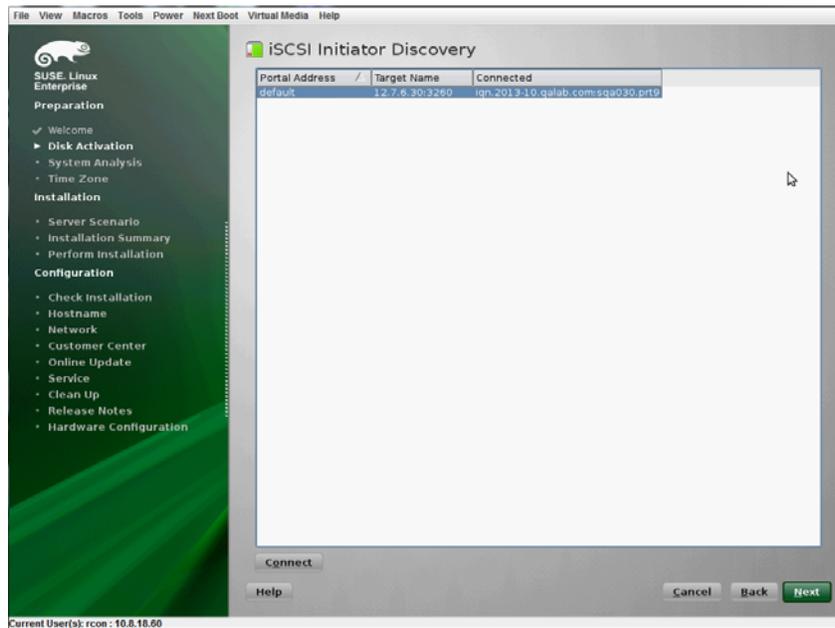
Step 12. Select **onboot** from drop-list.

Step 13. Click **Next** to exit the discovery screen.

Step 14. Go to the **Connected Targets** tab again to confirm iSCSI connection with target.
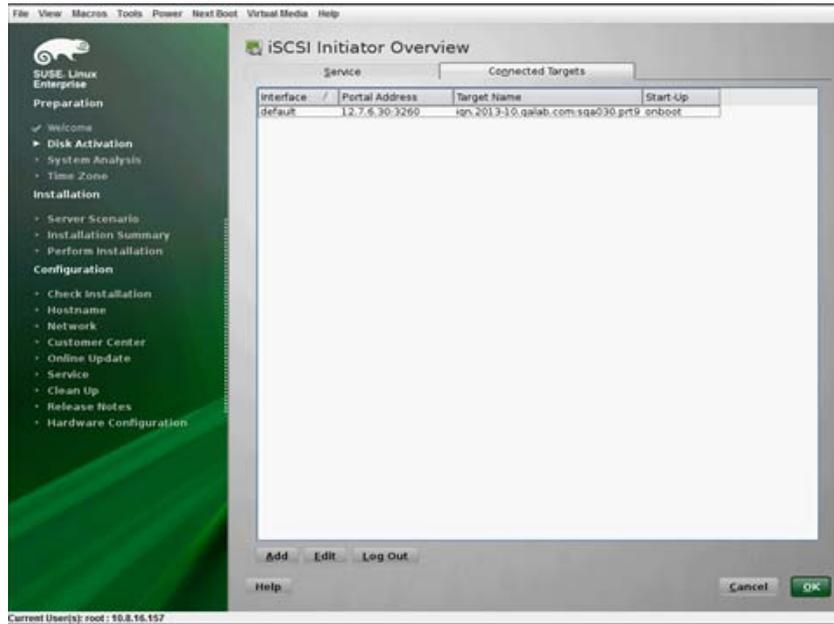
Step 15. Click **OK**.

Step 16. Click **Next** back at the Disk Activation screen.

Step 17. Select **New Installation**.

Step 18. Click **Next**.

Step 19. Complete **Clock and Time Zone** configuration.

Step 20. Select **Physical Machine**.

Step 21. Click **Next**.

Step 22. Click **Install**.
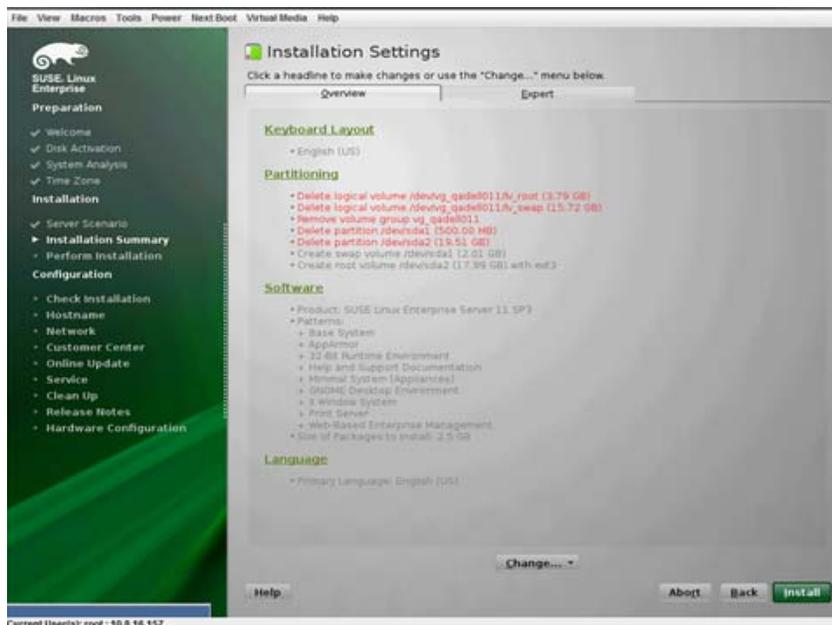
Make sure "`open-iscsi`" RPM is selected for the installation under "`Software`".

After the installation is completed, the system will reboot.

Choose "`SLES11.3x64_iscsi_boot`" label from the boot menu (See Section 5.1.3.3, "Installing SLES11 SP3 on a Remote Storage over iSCSI", on page 176).

Step 23. Complete post installation configuration steps.

> It is recommended to download and install the latest version of MLNX_OFED_LINUX available from
> http://www.mellanox.com/page/products_dyn?product_family=26&mtag=linux_sw_drivers

5.1.3.4 Using PXE Boot Services for Booting the SLES11 SP3 from the iSCSI Target

Once the installation is completed, the system will reboot. At this stage, it is expected from the client to perform another PXE network boot with FlexBoot®.

Choose the "SLES11.3x64 iSCSI boot" label from the boot menu (See Section 5.1.3.3, "Installing SLES11 SP3 on a Remote Storage over iSCSI", on page 176).

> In firmware version 2.33.50.50 and prior releases, the iSCSI boot initiator TCP/IP default parameters resulted in DHCP being enabled by default for the initiator. This can be determined because all the TCP fields in "iSCSI Initiator Parameters" page are blank. In firmware version 2.34.50.60 and newer releases, iSCSI boot initiator DHCP behavior is controlled by "TCP/IP Parameters via DHCP" on "iSCSI General Parameters" page and is disabled by default. Similarly, the iSCSI boot target DHCP behavior is controlled by "iSCSI Parameters via DHCP" on "iSCSI General Parameters" page and is disabled by default.

> iSCSI boot configuration parameters will be retained when upgrading to firmware version 2.34.50.60. When downgrading from firmware version 2.34.50.60 to older firmware versions, the differences in how the fields are in interpreted relative to DHCP usage for initiator and target may result in a change in iSCSI boot behavior. Users are advised to verify iSCSI boot parameters after such a downgrade.

5.2 PXE Boot

5.2.1 SLES11 SP3

5.2.1.1 Configuring the PXE Server

Step 1. Download SLES11SP3-kISO-VPI.tgz from
http://www.mellanox.com/page/products_dyn?&product_family=34&mtag=flexboot

Step 2. Extract the .tgz file on the PXE server, under the TFTP root directory.

For example:

```
[root@sqa030 ~]# cd /var/lib/tftpboot;  tar xzvf SLES11SP3-kISO-VPI.tgz
```

The following are examples of PXE server configuration:

• PXE server configuration:

```
SLES11SP3-kISO-VPI/pxelinux.cfg/default
```

• Kernel and initrd for the installation:

```
SLES11SP3-kISO-VPI/pxeboot-install/initrd
SLES11SP3-kISO-VPI/pxeboot-install/linux
```

• Kernel and initrd for the boot after the installation:

```
SLES11SP3-kISO-VPI/pxeboot/initrd
SLES11SP3-kISO-VPI/pxeboot/linux
```

• kISO installation medium that can be used to boot from instead of booting the installer program over the network.

If choosing this method:

• Boot the client into the below SLES11 SP3 iso and proceed with the installation until the client fully boots up the installer program.

• Discover and connect to a remote iSCSI storage.
During the installation process you will be asked to insert the original installation medium to continue with the installation.

```
SLES11SP3-kISO-VPI/sles11-sp3-x86_64-mlnx_ofed-2.1-1.0.6.iso
```

If the iso method above is not used, two different PXE server configurations are required (PXELINUX booting labels) for each phase discussed herein (booting the installer and post-installation boot)

• For booting the installer program off the TFTP server, please provide the client a path to the initrd and linux kernel as provided inside SLES11SP3-kISO-VPI/pxeboot-install/ in the tgz above.

The below is an example of such label.

```
LABEL SLES11.3x64_manual_installl
MENU LABEL ^1) Install SLES11.3
kernel SLES11SP3-kISO-VPI/pxeboot-install/linux
append initrd=SLES11SP3-kISO-VPI/pxeboot-install/initrd install=nfs://12.7.6.30/pxerepo/
SLES/11.3/x86_64/DVD1/?device=p4p2
IPAPPEND 2
```

- For post-installation boot (booting the SLES 11 SP3 off the iSCSI storage using PXE services) please provide the booting client a path to the initrd and linux kernel as provided inside SLES11SP3-kISO-VPI/pxeboot/ in the tgz above.

The below is an example of such label.

```
LABEL SLES11.3x64_iscsi_boot
MENU LABEL ^2) SLES11.3 iSCSI boot
kernel SLES11SP3-kISO-VPI/pxeboot/linux
append initrd= SLES11SP3-kISO-VPI/pxeboot/initrd net-root=iscsi:12.7.6.30::::iqn.2013-
10.qalab.com:sqa030.prt9 TargetAd-dress=12.7.6.30 TargetName=iqn.2013-
10.qalab.com:sqa030.prt9 TargetPort=3260 net_delay=10 rootfstype=ext3 rootdev=/dev/sda2
```

The steps described in this document do not refer to an unattended installation with autoyast. For official information on SLES unattented installation with autoyast, please refer to:

https://www.suse.com/documentation/sles11/book_autoyast/?page=/documentation/sles11/book_autoyast/data/book_autoyast.html

The following is known to work with Mellanox NIC:

```
append initrd=SLES-11-SP3-DVD-x86_64-GM-DVD1/boot/x86_64/loader/initrd install=nfs://
<NFS IP Address>/<path the the repository directory>/ autoyast=nfs://<NFS IP Address>/
<path to autoyast xml directory>/autoyast-unattended.xml biosdevname=0
IPAPPEND 2
```

6 Firmware

Firmware and update instructions for these cards can be obtained from the Dell support web page: http://www.dell.com/support.

Note: The firmware version on the adapter can be checked using the following methods:

1. System Setup > Device Settings
2. Dell iDRAC

6.1 Linux Firmware Update Package

For Linux, download the latest Linux Firmware Update Package for Mellanox ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex Ethernet adapters firmware package available at Dell's support site http://www.dell.com/support.

Run the binary package from the command line and follow the instructions to install Mellanox firmware.

6.2 Windows Firmware Update Package

For Windows, download and install the latest Windows Drivers for Mellanox ConnectX-3, ConnectX-3 Pro, ConnectX-4, ConnectX-4 Lx and ConnectX-5 Ex Ethernet adapters software package available at Dell's support site http://www.dell.com/support.

Run the binary package from the command line and follow the instructions to install Mellanox firmware

```
.\Network_Firmware_TPWYK_WN64_14.17.20.52.EXE
```

6.3 Updating Firmware using Dell iDRAC or Lifecycle Controller

Follow the below steps for updating firmware via iDRAC or Lifecycle Controller. These steps are useful when updating with Extended Secure Boot enabled or for operating systems which do not have firmware tools support.

6.3.1 Updating Firmware Using Dell Lifecycle Controller

Step 1. On boot, Press F10 to enter Lifecycle Controller

Step 2. Select "Firmware Update"

Step 3. Launch Firmware Update

Step 4. Connect USB Drive with Firmware DUP to the system

Step 5. Select "Local Drive (CD, DVD or USB)"

Step 6. Click "Next"

Step 7. Enter Path and Filename of the DUP

Step 8. Click "Next".

Step 9. Select Devices to update.

6.3.2 Updating Firmware Using Dell iDRAC

Step 1. Log into iDRAC GUI.
For iDRAC9 go to Maintenance => System Update.
For iDRAC8, go to "Update and Rollback" under "Quick Launch Tasks"

Step 2. Enter path of Dell update package

Step 3. Click "Upload"

Step 4. Follow on-screen instructions.

7 Troubleshooting

7.1 General

Server unable to find the adapter	• Ensure that the adapter is placed correctly • Make sure the adapter slot and the adapter are compatible • Install the adapter in a different PCI Express slot • Use the drivers that came with the adapter or download the latest • Make sure your motherboard has the latest BIOS • Try to reboot the server
The adapter no longer works	• Reseat the adapter in its slot or a different slot, if necessary • Try using another cable • Reinstall the drivers for the network driver files may be damaged or deleted • Reboot the server
Adapters stopped working after installing another adapter	• Try removing and re-installing all adapters • Check that cables are connected properly • Make sure your motherboard has the latest BIOS
Link indicator light is off	• Ensure that adapter driver/s is loaded • Try another port on the switch • Make sure the cable is securely attached • Check your are using the proper cables that do not exceed the recommended lengths • Verify that your switch and adapter port are compatible
Link light is on, but with no communication established	• Check that the latest driver is loaded • Check that both the adapter and its link are set to the same speed and duplex settings
Low Performance with RDMA over Converged Ethernet (RoCE)	• Check to make sure flow-control is enabled on the switch ports.
HII fails in server with UEFI 2.5	• For HII to work in servers with UEFI 2.5, the following are the minimum required firmware versions: • For ConnectX-3 Pro: version 2.40.50.48 • For ConnectX-4: version 12.17.20.52 • For ConnectX-4 Lx: version 14.17.20.52 • For ConnectX-5 Ex: version 16.23.1020
UEFI Secure Boot Known Issue	On RHEL and SLES12, the following error is displayed in dmesg if the Mellanox's x.509 Public Key is not added to the system: [4671958.383506] Request for unknown module key 'Mellanox Technologies signing key: 61feb074fc7292f958419386ffdd9d5ca999e403' err -11 For further information, please refer to the User Manual section "Enrolling Mellanox's x.509 Public Key On your System".

Slow PXE TFTP download	In ConnectX-4/ConnectX-4 Lx adapters, there is a delay in the PXE boot ROM data flow code which causes poor performance when downloading an operating system image from a TFTP PXE server in 14th Generation Dell EMC PowerEdge Servers. This has been resolved in firmware version 12.20.18.20 for ConnectX-4 and 14.20.18.20 for ConnectX-4 Lx (Flexboot 3.5.214) and newer.

7.2   Linux

Firmware Version Upgrade	To download the latest firmware version refer to the Dell support site http://www.dell.com/support
Environment Information	cat/etc/issue uname –a cat/proc/cupinfo | grep 'model name' | uniq ofed_info | head -1 ifconfig –a ethtool <interface> ethtool –i <interface_of_Mellanox_port_num> ibdev2netdev
Card Detection	lspci | grep –i Mellanox
Ports Information	ibstat lbv_devinfo
Collect Log File	cat /var/log/messages dmesg > system.log
Insufficient memory to be used by udev upon OS boot	Limit the udev instances running simultaneously per boot by adding udev.children-max=<number> to the kernel command line in grub. This is seen with SLES12 SP2 and SP3.
ConnectX-3 Pro adapter cards fail to load after installing RHEL 6.9	Re-install the "rdma" package after removing MLNX_OFED. To install the "rdma" package, run: "yum install rdma"

7.3   Windows

Firmware Version Upgrade	To download the latest firmware version refer to the Dell support site http://www.dell.com/support
Performance is Low	This can be due to non-optimal system configuration. See the section "Performance Tuning" to take advantage of Mellanox 10 GBit NIC performance
Driver Does Not Start	This can happen due to an RSS configuration mismatch between the TCP stack and the Mellanox adapter. To confirm this scenario, open the event log and look under "System" for the "mlx4eth5" or "mlx-4eth6" source. If found, enable RSS as follows: Run the following command: "netsh int tcp set global rss = enabled".
	This is a less recommended suggestion, and will cause low performance. Disable RSS on the adapter. To do this set RSS mode to "No Dynamic Rebalancing".
The Ethernet Driver Fails to Start	If in the Event log, under the mlx4_bus source, the following error message appears: RUN_FW command failed with error -22, this error message indicates that the wrong firmware image has been programmed on the adapter card.
	If a yellow sign appears near the Mellanox ConnectX Ethernet Adapter instance in the Device Manager display, this can happen due to a hardware error. Try to disable and re-enable "Mellanox ConnectX Adapter" from the Device Manager display.
No connectivity to a Fault Tolerance bundle while using network capture tools (e.g., Wireshark)	This can happen if the network capture tool captures the network traffic of the non-active adapter in the bundle. This is not allowed since the tool sets the packet filter to "promiscuous", thus causing traffic to be transferred on multiple interfaces. Close the network capture tool on the physical adapter card, and set it on the LBFO interface instead.
No Ethernet connectivity on 1Gb/100Mb adapters after activating Performance Tuning (part of the installation)	This can happen due to adding a TcpWindowSize registry value. To resolve this issue, remove the value key under HKEY_LO-CAL_MACHINE\SYSTEM\CurrentControl- Set\Ser-vices\Tcpip\Parameters\TcpWindowSize or set its value to 0xFFFF
System reboots on an I/O AT capable system	This may occur if you have an Intel I/OAT capable system with Direct Cache Access enabled, and 9K jumbo frames enabled. To resolve this issue, disable 9K jumbo frames.
Packets are being lost	This may occur if the port MTU has been set to a value higher than the maximum MTU supported by the switch.
BSOD when installing Windows Server 2016 on iSCSI LUN in UEFI boot mode	A BSOD may occur when trying to UEFI iSCSI boot from port 2 using Windows Server 2016 with Inbox or WinOF-2 1.80 drivers. This has been resolved in WinOF-2 1.90 and newer drivers.

RDMA Connection refusal	The operating system fails to create an NDK listener during driver load due to an address conflict error. This causes a failure to listen on this interface until another notification requesting to listen on that interface is received (i.e. an RDMA enable event, an IP arrival event, a system bind event, etc.).
	Workaround: unbind and rebind TCP/IPv4 protocol.
	The following is an example:
	`Set-NetAdapterBinding -Name MyAdapter -Dis-`
	`playName "Internet Protocol Version 4 (TCP/`
	`IPv4)" -Enabled $false`
	`Set-NetAdapterBinding -Name MyAdapter -Dis-`
	`playName "Internet Protocol Version 4 (TCP/`
	`IPv4)" -Enabled $true`

AMD EPYC based Systems Requirement	For AMD EPYC based system, WinOF-2 version 1.70 and higher are required.
Windows System Event Viewer Events	Failed to reset the Mellanox ConnectX EN 10Gbit Ethernet NIC. Try disabling then re-enabling the "Mellanox Ethernet Bus Driver" device via the Windows device manager.
	Failed to initialize the Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> because it uses old firmware version (<old firmware version>). You need to update the adapter to firmware version <new firmware version> or higher, and to restart your computer.
	Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> device detected that the link connected to port <Y> is down. This can occur if the physical link is disconnected or damaged, or if the other end-port is down.
	Mismatch in the configurations between the two ports may affect the performance. When Using MSI-X, both ports should use the same RSS mode. To fix the problem, configure the RSS mode of both ports to be the same in the driver GUI.
	Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> device failed to create enough MSI-X vectors. The Network interface will not use MSI-X interrupts. This may affects the performance. To fix the problem, configure the number of MSI-X vectors in the registry to be at least <Y>.

Driver fails to load on a server with more than 4TB Ram	To load a server with more than 4TB Ram, set the following registry key:
	    Reg path: HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\mlx4_bus\Parameters
	    Reg key name: LogNumMttOverride
	    Reg key type: DWORD
	    Reg key value: 1
	    Restart the device.

Operating system crashes when unloading inbox drivers in systems with more than 64 cores (or 128 logical cores)	For loading inbox drivers in system with more than 64 cores, reduce the number of cores via BIOS. Then, inject out of box drivers (WinOF 5.22 or 5.25) during Operating System installation.

8   Specifications

Table 30 - Mellanox ConnectX-3 Dual 40GbE QSFP+ Network Adapter Specifications

Physical	**Board Size:** 2.71in. x 5.6in. (68.90 mm x 142.25 mm) **Full height Bracket Size:** 4.5 in. (116 mm) **Low profile Bracket Size:** 3.16 in. (80.3 mm)		
	Connector: QSFP+ 40Gb/s		
Protocol Support	**Ethernet:** 10GBASE-CR, 40GBASE-CR4 /-SR4		
	Data Rate: 10/40Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8lanes (2.0 and 1.1 compatible)		
Power and Environmental[a]	**Voltage:** 12V, 3.3V, 3.3V_AUX		
	Power	**Cable Type**	
	Typical Power	**Passive Cables**	7.34W
		Optical Cables	8.87W
	Maximum Power	**Passive Cables**	8.87W
		Optical Cables	10.64W
	Temperature[b]	**Operational**	0°C to 55°C
		Non-operational	0°C to 70°C
	Humidity: 90% relative humidity [c]		
	Air Flow: 120LFM		
Regulatory	**EMC:** Refer to Chapter 8.2, "Regulatory Statements," on page 204		

	Safety: IEC/EN 60950-1:2006
	            ETSI EN 300 019-2-2
	            IEC 60068-2- 64, 29, 32		

	RoHS: RoHS-R6		

a. Thermal and power characteristics with optical modules only supported with Mellanox QSFP+ optical module, MC2210411-SR4 (Dell Part Number 2MJ5F)
b. Thermal spec covers Power Level 1 QSFP modules
c. For both operational and non-operational states

Table 31 - Mellanox ConnectX-3 Dual 10GbE SFP+ Network Adapter Specifications

Physical	**Size:** 2.71in. x 5.6in. (68.90 mm x 142.25 mm) **Full height Bracket Size:** 3.8in (96.52 mm) **Low profile Bracket Size:** 3.16in (80.3 mm)		
	Connector: SFP+ 10Gb/s		
Protocol Support	**Ethernet:** 10GBASE-CR, 10GBASE-R, and 1000BASE-R Note: No 1Gb/s optics are currently supported.		
	Data Rate:10Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8lanes (2.0 and 1.1 compatible)		
Power and Environmental[a]	**Voltage:** 12V, 3.3V. 3.3V_AUX		
	Power	**Cable Type**	
	Typical Power	**Passive Cables**	5.11W
		Optical Cables	6.75W
	Maximum Power	**Passive Cables**	6.14W
		Optical Cables	8.5W
	Temperature[b]	**Operational**	0°C to 55°C
		Non-operational	0°C to 70°C
	Humidity: 90% relative humidity[c]		
	Air Flow: 100LFM[d]		
Regulatory	**EMC:** Refer to Chapter 8.2, "Regulatory Statements," on page 204.		

	Safety: IEC/EN 60950-1:2006
	          ETSI EN 300 019-2-2
	          IEC 60068-2- 64, 29, 32		

	RoHS: RoHS-R6		

a. Thermal and power characteristics with optical modules only supported with Mellanox SFP+ optical module, MFM1T02A-SR (Dell Part Number T16JY).
b. Thermal spec covers Power Level 1 SFP+ modules
c. For both operational and non-operational states
d. Air flow is measured ~1" from the heat sink between the heat sink and the cooling air inlet.

Table 32 - Mellanox ConnectX-3 Dual 10GbE KR Blade Mezzanine Card Specifications

Physical	**Size:** 3.4in. x 3.3in. (88 mm x 84mm)		
Protocol Support	**Ethernet:** 10GBASE-KR, 10GBASE-KX4, 1000BASE-KX Note: No 10GBASE-KX4 IOMs are currently supported.		
	Data Rate: 1/10Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8 lanes (2.0 and 1.1 compatible)		
Power and Environmental	**Voltage:** 12V, 3.3Vaux		
	Typical Power	4.84W	
	Maximum Power	5.87W	
	Temperature	**Operational**	0°C to 65°C
		Non-operational	0°C to 70°C
	Humidity: 90% relative humidity[a]		
	Air Flow: 200LFM[b]		
Regulatory	**Safety:** IEC/EN 60950-1:2006 ETSI EN 300 019-2-2 IEC 60068-2- 64, 29, 32		
	RoHS: RoHS-R6		

a. For both operational and non-operational states
b. Air flow is measured ~1" from the heat sink between the heat sink and the cooling air inlet.

Table 33 - Mellanox ConnectX-3 Pro Dual 40GbE QSFP+ Network Adapter Specifications

Physical	**Board Size:** 2.71in. x 5.6in. (68.90 mm x 142.25 mm) **Full height Bracket Size:** 4.5 in. (116 mm) **Low profile Bracket Size:** 3.16 in. (80.3 mm)		
	Connector: QSFP+ 40Gb/s		
Protocol Support	**Ethernet:** 10GBASE-CR, 40GBASE-CR4 /-SR4		
	Data Rate: 10/40Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8lanes (2.0 and 1.1 compatible)		
Power and Environmental[a]	**Voltage:** 12V, 3.3V, 3.3V_AUX		
	Power	**Cable Type**	
	Typical Power	**Passive Cables**	7.57W
		Optical Cables	9.10W
	Maximum Power	**Passive Cables**	9.11W
		Optical Cables	10.87W
	Temperature[b]	**Operational**	0°C to 55°C
		Non-operational	0°C to 70°C
	Humidity: 90% relative humidity [c]		
	Air Flow: 120LFM[d]		
Regulatory	**EMC:** Refer to Chapter 8.2, "Regulatory Statements," on page 204		

	Safety: IEC/EN 60950-1:2006
	    ETSI EN 300 019-2-2
	    IEC 60068-2- 64, 29, 32		

	RoHS: RoHS-R6		

a. Thermal and power characteristics with optical modules only supported with Mellanox QSFP+ optical module, MC2210411-SR4 (Dell Part Number 2MJ5F)
b. Thermal spec covers Power Level 1 QSFP modules
c. For both operational and non-operational states
d. Air flow is measured ~1" from the heat sink between the heat sink and the cooling air inlet.

Table 34 - Mellanox ConnectX-3 Pro Dual 10GbE SFP+ Network Adapter Specifications

Physical	**Size:** 2.71in. x 5.6in. (68.90 mm x 142.25 mm) **Full height Bracket Size:** 3.8in (96.52 mm) **Low profile Bracket Size:** 3.16in (80.3 mm)
	Connector: SFP+ 10Gb/s
Protocol Support	**Ethernet:** 10GBASE-CR, 10GBASE-R, and 1000BASE-R Note: No 1Gb/s optics are currently supported.
	Data Rate:10Gb/s – Ethernet
	PCI Express Gen3: SERDES @ 8.0GT/s, 8lanes (2.0 and 1.1 compatible)

Voltage: 12V, 3.3V. 3.3V_AUX

Power	Cable Type	
Typical Power	**Passive Cables**	5.45W
	Optical Cables	7.1W
Maximum Power	**Passive Cables**	6.41W
	Optical Cables	8.76W
Temperature[b]	**Operational**	0°C to 55°C
	Non-operational	0°C to 70°C

Power and Environmental[a]

Humidity: 90% relative humidity[c]

Air Flow: 100LFM[d]

Regulatory	**EMC:** Refer to Chapter 8.2, "Regulatory Statements," on page 204
	Safety: IEC/EN 60950-1:2006 ETSI EN 300 019-2-2 IEC 60068-2- 64, 29, 32
	RoHS: RoHS-R6

a. Thermal and power characteristics with optical modules only supported with Mellanox SFP+ optical module, MFM1T02A-SR (Dell Part Number T16JY).
b. Thermal spec covers Power Level 1 SFP+ modules
c. For both operational and non-operational states
d. Air flow is measured ~1" from the heat sink between the heat sink and the cooling air inlet.

Table 35 - Mellanox ConnectX-3 Pro Dual 10GbE KR Blade Mezzanine Card Specifications

Physical	Size: 3.4in. x 3.3in. (88 mm x 84mm)		
Protocol Support	Ethernet: 10GBASE-KR, 10GBASE-KX4, 1000BASE-KX Note: No 10GBASE-KX4 IOMs are currently supported.		
	Data Rate: 1/10Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8 lanes (2.0 and 1.1 compatible)		
Power and Environmental	Voltage: 12V, 3.3Vaux		
	Power	Cable Type	
	Typical Power	Passive Cables	5.04W
	Maximum Power	Passive Cables	6.13W
	Temperature	Operational	0°C to 65°C
		Non-operational	0°C to 70°C
	Humidity: 90% relative humidity[a]		
	Air Flow: 200LFM[b]		
Regulatory	EMC: Refer to Chapter 8.2, "Regulatory Statements," on page 204		

	Safety: IEC/EN 60950-1:2006
	       ETSI EN 300 019-2-2
	       IEC 60068-2- 64, 29, 32		

	RoHS: RoHS-R6		

a. For both operational and non-operational states
b. Air flow is measured ~1" from the heat sink between the heat sink and the cooling air inlet.

Table 36 - Mellanox ConnectX-4 Dual Port 100 GbE QSFP Network Adapter Specifications

Physical	**Size:** 2.71 in. x 5.6 in. (68.90mm x 142.24mm) **Full height Bracket Size**: 4.76 in. (121 mm) **Low profile Bracket Size:** 1.036 in. (79.3 mm)		
	Connector: Dual QSFP28 (Copper Cables)		
Protocol Support	Ethernet:100GBASE-CR4, 50GBASE-R2, 50GBASE-R4, 40GBASE-CR4, 25GBASE-CR, 10GBASE-CR		
	Data Rate: 1/10/25/40/100Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 16 lanes (2.0 and 1.1 compatible)		
Power and Environmental	**Voltage:** 12V, 3.3VAUX		
	Power	**Cable Type**	
	Typical Power[a]	**Passive Cables**	16.3W
		Optical Cables	21.9W
	Maximum Power	**Passive Cables**	19.0W
		Optical Cables	24.5W
	Temperature	**Operational**	0°C to 55°C
		Non-operational	-40°C to 70°C
	Humidity: 90% relative humidity [b]		
	Air Flow: 300LFM[c]		
Regulatory	**Safety**: CB / cTUVus / CE		
	EMC: Refer to Chapter 8.2, "Regulatory Statements," on page 204		
	RoHS: RoHS-R6		

a. Typical power for ATIS traffic load
b. For both operational and non-operational states
c. For Passive Cables only. Air flow is measured ~1" from the heat sink to the port.

Table 37 - Mellanox ConnectX-4 Lx Dual Port SFP28 25GbE for Dell Rack NDC

Physical	**Size:** 3.66 in. x 4.33 in. (93mm x 110mm)		
	Connector: Dual SFP28 (Copper Cables)		
Protocol Support	**Ethernet:** 25GBASE-CR, 10GBASE-CR		
	Data Rate: 10/25 Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8 lanes (2.0 and 1.1 compatible)		
Power and Environmental	**Voltage:** 12V, 3.3V		
	Power	**Cable Type**	
	Typical Power[a]	**Passive Cables**	9.8W
		Optical Cables	13.1W
	Maximum Power	**Passive Cables**	11.8W
		Optical Cables	15.2W
	Temperature	**Operational**	0°C to 60°C
		Non-operational	-40°C to 70°C
	Humidity: 90% relative humidity [b]		
	Air Flow: 100LFM		
Regulatory	**Safety:** CB / cTUVus / CE		
	EMC: Refer to Chapter 8.2, "Regulatory Statements," on page 204		
	RoHS: RoHS-R6		

a. Typical power for ATIS traffic load
b. For both operational and non-operational states

Table 38 - Mellanox ConnectX-4 Lx Dual 25GbE SFP28 Network Adapter Specifications

Physical	**Size:** 2.71 in. x 5.6 in. (68.90mm x 142.24mm) **Full height Bracket Size:** 4.5 in. (116 mm) **Low profile Bracket Size:** 3.16 in. (80.3 mm)		
	Connector: Dual SFP28 (Copper Cables)		
Protocol Support	**Ethernet:** 25GBASE-CR, 10GBASE-CR		
	Data Rate: 10/25 Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8 lanes (2.0 and 1.1 compatible)		
Power and Environmental	**Voltage: 12V, 3.3V**		
	Power	**Cable Type**	
	Typical Power[a]	**Passive Cables**	9.6W
		Optical Cables	12.9W
	Maximum Power	**Passive Cables**	11.6W
		Optical Cables	15.0W
	Temperature	**Operational**	0°C to 55°C
		Non-operational	-40°C to 70°C
	Humidity: 90% relative humidity [b]		
	Air Flow: 100LFM		
Regulatory	**Safety**: CB / cTUVus / CE		
	EMC: Refer to Chapter 8.2, "Regulatory Statements," on page 204		
	RoHS: RoHS-R6		

a. Typical power for ATIS traffic load
b. For both operational and non-operational states

Table 39 - Mellanox ConnectX-4 Lx Dual Port 25GbE KR Mezzanine Card Specifications

Physical	**Size:** 3.17in. x 3.14in. (80.60mm x 80.00mm)		
	Connector: Two IMPEL connectors to Two PTMs or Switch Modules		
Protocol Support	**Ethernet:** 25GBASE-KR, 10GBASE-KR		
	Data Rate: 10/25 Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 8 lanes (2.0 and 1.1 compatible)		
Power and Environmental	**Voltage:** 12V		
	Typical Power[a]	10.4W	
	Maximum Power:	13.0W	
	Temperature	**Operational**	0°C to 65°C
		Non-operational	-40°C to 70°C
	Humidity: 90% relative humidity[b]		
	Air Flow	**At 55C**	**At 60C**	**At 65C**
		100LFM	150LFM	200LFM
Regulatory	**Safety:** CB / cTUVus / CE		
	EMC: ACMA / IC / CE / VCCI / KC / FCC		
	RoHS: RoHS-R6		

a. Typical power for ATIS traffic load.
b. For both operational and non-operational states.

Table 40 - Mellanox ConnectX-5 Ex Dual Port 100GbE QSFP Network Adapter Specifications

Physical	**Size:** 2.71 in. x 5.6 in. (68.90mm x 142.24mm) **Full height Bracket Size**: 4.76 in. (121 mm) **Low profile Bracket Size:** 1.036 in. (79.3 mm)		
	Connector: Dual QSFP28 (Copper Cables)		
Protocol Support	Ethernet:100GBASE-CR4, 50GBASE-R2, 50GBASE-R4, 40GBASE-CR4, 25GBASE-CR, 10GBASE-CR		
	Data Rate: 1/10/25/40/100Gb/s – Ethernet		
	PCI Express Gen3: SERDES @ 8.0GT/s, 16 lanes (2.0 and 1.1 compatible)		
Power and Environmental	**Voltage:** 12V		
	Power	**Cable Type**	**Airflow (LFM)[a]**
	Typical Power[b]	**Passive Cables**	18.0W	300
	Maximum Power	**Passive Cables**	22.0W	350
	Temperature	**Operational**	0°C to 55°C
		Non-operational	-40°C to 70°C
	Humidity: 90% relative humidity [c]		
Regulatory	**Safety**: CB / cTUVus / CE		
	EMC: Refer to Chapter 8.2, "Regulatory Statements," on page 204		
	RoHS: RoHS-R6		

a. For Passive Cables only. Air flow is measured ~1" from the heat sink to the port.
b. Typical power for ATIS traffic load
c. For both operational and non-operational states

8.1   Regulatory

Table 41 - Ethernet Network Adapter Certifications

OPN	FCC	VCCI	EN	ICES	CE	CB	cTUVus	KCC	CCC	GOST-R	S-MARK	RCM
Mellanox ConnectX-3 Dual 40GbE QSFP+ Network Adapter	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-3 Dual 10GbE SFP+ Network Adapter	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-3 Pro Dual 40GbE QSFP+ Network Adapter	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-3 Pro Dual 10GbE SFP+ Network Adapter	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-4 Dual Port 100GbE QSFP Network Adapter	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-4 Lx EN Dual Port SFP28, 25GbE for Dell rack NDC	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-4 Lx Dual 25GbE SFP28 Network Adapter Specifications	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX-4 Lx Dual Port 25GbE KR Mezzanine Card	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES
Mellanox ConnectX®-5 Ex Dual Port 100GbE QSFP Network Adapter	YES	YES	YES	YES	YES	YES	YES	YES	Exemption letter	N/A	N/A	YES

8.2 Regulatory Statements

8.2.1 FCC Statements (USA)

Class A Statements:

§ 15.19(a)(4)

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.

§ 15.21

Statement

Warning!

Changes or modifications to this equipment not expressly approved by the party responsible for compliance (Mellanox Technologies) could void the user's authority to operate the equipment.

§15.105(a)

Statement

NOTE: This equipment has been tested and found to comply with the limits for a Class A digital device, pursuant to Part 15 of the FCC Rules. These limits are designed to provide reasonable protection against harmful interference when the equipment is operated in a commercial environment. This equipment generates, uses, and can radiate radio frequency energy and, if not installed and used in accordance with the instruction manual, may cause harmful interference to radio communications. Operation of this equipment in a residential area is likely to cause harmful interference in which case the user will be required to correct the interference at his own expense.

8.2.2 EN Statements (Europe)

EN55022 Class A Statement:

Warning

This is a class A product. In a domestic environment this product may cause radio interference in which case the user may be required to take adequate measures.

8.2.3 ICES Statements (Canada)

Class A Statement:

CAN ICES-3 (A)/NMB-3(A)

8.2.4   VCCI Statements (Japan)

Class A Statement:

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

(Translation - "This is a Class A product. In a domestic environment, this product may cause radio interference, in which case the user may be required to take corrective actions..")

8.2.5   KCC Certification (Korea)

Class A Statement:

A급 기기
(업무용 방송통신기자재

이 기기는 업무용(A급) 전자파적합기기로서 판매자 또는 사용자는 이 점을 주의하시기 바라며, 가정외의 지역에서 사용하는 것을 목적으로 합니다.

(Translation: "This equipment has been tested to comply with the limits for a Class A digital device. This equipment should be operated in a commercial environment. Please exchange if you purchased this equipment for noncommercial purposes")

Appendix A: Configuration for Mellanox Adapters through System Setup

This section covers the main configuration options in Dell EMC PowerEdge System Setup which can be accessed through BIOS or through Lifecycle Controller.

Figure 20: System Setup Menu
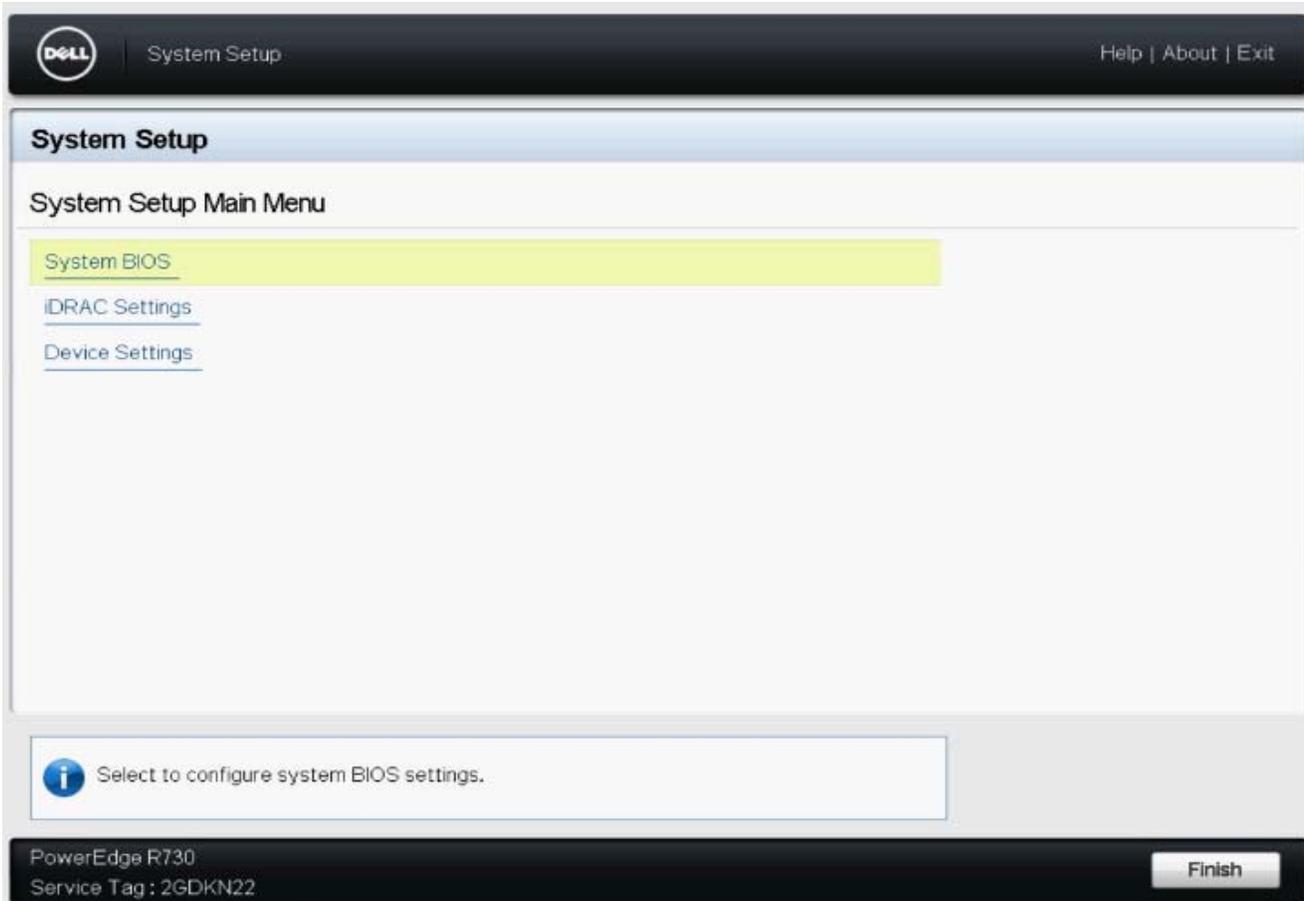
Setup menu options:

1. System BIOS
2. iDRAC Settings
3. Device Settings

To configure the Mellanox Adapter, choose 'Device settings' and the relevant Mellanox adapter:

Figure 21: Main Configuration Page Options

1. Shows general info regarding the adapter.
2. Allows configuration of SR-IOV on the adapter - see Appendix A.5, "SR-IOV Configuration," on page 216.
3. Allows setting the Blink LEDs to allow physical identification of the card

a. To trigger Blink LEDs configure the number of seconds for it to blink (Max is 15 seconds).
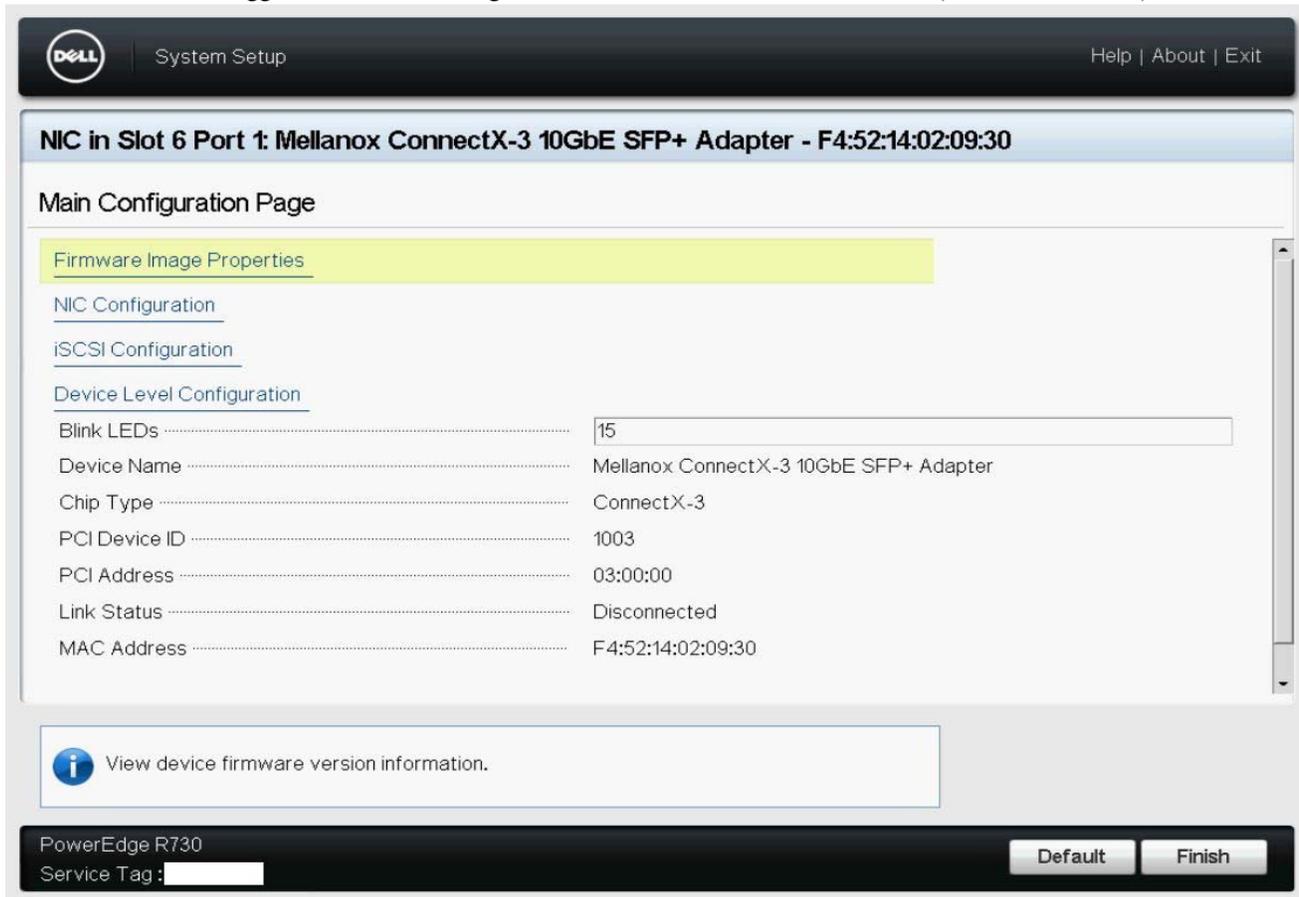
A.1 Main Configuration Page - Firmware Image Properties

The below provides Firmware the uEFI versions numbers.[1]

---

1. These version numbers are just an example.

A.2 Main Configuration Page - NIC Configuration

1. Allows configuration of Legacy Banner Message Timeout
2. Allows configuration of Legacy Boot Protocol: None, PXE, iSCSI, PXE without fail-over, iSCSI without fail-over
3. Allows configuration of Wake on LAN
4. Allows configuration of Virtual LAN Mode and Virtual LAN ID
5. Allows configuration of Boot Retry Count.

A.3 Main Configuration Page - iSCSI Configuration

This section allows to override the default configurations of iSCSI and replaces DHCP configuration of iSCSI.
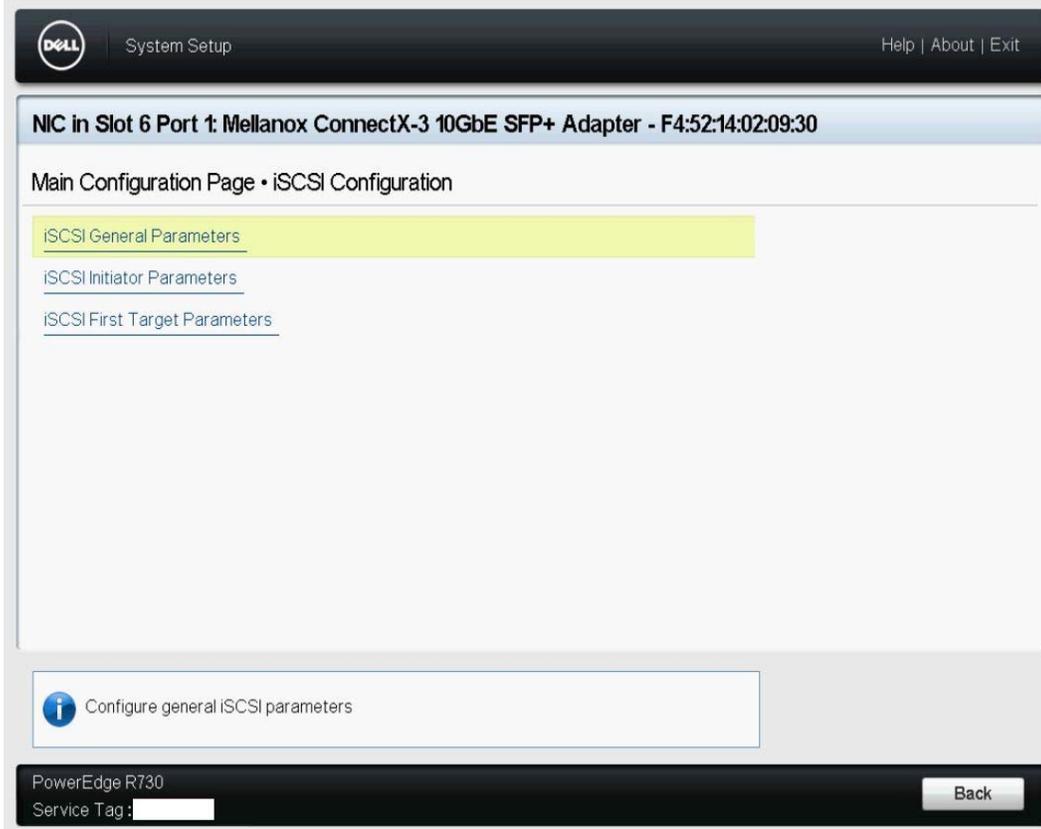
Figure 22: Main Configuration Page - iSCSI Configuration - iSCSI General Parameters

Figure 23: Main Configuration Page - iSCSI Configuration - iSCSI Initiator Parameters

Figure 24: Main Configuration Page - iSCSI Configuration - iSCSI Target Parameters

A.4 Main Configuration Page - Device Level Configuration

Allows setting Global Flow control settings for the adapter's port. Flow Control is deprecated. The only thing on the Device Level Configuration page is the SR-IOV configuration.

A.5   SR-IOV Configuration

Enabling SRIOV requires configuration both for the system and the specific Mellanox adapter.

To enable SR-IOV - follow steps 1-4.

To disable SR-IOV - set the configuration in steps 1-3 to disabled.

Step 1. Enable **"Virtualization Technology"** in System BIOS => Processor setting:
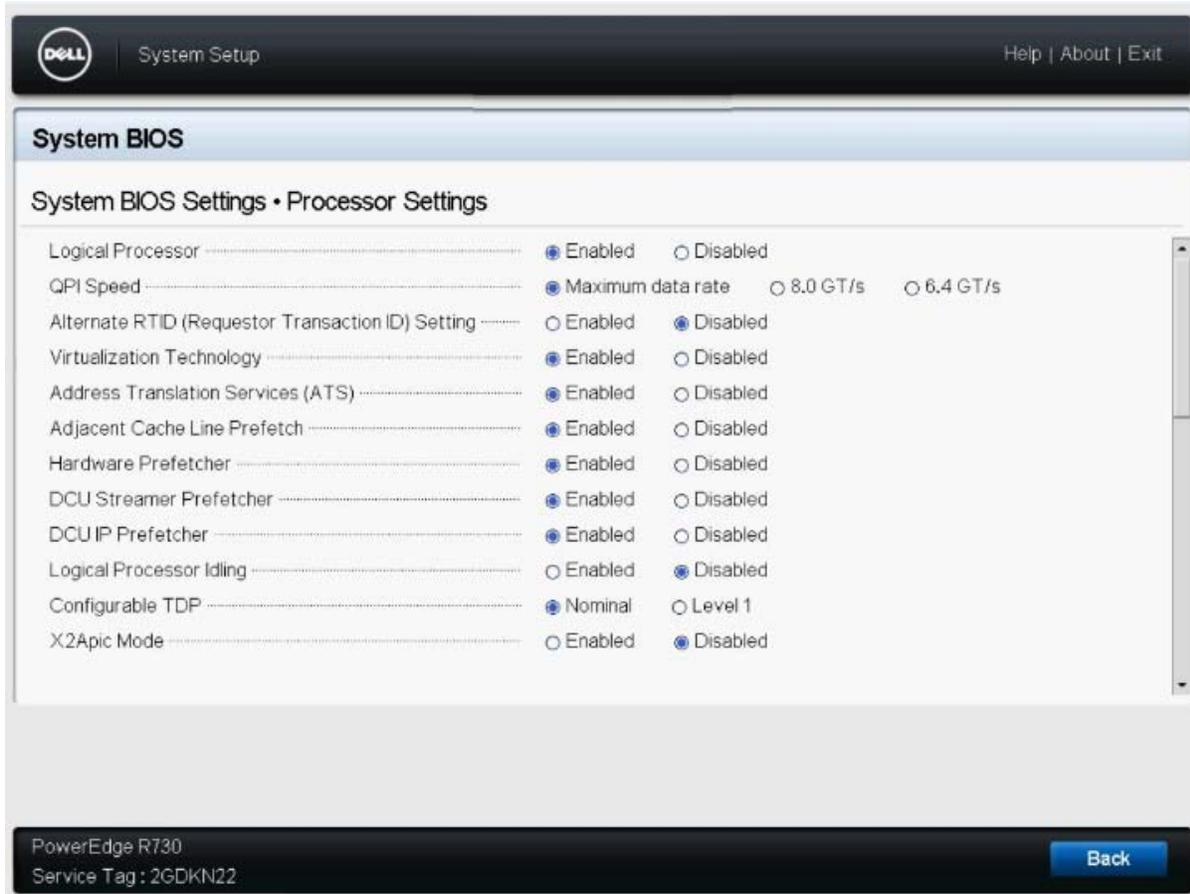
Step 2. Enable **"SR-IOV Global Enable"**

Go to: System BIOS => integrated Devices section
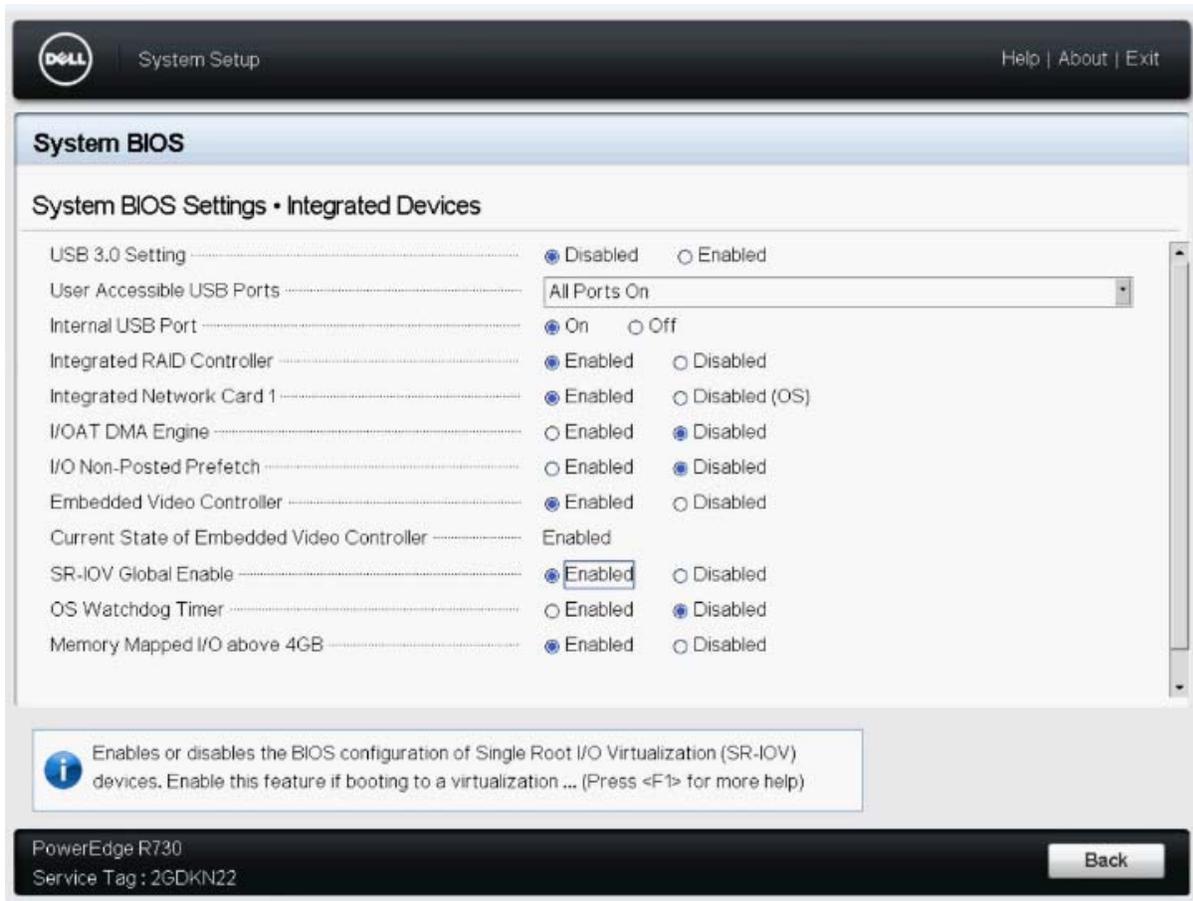
Step 3. Enable SR-IOV on the relevant adapter and set the number of required virtual functions:

Go to: Device settings => Select the relevant Mellanox adapter.

By default, Mellanox adapter is set to SR-IOV disabled.

Note the maximum number of virtual functions supported by the adapter PCIe.

Refer to the relevant driver user manual for support for SR-IOV and number of supported functions.
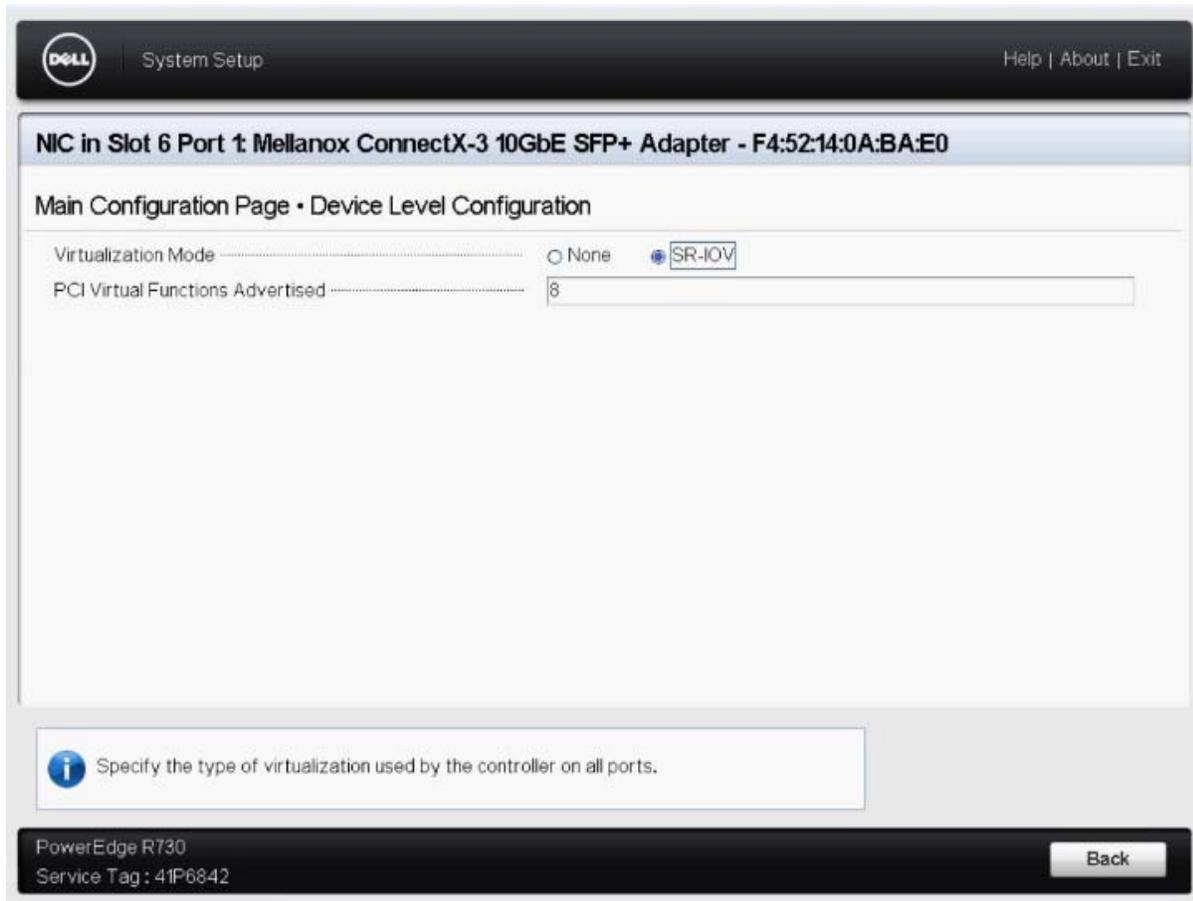
Step 4. Reboot the server for the SR-IOV configuration to take effect.

A.6   Wake on LAN Configuration

Applies to the following adapter cards:
- Mellanox ConnectX-3 10 GbE Mezzanine card
- Mellanox ConnectX-3 Pro 10 GbE Mezzanine card
- Mellanox ConnectX-4 Lx 25 GbE Rack NDC

Enabling Wake on LAN requires configuration for the specific Mellanox card.

Step 1.   On boot, press F2 to enter "System Setup"

Step 2.   Select "Device Settings"

Step 3.   Select the Wake on LAN capable Mellanox Adapter

Step 4.   Select "NIC Configuration"

Step 5.   Enable Wake on LAN setting

Step 6.   Exit and save.

Appendix B: Safety Warnings

1. Installation Instructions

Read all installation instructions before connecting the equipment to the power source.

2. Over-temperature

This equipment should not be operated in an area with an ambient temperature exceeding the maximum recommended: 55°C (131°F).
To guarantee proper air flow, allow at least 8cm (3 inches) of clearance around the ventilation openings.

3. During Lightning - Electrical Hazard

During periods of lightning activity, do not work on the equipment or connect or disconnect cables.

4. Copper Cable Connecting/Disconnecting

Some copper cables are heavy and not flexible, as such they should be carefully attached to or detached from the connectors. Refer to the cable manufacturer for special warnings and instructions.

5. Equipment Installation

This equipment should be installed, replaced, or serviced only by trained and qualified personnel.

6. Equipment Disposal

Disposal of this equipment should be in accordance to all national laws and regulations.

7. Local and National Electrical Codes

This equipment should be installed in compliance with local and national electrical codes.

8. Hazardous Radiation Exposure

Caution – Use of controls or adjustment or performance of procedures other than those specified herein may result in hazardous radiation exposure.

CLASS 1 LASER PRODUCT and reference to the most recent laser standards: IEC 60 825-1:1993 + A1:1997 + A2:2001 and EN 60825-1:1994+A1:1996+ A2:20.

Appendix C: Avertissements de sécurité d'installation (Warnings in French)

1.   **Instructions d'installation**

Lisez toutes les instructions d'installation avant de brancher le matériel à la source d'alimentation électrique.

2.   **Température excessive**

Ce matériel ne doit pas fonctionner dans une zone avec une température ambiante dépassant le maximum recommandé de 55°C (131°F). Un flux d'air de 200LFM à cette température ambiante maximale est nécessaire. En outre, pour garantir un bon écoulement de l'air, laissez au moins 8 cm (3 pouces) d'espace libre autour des ouvertures de ventilation.

3.   **Orages – dangers électriques**

Pendant un orage, il ne faut pas utiliser le matériel et il ne faut pas brancher ou débrancher les câbles.

4.   **Branchement/débranchement des câbles en cuivre**

Les câbles en cuivre sont lourds et ne sont pas flexibles, il faut donc faire très attention en les branchant et en les débranchant des connecteurs. Consultez le fabricant des câbles pour connaître les mises en garde et les instructions spéciales.

5.   **Installation du matériel**

Ce matériel ne doit être installé, remplacé ou entretenu que par du personnel formé et qualifié.

6.   **Elimination du matériel**

L'élimination de ce matériel doit s'effectuer dans le respect de toutes les législations et réglementations nationales en vigueur.

7.   **Codes électriques locaux et nationaux**

Ce matériel doit être installé dans le respect des codes électriques locaux et nationaux.

8.   Exposition au rayonnement grave

Mise en garde – l'utilisation de commandes ou de réglages ou l'exécution de procédures autres que ce qui est spécifié dans les présentes peut engendrer une exposition au rayonnement grave.

PRODUIT LASER DE CLASSE 1 » et références aux normes laser les plus récentes CEI 60 825-1

Appendix D: Sicherheitshinweise (Warnings in German)

1. Installationsanleitungen

Lesen Sie alle Installationsanleitungen, bevor Sie das Gerät an die Stromversorgung anschließen.

2. Übertemperatur

Dieses Gerät sollte nicht in einem Bereich mit einer Umgebungstemperatur über der maximal empfohlenen Temperatur von 55°C (131°F) betrieben werden. Es ist ein Luftstrom von 200 LFM bei maximaler Umgebungstemperatur erforderlich. Außerdem sollten mindestens 8 cm (3 in.) Freiraum um die Belüftungsöffnungen sein, um einen einwandfreien Luftstrom zu gewährleisten.

3. Bei Gewitter - Elektrische Gefahr

Arbeiten Sie während eines Gewitters und Blitzschlag nicht am Gerät, schließen Sie keine Kabel an oder ab.

4. Anschließen/Trennen von -Kupferkabel

Kupferkabel sind schwer und nicht flexible. Deshalb müssen sie vorsichtig an die Anschlüsse angebracht bzw. davon getrennt werden. Lesen Sie die speziellen Warnungen und Anleitungen des Kabelherstellers.

5.   Geräteinstallation

Diese Gerät sollte nur von geschultem und qualifiziertem Personal installiert, ausgetauscht oder gewartet werden.

6. Geräteentsorgung

Die Entsorgung dieses Geräts sollte unter Beachtung aller nationalen Gesetze Bestimmungen erfolgen.

7. Regionale und nationale elektrische Bestimmungen t

Dieses Gerät sollte unter Beachtung der regionalen und nationalen elektrischen Bestimmungen installiert werden.

8. Strahlenkontak

Achtung – Nutzung von Steuerungen oder Einstellungen oder Ausführung von Prozeduren, die hier nicht spezifiziert sind, kann zu gefährlichem Strahlenkontakt führen.

Klasse 1 Laserprodukt und Referenzen zu den aktuellsten Lasterstandards :
ICE 60 825-1

Appendix E: Advertencias de seguridad para la instalación (Warnings in Spanish)

1. Instrucciones de instalación

Antes de conectar el equipo a la fuente de alimentación, leer todas las instrucciones de instalación.

2. Sobrecalentamiento

No se debe utilizar el equipo en un área con una temperatura ambiente superior a la máxima recomendada: 55°C(131°F). Además, para garantizar una circulación de aire adecuada, se debe dejar como mínimo un espacio de 8 cm (3 pulgadas) alrededor de las aberturas de ventilación.

3. Cuando hay rayos: peligro de descarga eléctrica

No utilizar el equipo ni conectar o desconectar cables durante períodos de actividad de rayos.

4. Conexión y desconexión del cable Copper

Dado que los cables de cobre son pesados y no son flexibles, su conexión a los conectores y su desconexión se deben efectuar con mucho cuidado. Para ver advertencias o instrucciones especiales, consultar al fabricante del cable.

5. Instalación de equipos

La instalación, el reemplazo y el mantenimiento de este equipo estarán a cargo únicamente de personal capacitado y competente.

6. Eliminación de equipos

La eliminación definitiva de este equipo se debe efectuar conforme a todas las leyes y reglamentaciones nacionales.

7. Códigos eléctricos locales y nacionales

Este equipo se debe instalar conforme a los códigos eléctricos locales y nacionales.

8. Exposición a niveles de radiación peligrosos

Precaución: el uso de controles o ajustes o la realización de procedimientos distintos de los que aquí se especifican podrían causar exposición a niveles de radiación peligrosos.

PRODUCTO LÁSER DE CLASE 1 y referencia a las normas de láser más recientes: IEC 60825-1