Mellanox ConnectX-4 onwards NICs NATIVE ESXi 6.5 Driver for VMware vSphere Documentation

Rev. 4.16.70.1

Table of Contents

Overview

Mellanox native ESXi drivers enable industry-leading performance and efficiency as non-virtualized environments using hardware offloads such as RDMA over Converged Ethernet (RoCE) on VMware vSphere. Mellanox ConnectX-4 onwards deliver 10/25/40/50/100 and 200GbE network speeds with ESXi 6.5 onwards, allowing the highest port rate on ESXi today.

The documentation here relates to:

- Release Notes
- User Manual

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Software Download

Please visit http://www.mellanox.com → Products → Ethernet Drivers → VMware Driver

Document Revision History

A list of the changes made to the User Manual are provided in Document Revision History.

Common Abbreviations and Acronyms

Abbreviation/Acronym	Description
B	(Capital) 'B' is used to indicate size in bytes or multiples of bytes (e.g., 1KB = 1024 bytes, and 1MB = 1048576 bytes)
b	(Small) 'b' is used to indicate size in bits or multiples of bits (e.g., 1Kb = 1024 bits)
FW	Firmware
HCA	Host Channel Adapter
HW	Hardware
LSB	Least significant *byte*
lsb	Least significant *bit*
MSB	Most significant *byte*
msb	Most significant *bit*
NIC	Network Interface Card
SW	Software
VPI	Virtual Protocol Interconnect
PR	Path Record
RDS	Reliable Datagram Sockets
SDP	Sockets Direct Protocol
SL	Service Level
MPI	Message Passing Interface
QoS	Quality of Service
ULP	Upper Level Protocol

Abbreviation/Acronym	Description
vHBA	Virtual SCSI Host Bus adapter
uDAPL	User Direct Access Programming Library

Related Documentation

Document Name	Description
IEEE Std 802.3ae™-2002 (Amendment to IEEE Std 802.3-2002) Document # PDF: SS94996	Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment: Media Access Control (MAC) Parameters, Physical Layers, and Management Parameters for 10 Gb/s Operation
Firmware Release Notes for Mellanox adapter devices	See the Release Notes relevant to your adapter device. For further information please refer to the Mellanox website.
MFT Documentation	Mellanox Firmware Tools User Manual amd Release Notes. For further information please refer to the Mellanox website.
VMware vSphere Documentation Center	VMware website

1 Release Notes

Release Notes Update History

Revision	Date	Description
4.16.70.1	April 21, 2020	Initial release of this Release Notes version.

1.1 Overview

These are the release notes of Mellanox ConnectX-4/ConnectX-5 NATIVE ESXi Driver for VMware vSphere 6.5. Mellanox ConnectX-4/ConnectX-5 NATIVE ESXi Driver for VMware vSphere 6.5 supports the following uplinks to servers.

Version	OS	Uplink Speed
4.16.70.1	ESXi 6.5	10/25/40/50/100GbE

Content of MLNX-NATIVE-ESX Driver Package

ESXi 6.5:
MLNX-NATIVE-ESX-ConnectX-4-5_4.16.70.1-10EM-650.0.0.4598673.zip - Hypervisor bundle for ESXi 6.5 contains the following kernel modules:

- nmlx5_core
- nmlx5_rdma

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

1.1.1 Supported HCAs Firmware Versions

MLNX-NATIVE-ESX Rev 4.16.70.1 supports the following Mellanox Ethernet HCA and their corresponding firmware version:

HCAs	Minimal Recommended Firmware Rev.
ConnectX-4	12.27.1016
ConnectX-4 Lx	14.27.1016
ConnectX-5 / ConnectX-5 Ex	16.27.2008
ConnectX-6	20.27.2008
ConnectX-6 Dx	22.27.2008

For the latest firmware versions, visit: https://www.mellanox.com/support/firmware/firmware-downloads

1.1.2 Tested Hypervisors in Paravirtualized and SR-IOV Environments

Tested Hypervisors	HCAs	Guest Operating System
SR-IOV	ConnectX-4	Windows Server 2016 DC
	ConnectX-4 Lx	RedHat 8.0
	ConnectX-5/ConnectX-5 Ex	RedHat 7.5
	ConnectX-6	RedHat 7.3
	ConnectX-6 Dx	RedHat 6.10

Tested Hypervisors	HCAs	Guest Operating System
		RedHat 6.3 SLES 12 SP4 SLES 12 SP3
Paravirtualized[a] (Ethernet Only)	ConnectX-4 ConnectX-4 Lx ConnectX-5/ConnectX-5 Ex ConnectX-6 ConnectX-6 Dx	Windows Server 2016 DC RedHat 8.0 RedHat 7.5 RedHat 7.3 RedHat 6.10 RedHat 6.3 SLES 12 SP4 SLES 12 SP3

a. Paravirtualized RDMA is supported only in Linux Operating Systems and in this release it was tested for RedHat 7.5 only.

1.2 Changes and New Features

Feature/Change	Description
4.16.70.1	
Power Limitation	An event will be sent to notify the administrator if the power required by the network adapter is higher than that available on the PCIe slot.
Differentiated Services Code Point (DSCP)	Added support for trusting Differentiated Services Code Point (DSCP) and setting default value for RoCE traffic.
SR-IOV VF Counters	Added a new counter that enables the user to query per Virtual Function counters.
RX Counters	Added the RX out-of-buffer counter to indicate any lack of software receive buffers.
RoCE, RDMA	Added a module parameter to enforce specific RoCE version.
SR-IOV	SR-IOV InfiniBand is at beta level.

1.3 Bug Fixes in the Version

The table below lists the bugs fixed in this release. For older issues, please refer to Bug Fixes History.

Internal Ref.	Description
781277	**Description:** The "esxcli network sriovnic vf stats" command is not supported. When running this command on a vmknic, a failure message is displayed.
	Keywords: esxcli SR-IOV
	Discovered in Version: 4.6.10.3
	Fixed in Release: 4.16.70.1

1.4 Known Issues

The following is a list of general limitations and known issues of the various components of this MLNX-NATIVE-ESX release.

Internal Ref.	Description
2120216	**Description:** The maximum number of established active RDMA connections (QPs) is currently 5000.
	Workaround: N/A
	Keywords: QPs, RDMA
	Discovered in Version: 4.16.70.1
2130911	**Description:** Setting ETS value to 0 may cause WQE timeout.
	Workaround: Set ETS value of 1 instead of 0.
	Keywords: ETS, QOS
	Discovered in Version: 4.16.70.1
1446060	**Description:** Although the max_vfs module parameter range is "0-128", due to firmware limitations, the following are the supported VFs per single port devices: • ConnectX-4 / ConnectX-5: up to 127
	Workaround: N/A
	Keywords: SR-IOV, VFs per port
	Discovered in Version: 4.16.14.2
1340275	**Description:** ECN tunable parameter initialAlphaValue for the Reaction Point protocol cannot be modified.
	Workaround: N/A
	Keywords: nmlx5 ecn nmlxcli
	Discovered in Version: 4.16.13.5
1340255	**Description:** ECN statistic counters accumulatorsPeriod and ecnMarkedRocePackets display wrong values and cannot be cleared.
	Workaround: N/A
	Keywords: nmlx5 ecn nmlxcli
	Discovered in Version: 4.16.13.5
-	**Description:** The hardware can offload only up to 256B of headers.
	Workaround: N/A
	Keywords: Hardware offload
	Discovered in Version: 4.16.10.3
685558	**Description:** There is no traffic between PV and SR-IOV VF connected to different ports on the same HCA. This issue is applicable to ESXi 6.5 & ESXi 6.5 UP1. The issue is solved in ESXi 6.5 UP2.

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Internal Ref.	Description
	Workaround: N/A
	Keywords: PV, SR-IOV VF, HCA
858972	**Description:** Setting the "Allow Guest MTU Change" option in vSphere Client is currently not functional. Although guest MTU changes in SR-IOV are allowed, they do not affect the port's MTU and the guest's MTU remains the same as the PF MTU.
	Workaround: N/A
	Keywords: MTU, SR-IOV
-	**Description:** Geneve options length support is limited to 56B. Received packets with options length bigger than 56B are dropped.
	WA: N/A
	Keywords: Geneve
910292	**Description:** Running with ConnectX-4/ConnectX-4 Lx older firmware versions, might result in the following internal firmware errors: • Device health compromised • synd 0x1: firmware internal error • extSync 0x94ee
	Workaround: Upgrade your firmware to the latest version 12.17.2020/14.17.2020
	Keywords: Firmware
746100	Description: The 'esxcli mellanox uplink link info -u <vmnic_name>' command reports the 'Auto negotiation' capability always as 'true'.
	Workaround: N/A
	Keywords: 'Auto negotiation' capability
1072640	Description: ESXi v4.16.10.3 cannot updated from v4.16.8.8 (GA) or from the Inbox driver using the "esxcli software vib update" command.
	Workaround: To update it, run the "esxcli software vib install" command.
	Keywords: Driver update
1068621	Description: SMP MADs (ibnetdiscover, sminfo, iblinkinfo, smpdump, ibqueryerr, ibdiagnet and smpquery) are not supported on the VFs.
	Workaround: N/A
	Keywords: SMP MADs
778371	**Description:** Wake-on-LAN does not notify when invalid parameters are provided.
	Workaround: N/A
	Keywords: WoL
778572	**Description:** Nested ESXi might not function properly.
	Workaround: N/A

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Internal Ref.	Description
	Keywords: Nested ESXi
765008	**Description:** Device RSS fails to hash traffic to sufficient RX rings with Broadcast traffic.
	Workaround: N/A
	Keywords: RSS, RX rings
852883	**Description:** In stress condition 'Watchdog' may appear, leading to uplink going up and down.
	Workaround: N/A
	Keywords: uplink, watchdog

1.5 Bug Fixes History

The table below lists the bugs fixed in this release.

Internal Ref.	Description
1358381	**Description:** Fixed an issue that prevented ESXi from being discovered via the CDP protocol on ConnectX-4 Lx adapter cards.
	Keywords: CDP protocol, ConnectX-4 Lx
	Discovered in Release: 4.16.12.12
	Fixed in Release: 4.16.14.2
1253564	**Description:** Disabled multicast loopback to avoid a scenario that prevented MAC learning in some configurations.
	Keywords: MAC, multicast loopback
	Discovered in Release: 4.16.12.12
	Fixed in Release: 4.16.13.5
958154	**Description:** NetQ RSS for encapsulated traffic is currently not supported. Encapsulated traffic (VXLAN/Geneve) directed to NetQ RSS queue will not be distributed through all queues' channels, thus will not utilize the RSS feature. **Note:** It is highly recommended to avoid requesting RSS for encapsulated interfaces, i.e. refrain from defining the following in the VM configuration file: <iface_name>.pnicFeatures=4
	Keywords: NetQ RSS, encapsulated traffic
	Discovered in Release: 4.16.8.8
	Fixed in Release: 4.16.12.12
698142/637104	**Description:** Traffic loss of large packets might occur after MTU change.
	Keywords: MTU, Traffic loss

Internal Ref.	Description
	Discovered in Release: 4.16.7.8
	Fixed in Release: 4.16.10.3
846359	**Description:** Fixed an issue which caused the adapter card to get stuck in Down state after setting the ring size to 8192.
	Keywords: Ring size
	Discovered in Release: 4.16.7.8
	Fixed in Release: 4.16.8.8

2 Introduction

Mellanox ConnectX®-4 onwards NATIVE ESXi is a software stack which operates across all Mellanox network adapter solutions supporting up to 200Gb/s Ethernet (ETH) and 2.5 or 5.0 GT/s PCI Express 2.0 and 3.0 uplinks to servers.

The following sub-sections briefly describe the various components of the Mellanox ConnectX-4 onwards NATIVE ESXi stack.

2.1 nmlx5 Driver

nmlx5 is the low-level driver implementation for the ConnectX-4/ConnectX-5 adapter cards designed by Mellanox Technologies. ConnectX-4/ConnectX-5 adapter cards can operate as an InfiniBand adapter, or as an Ethernet NIC. The ConnectX-4/ConnectX-5 NATIVE ESXi driver supports Ethernet NIC configurations exclusively. In addition, the driver provides RDMA over Converged Ethernet (RoCE) functionality through ESXi RDMA layer APIs (kernel-space only) and SR-IOV.

2.2 Mellanox NATIVE ESXi Package

2.2.1 Software Components

MLNX-NATIVE-ESX-ConnectX-4/ConnectX-5 contains the following software components:

- Mellanox Host Channel Adapter Drivers
    - **nmlx5_core** (Ethernet): Handles Ethernet specific functions and plugs into the ESXi uplink layer
- **nmlx5_rdma**: Enables RoCE functionality by plugging into the ESXi RDMA layer

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 2 + Aligned at: 0.5" + Tab after: 0.75" + Indent at: 0.75"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

2.3 Module Parameters

2.3.1 Module Parameters

To set **nmlx5_core** parameters:

```
esxcli system module parameters set -m nmlx5_core -p <parameter>=<value>
```

To set nmlx5_rdma parameters:

```
esxcli system module parameters set -m nmlx5_rdma -p <parameter>=<value>
```

To show the values of the parameters:

```
esxcli system module parameters list -m  <module name>
```

For the changes to take effect, reboot the host.

2.3.1.1 nmlx5_core Module Parameters

Name	Description	Values
DRSS	Number of hardware queues for Default Queue (DEFQ) RSS. **Note:** This parameter replaces the previously used “drss” parameter which is now obsolete.	• 2-16 • 0 - disabled When this value is != 0, DEFQ RSS is enabled with 1 RSS Uplink queue that manages the 'drss' hardware queues. **Notes:** • The value must be a power of 2. • The value must not exceed num. of CPU cores. • Setting the DRSS value to 16, sets the Steering Mode to device RSS
ecn	Enables the ECN feature	• 1 - enabled • 0 - disabled (Default)
enable_nmlx_debug	Enables debug prints for the core module.	• 1 - enabled • 0 - disabled (Default)
geneve_offload_enable	Enables GENEVE HW Offload	• 1 - enabled • 0 - disabled (Default)
max_vfs	max_vfs is an array of comma separated integer values, that represent the amount of VFs to open from each port. For example: max_vfs = 1,1,2,2, will open a single VF per port on the first NIC and 2 VFs per port on second NIC. The order of the NICs is determined by pci SBDF number. **Note:** VFs creation based on the system resources limitations.	• 0 - disabled (Default) N number of VF to allocate over each port **Note:** The amount of values provided in the max_vfs array should not exceed the supported_num_ports module parameter value.
mst_recovery	Enables recovery mode (only NMST module is loaded).	• 1 - enabled • 0 - disabled (Default)
pfcrx	Priority based Flow Control policy on RX.	• 0-255 • 0 - default It is an 8 bits bit mask, where each bit indicates a priority [0-7]. Bit values:

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Name	Description	Values
		• 1 - respect incoming PFC pause frames for the specified priority. • 0 - ignore incoming pause frames on the specified priority. **Note**: The pfcrx and pfctx values must be identical.
pfctx	Priority based Flow Control policy on TX.	• 0-255 • 0 - default It is an 8 bits bit mask, where each bit indicates a priority [0-7]. Bit values: • 1 - generate pause frames according to the RX buffer threshold on the specified priority. • 0 - never generate pause frames on the specified priority. **Note**: The pfcrx and pfctx values must be identical.
RSS	Number of hardware queues for NetQ RSS. **Note:** This parameter replaces the previously used “rss” parameter which is now obsolete.	• 2-8 • 0 - disabled When this value is != 0, NetQ RSS is enabled with 1 RSS uplink queue that manages the 'rss' hardware queues. Notes: • The value must be a power of 2 • The maximum value must be lower than the number of CPU cores.
supported_num_ports	Sets the maximum supported ports.	2-8 Default 4 **Note:** Before installing new cards, you must modify the maximum number of the supported ports to include the additional new ports.

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

3 Installation

This chapter describes how to install and test the Mellanox ConnectX-4/ConnectX-5 NATIVE ESXi package on a single host machine with Mellanox Ethernet adapter hardware installed.

3.1 Hardware and Software Requirements

Requirements	Description
Platforms	A server platform with an adapter card based on one of the following Mellanox Technologies' HCA devices: • ConnectX®-4 (EN) (firmware: fw-ConnectX4) • ConnectX®-4 Lx (EN) (firmware: fw-ConnectX4-Lx) • ConnectX®-5 (VPI) (firmware: fw-ConnectX5) • ConnectX®-5 Ex (VPI) (firmware: fw-ConnectX5)
Device ID	For the latest list of device IDs, please visit Mellanox website.
Operating System	ESXi 6.5: 4.16.10.3
Installer Privileges	The installation requires administrator privileges on the target machine.

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

3.2 Installing Mellanox NATIVE ESXi Driver for VMware vSphere

Uninstall any previous Mellanox driver packages prior to installing the new version. See "Removing the Previous Mellanox Driver" for further information.

To install the driver, do the following:

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 /tmp/MLNX-NATIVE-ESX-ConnectX-4-
5_4.16.10.3-10EM- 650.0.0.2768847.zip
```

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

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

```
# esxcli software vib list | grep nmlx
nmlx5-core  4.16.10.3-1OEM.650.0.0.4598673  MEL PartnerSupported 2017-01-
31
nmlx5-rdma  4.16.10.3-1OEM.650.0.0.4598673  MEL PartnerSupported 2017-01-
31
```

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

3.3 Removing the Previous Mellanox Driver

Unload the driver before removing it.

To remove all the drivers, do the following:

1. Log into the ESXi server with root permissions.
2. List all the existing NATIVE ESXi driver modules. (see Step 4 in "Installing Mellanox NATIVE ESXi Driver for VMware vSphere" section above)
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.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

3.4 Downgrading to an Older Mellanox Driver Version

Unload the driver before removing it.

To downgrade to the previous ESXi version, do the following:

Automatic downgrade flow in currently unavailable for current driver version due to a change in the number of driver modules. Using "esxcli

software update" command to downgrade may cause unexpected result. In order to safely downgrade to any previous version (e.g. 4.16.8.8), you must **manually** remove the current version and install the previous one as described in the process below.

1. Log into the ESXi server with root permissions.
2. List all the existing NATIVE ESXi driver modules. (see Step 4 in "Installing Mellanox NATIVE ESXi Driver for VMware vSphere" section above)
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. Install the desired driver version.
5. Reboot the machine.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

3.5 Firmware Programming

1. Download the VMware bootable binary images v4.11.0 from the Mellanox Firmware Tools (MFT) site.

   ESXi 6.5 File: mft-4.8.0.26-10EM-650.0.0.4598673.x86_64.vib

   MD5SUM: 6f4a1c1ef2482f091bee4086cbec5caf
2. Install the image according to the steps described in the MFT User Manual.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4 Features Overview and Configuration

The chapter contains the following sections:

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4.1 Ethernet Network

4.1.1 Port Type Management

ConnectX®-4 onward adapter cards' ports can be individually configured to work as InfiniBand or Ethernet ports. The port type depends on the card type. In case of a VPI card, the default type is IB. If you wish to change the port type use the mlxconfig script.

To use a VPI card as an Ethernet only card, run:

```
/opt/mellanox/bin/mlxconfig -d /dev/mt4115_pciconf0 set LINK_TYPE_P1=2
LINK_TYPE_P2=2
```

The protocol types are:

- Port Type 1 = IB
- Port Type 2 = Ethernet

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

For further information on how to set the port type in ConnectX®-4/ConnectX®-4 Lx/ ConnectX®-5, please refer to the MFT User Manual (www.mellanox.com → Products → Software → InfiniBand/VPI Software → MFT - Firmware Tools).

4.1.2 Wake-on-LAN (WoL)

Wake-on-LAN (WoL) is applicable only to adapter cards that support this feature.

Wake-on-LAN (WoL) is a technology that allows a network professional to remotely power on a computer or to wake it up from sleep mode.

- To enable WoL:

```
esxcli network nic set -n <nic name> -w g
```

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

or

```
set /net/pNics/<nic name>/wol g
```

- To disable WoL:

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
vsish -e set /net/pNics/<nic name>/wol d
```

- To verify configuration:

```
esxcli network nic get -n vmnic5
   Advertised Auto Negotiation: true
   Advertised Link Modes: 10000baseT/Full, 40000baseT/Full,
100000baseT/Full, 100baseT/Full, 1000baseT/Full, 25000baseT/Full,
50000baseT/Full
   Auto Negotiation: false
   Cable Type: DA
   Current Message Level: -1
   Driver Info:
         Bus Info: 0000:82:00:1
         Driver: nmlx5_core
         Firmware Version: 12.20.1010
         Version: 4.15.10.3
   Link Detected: true
   Link Status: Up
   Name: vmnic5
   PHYAddress: 0
   Pause Autonegotiate: false
   Pause RX: false
   Pause TX: false
   Supported Ports:
   Supports Auto Negotiation: true
   Supports Pause: false
   Supports Wakeon: false
   Transceiver:
   Wakeon: MagicPacket(tm)
```

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4.1.3 Set Link Speed

The driver is set to auto-negotiate by default. However, the link speed can be forced to a specific link speed supported by ESXi using the following command:

```
esxcli network nic set -n <vmnic> -S <speed> -D <full, half>
```

Example:

```
esxcli network nic set -n vmnic4 -S 10000 -D full
```

Where:

- <speed> can be 10/100/1000/2500/5000/10000/20000/25000/40000/50000/56000/100000/200000Mb/s.
- <vmnic> is the vmnic for the Mellanox card as provided by ESXi
- <full, half> The duplex to set this NIC to. Acceptable values are: [full, half]

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

The driver can be reset to auto-negotiate using the following command:

```
esxcli network nic set -n <vmnic> -a
```

Example:

```
esxcli network nic set -n vmnic4 -a
```

where <vmnic> is the vmnic for the Mellanox card as provided by ESXi.

4.1.4 Priority Flow Control (PFC)

Priority Flow Control (PFC) IEEE 802.1Qbb applies pause functionality to specific classes of traffic on the Ethernet link. PFC can provide different levels of service to specific classes of Ethernet traffic (using IEEE 802.1p traffic classes).

When PFC is enabled, Global Pause will be operationally disabled, regardless of what is configured for the Global Pause Flow Control.

To configure PFC, do the following:

1. Enable PFC for specific priorities.

```
esxcfg-module nmlx5_core -s "pfctx=0x08 pfcrx=0x08"
```

The parameters, “pfctx” (PFC TX) and “pfcrx” (PFC RX), are specified per host. If you have more than a single card on the server, all ports will be enabled with PFC (Global Pause will be disabled even if configured).

The value is a bitmap of 8 bits = 8 priorities. We recommend that you enable only lossless applications on a specific priority.

To run more than one flow type on the server, turn on only one priority (e.g. priority 3), which should be configured with the parameters "0x08" = 00001000b (binary). Only the 4th bit is on (starts with priority 0,1,2 and 3 -> 4th bit).

The values of “pfctx” and “pfcrx” must be identical.

2. Restart the host for changes to the module parameters to take effect.

```
reboot
```

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4.1.5 Receive Side Scaling (RSS)

Receive Side Scaling (RSS) technology allows spreading incoming traffic between different receive descriptor queues. Assigning each queue to different CPU cores allows better load balancing of the incoming traffic and improve performance.

4.1.5.1 Default Queue Receive Side Scaling (DRSS)

Default Queue RSS (DRSS) allows the user to configure multiple hardware queues backing up the default RX queue. DRSS improves performance for large scale multicast traffic between hypervisors and Virtual Machines interfaces.

To configure DRSS, use the 'DRSS' module parameter which replaces the previously advertised 'device_rss' module parameter ('device_rss' is now obsolete). The 'drss' module parameter and 'device_rss' are mutually exclusive

If the 'device_rss' module parameter is enabled, the following functionality will be configured:

- The new Default Queue RSS mode will be triggered and all hardware RX rings will be utilized, similar to the previous 'device_rss' functionality
- Module parameters 'DRSS' and 'RSS' will be ignored, thus the NetQ RSS, or the standard NetQ will be active

To query the 'DRSS' module parameter default, its minimal or maximal values, and restrictions, run a standard esxcli command.

For example:

```
#esxcli system module parameters list -m nmlx5_core
```

4.1.5.2 NetQ RSS

NetQ RSS is a new module parameter for ConnectX-4 adapter cards providing identical functionality as the ConnectX-3 module parameter 'num_rings_per_rss_queue'. The new module parameter allows the user to configure multiple hardware queues backing up the single RX queue. NetQ RSS improves vMotion performance and multiple streams of IPv4/IPv6 TCP/ UDP/IPSEC bandwidth over single interface between the Virtual Machines.

To configure NetQ RSS, use the 'RSS' module parameter. To query the 'RSS' module parameter default, its minimal or maximal values, and restrictions, run a standard esxcli command.

For example:

```
#esxcli system module parameters list -m nmlx5_core
```

Using NetQ RSS is preferred over the Default Queue RSS. Therefore, if both module parameters are set but the system lacks resources to support both, NetQ RSS will be used instead of DRSS.

4.1.5.2.1 Important Notes

If the 'DRSS' and 'RSS' module parameters set by the user cannot be enforced by the system due to lack of resources, the following actions are taken in a sequential order:

1. The system will attempt to provide the module parameters default values instead of the ones set by the user
2. The system will attempt to provide 'RSS' (NetQ RSS mode) default value. The Default Queue RSS will be disabled
3. The system will load with only standard NetQ queues
4. 'DRSS' and 'RSS' parameters are disabled by default, and the system loads with standard NetQ mode

4.1.6 RDMA over Converged Ethernet (RoCE)

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 to provide this efficient

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

data transfer with very low latencies on lossless Ethernet networks. With advances in data center convergence over reliable Ethernet, ConnectX® EN with RoCE uses the proven and efficient RDMA transport to provide the platform for deploying RDMA technology in mainstream data center application at 10GigE and 40GigE link-speed. ConnectX® EN with its hardware offload support takes advantage of this efficient RDMA transport (InfiniBand) services over Ethernet to deliver ultra-low latency for performance-critical and transaction intensive applications such as financial, database, storage, and content delivery networks.

When working with RDMA applications over Ethernet link layer the following points should be noted:

- The presence of a Subnet Manager (SM) is not required in the fabric. Thus, operations that require communication with the SM are managed in a different way in RoCE. This does not affect the API but only the actions such as joining multicast group, that need to be taken when using the API
- Since LID is a layer 2 attribute of the InfiniBand protocol stack, it is not set for a port and is displayed as zero when querying the port
- With RoCE, the alternate path is not set for RC QP and therefore APM is not supported
- GID format can be of 2 types, IPv4 and IPv6. IPv4 GID is a IPv4-mapped IPv6 address while IPv6 GID is the IPv6 address itself
- VLAN tagged Ethernet frames carry a 3-bit priority field. The value of this field is derived from the InfiniBand SL field by taking the 3 least significant bits of the SL field
- RoCE traffic is not shown in the associated Ethernet device's counters since it is offloaded by the hardware and does not go through Ethernet network driver. RoCE traffic is counted in the same place where InfiniBand traffic is counted:

```
esxcli rdma device stats get -d [RDMA device]
```

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

It is recommended to use RoCE with PFC enabled in driver and network switches.

For how to enable PFC in the driver see "Priority Flow Control (PFC)" section.

4.1.6.1 RoCE Modes

RoCE encapsulates InfiniBand transport in one of the following Ethernet packet

- RoCEv1 - dedicated ether type (0x8915)
- RoCEv2 - UDP and dedicated UDP port (4791)

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4.1.6.1.1 RoCEv1

RoCE v1 protocol is defined as RDMA over Ethernet header (as shown in the figure above). It uses ethertype 0x8915 and may can be used with or without the VLAN tag. The regular Ethernet MTU applies on the RoCE frame.

4.1.6.1.2 RoCEv2

A straightforward extension of the RoCE protocol enables traffic to operate in IP layer 3 environments. This capability is obtained via a simple modification of the RoCE packet format. Instead of the GRH used in RoCE, IP routable RoCE packets carry an IP header which allows traversal of IP L3 Routers and a UDP header (RoCEv2 only) that serves as a stateless encapsulation layer for the RDMA Transport Protocol Packets over IP.

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.

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, in a completely transparent way.

4.1.6.2 GID Table Population

The GID table is automatically populated by the ESXi RDMA stack using the 'binds' mechanism, and has a maximum size of 128 entries per port. Each bind can be of type RoCE v1 or RoCE v2, where entries of both types can coexist on the same table. Binds are created using IP-based GID generation scheme.

For more information, please refer to the "VMkernel APIs Reference Manual."

4.1.6.3 Prerequisites

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

- ConnectX®-4 firmware version 12.17.2020 and above
- ConnectX®-4 Lx firmware version 14.17.2020 and above
- ConnectX®-5 firmware version 16.20.1000 and above
- All InfiniBand verbs applications which run over InfiniBand verbs should work on RoCE links if they use GRH headers.
- All ports must be set to use Ethernet protocol

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4.1.6.4 Running and Configuring RoCE on ESXi VMs

RoCE on ESXi VMs can run on VMs which are associated with either SR-IOV EN Virtual Functions or passthrough.

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.

On ESXi, the PFC settings should be set on the ESXi host only and not on the VMs as the ESXi host is the one to control PFC settings. PFC settings can be changed using the mlx5_core parameters pfctx and pfcrx. For further information, please refer to "nmlx5_core Parameters".

For further information on how to use and run RoCE on the VM, please refer t the VM's driver User Manual. Additional information can be found at the *RoCE Over L2 Network Enabled with PFC* User Guide:

http://www.mellanox.com/related-docs/prod_software/RoCE_with_Priority_Flow_Control_Application_Guide.pdf

4.1.7 Overlay Networking Stateless Hardware Offload

VXLAN/Geneve hardware offload enables the traditional offloads to be performed on the encapsulated traffic. With ConnectX® family adapter cards, data center operators can decouple the overlay network layer from the physical NIC performance, thus achieving native performance in the new network architecture.

4.1.7.1 Configuring Overlay Networking Stateless Hardware Offload

VXLAN/Geneve hardware offload includes:

- TX: Calculates the Inner L3/L4 and the Outer L3 checksum
- RX:

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

- Checks the Inner L3/L4 and the Outer L3 checksum
- Maps the VXLAN traffic to an RX queue according to:
  - Inner destination MAC address
  - Outer destination MAC address
  - VXLAN ID

VXLAN/Geneve hardware offload is enabled by default and its status cannot be changed.

VXLAN/Geneve configuration is done in the ESXi environment via VMware NSX manager. For additional NSX information, please refer to VMware documentation: http://pubs.vmware.com/NSX-62/index.jsp#com.vmware.nsx.install.doc/GUID-D8578F6E-A40C-493A-9B43-877C2B75ED52.html.

4.1.8 Packet Capture Utility

Packet Capture utility duplicates all traffic, including RoCE, in its raw Ethernet form (before stripping) to a dedicated "sniffing" QP, and then passes it to an ESX drop capture point.
It allows gathering of Ethernet and RoCE bidirectional traffic via pktcap-uw and viewing it using regular Ethernet tools, e.g. Wireshark.

> By nature, RoCE traffic is much faster than ETH. Meaning there is a significant gap between RDMA traffic rate and Capture rate.
> Therefore actual "sniffing" RoCE traffic with ETH capture utility is not feasible for long periods.

4.1.8.1 Components

Packet Capture Utility is comprised of two components:

- ConnectX-4 RDMA module sniffer:
  This component is part of the Native ConnectX-4 RDMA driver for ESX and resides in Kernel space.
- RDMA management interface:
  User space utility which manages the ConnectX-4 Packet Capture Utility

4.1.8.2 Usage

1. Installed the latest ConnectX-4 driver bundle.
2. Make sure all Native nmlx5 drivers are loaded

```
esxcli system module list | grep nmlx
nmlx5_core                     true        true
nmlx5_rdma                     true        true
```

3. Install the nmlxcli management tool (esxcli extension) using the supplied bundle MLNX-NATIVE-NMLXCLI_<version>.zip

```
esxcli software vib install -d <path to bundle>/MLNX-NATIVE-
NMLXCLI_1.16.12.11-10EM-650.0.0.4598673.zip
```

When the nmlxcli management tool is installed, the following esxli commands namespace is available:

Formatted: Bulleted + Level: 2 + Aligned at: 0.5" + Tab after: 0.75" + Indent at: 0.75", Tab stops: Not at 1"

Formatted: Bulleted + Level: 3 + Aligned at: 1" + Tab after: 1.25" + Indent at: 1.25", Tab stops: Not at 1.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
# esxcli mellanox uplink sniffer
```

This namespace allows user basic packet capture utility operations such as: query, enable or disable.
Usage of the tool is shown by running one of the options below:

```
snifferesxcli mellanox uplink sniffer {cmd} [cmd options]
```

Options:

```
disable      Disable sniffer on specified uplink
             * Requires -u/--uplink-name parameter
enable       Enable sniffer on specified uplink
             * Requires -u/--uplink-name parameter
query        Query operational state of sniffer on specified uplink
             * Requires -u/--uplink-name parameter
```

4. Determine the uplink device name.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
Name    PCI Device    Driver      Admin Status  Link Status   Speed
Duplex  MAC Address         MTU  Description
------  ------------  ----------  ------------  -----------  ------  -----
-  -----------------  ----  ---------------------------------------------
---------
vmnic4  0000:07:00.0  nmlx5_core  Up            Up           100000  Full
7c:fe:90:63:f2:d6  1500  Mellanox Technologies MT27700 Family [ConnectX-4]
vmnic5  0000:07:00.1  nmlx5_core  Up            Up           100000  Full
7c:fe:90:63:f2:d7  1500  Mellanox Technologies MT27700 Family [ConnectX-4]
```

5. Enable the packet capture utility for the required device(s).

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
esxcli mellanox uplink sniffer enable -u <vmnic_name>
```

6. Use the ESX internal packet capture utility to capture the packets.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
pktcap-uw --capture Drop --o <capture_file>
```

7. Generate the RDMA traffic through the RDMA device.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

8. Stop the capture.
9. Disable the packet capture utility.

```
esxcli mellanox uplink sniffer disable -u <vmnic_name>
```

10. Query the packet capture utility.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
esxcli mellanox uplink sniffer query -u <vmnic_name>
```

4.1.8.3 Limitations

- **Capture duration**: Packet Capture Utility is a debug tool, meant to be used for bind failure diagnostics and short period packet sniffing. Running it for a long period of time with stress RDMA traffic will cause undefined behavior. Gaps in capture packets may appear.

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

- **Overhead**: A significant performance decrease is expected when the tool is enabled:

- The tool creates a dedicated QP and HW duplicates all RDMA traffic to this QP, before stripping the ETH headers.
- The captured packets reported to ESX are duplicated by the network stack adding to the overhaul execution time

- **Drop capture point:** The tool uses the VMK_PKTCAP_POINT_DROP to pass the captured traffic. Meaning whomever is viewing the captured file will see all RDMA capture in addition to all the dropped packets reported to the network stack.
- **ESX packet exhaustion:** During the enable phase (/opt/mellanox/bin/ nmlx4_sniffer_mgmt-user -a vmrdma3 -e) the Kernel component allocates sniffer resources, and among these are the OS packets which are freed upon tool's disable. Multiple consecutive enable/disable calls may cause temporary failures when the tool requests to allocate these packets. It is recommended to allow sufficient time between consecutive disable and enable to fix this issue.

4.2 Single Root IO Virtualization (SR-IOV)

SR-IOV InfiniBand is at beta level.

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. Mellanox adapters are capable of exposing in ConnectX-4 onwards adapter cards up to 63/127 virtual instances called Virtual Functions (VFs) depending on the firmware capabilities. These virtual functions can then be provisioned separately. Each VF can be seen as an addition device connected to the Physical Function. It shares the same resources with 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.

In this chapter we will demonstrate setup and configuration of SR-IOV in a ESXi environment using Mellanox ConnectX® adapter cards family.

4.2.1 System Requirements

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

- nmlx5_core Driver
- A server/blade with an SR-IOV-capable motherboard BIOS
- Mellanox ConnectX® Adapter Card family with SR-IOV capability
- Hypervisor that supports SR-IOV such as: ESXi6.5

4.2.2 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:

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

Formatted: Bulleted + Level: 2 + Aligned at: 0.5" + Tab after: 0.75" + Indent at: 0.75", Tab stops: Not at 1"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

2. Enable "Intel Virtualization Technology."

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

4.2.2.1 Configuring SR-IOV for ConnectX-4 onward

To configure SR-IOV for ConnectX-4 onward, perform the following steps:

1. Install the MLNX-NATIVE-ESX-ConnectX-4/ConnectX-5 driver for ESXi that supports SR-IOV.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

2. Download the MFT package. Go to:
   www.mellanox.com → Products → Software → InfiniBand/VPI Drivers → MFT (http://www.mellanox.com/page/management_tools)
3. Install MFT.

```
# esxcli software vib install -v <MST Vib>
# esxcli software vib install -v <MFT Vib>
```

4. Reboot system.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

5. Start the mst driver.

```
# /opt/mellanox/bin/mst start
```

6. Check if SR-IOV is enabled in the firmware.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

```
/opt/mellanox/bin/mlxconfig -d /dev/mst/mt4115_pciconf0 q

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

  Device type:    ConnectX4
  PCI device:     /dev/mst/mt4115_pciconf0
  Configurations:         Current
     SRIOV_EN             1
     NUM_OF_VFS           8
     FPP_EN               1
```

If not, use mlxconfig to enable it.
Note: The example below shows how to enable SR-IOV and allow the creation of 16 VFs.

```
mlxconfig -d /dev/mst/mt4115_pciconf0 set SRIOV_EN=1 NUM_OF_VFS=16
```

7. Set the number of Virtual Functions you need to create for the PF using the max_vfs module parameter.

   Note: The example below shows the creation of 8 VFs.

```
esxcli system module parameters set -m nmlx5_core -p "max_vfs=8"
```

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

8. Reboot the server.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

> The number of max_vf is set per port. See the "nmlx5_core Module Parameters" table in the introduction, for more information.

4.2.3 Assigning a Virtual Function to a Virtual Machine in the vSphere Web Client

After you enable the Virtual Functions on the host, each of them becomes available as a PCI device.

➢ ***To assign Virtual Function to a Virtual Machine in the vSphere Web Client:***

1. Locate the Virtual Machine in the vSphere Web Client.
   a. Select a data center, folder, cluster, resource pool, or host and click the Related Objects tab.
   b. Click Virtual Machines and select the virtual machine from the list.
2. Power off the Virtual Machine.
3. On the **Manage** tab of the Virtual Machine, select **Settings > VM Hardware**.
4. Click **Edit** and choose the **Virtual Hardware** tab.
5. From the **New Device** drop-down menu, select **Network** and click **Add**.
6. Expand the **New Network** section and connect the Virtual Machine to a port group.
   The virtual NIC does not use this port group for data traffic. The port group is used to extract the networking properties, for example VLAN tagging, to apply on the data traffic.
7. From the **Adapter Type** drop-down menu, select **SR-IOV passthrough**.
8. From the **Physical Function** drop-down menu, select the **Physical Adapter** to back the passthrough Virtual Machine adapter.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 2 + Numbering Style: a, b, c, … + Start at: 1 + Alignment: Left + Aligned at: 0.5" + Tab after: 0.75" + Indent at: 0.75", Tab stops: Not at 1"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

9. **[Optional]** From the **MAC Address** drop-down menu, select **Manual** and type the static MAC address.
10. Use the **Guest OS MTU Change** drop-down menu to allow changes in the MTU of packets from the guest operating system.
    Note: This step is applicable only if this feature is supported by the driver.
11. Expand the **Memory** section, select **Reserve all guest memory (All locked)** and click **OK**. I/O memory management unit (IOMMU) must reach all Virtual Machine memory so that the passthrough device can access the memory by using direct memory access (DMA).
12. Power on the Virtual Machine.

4.3 Mellanox NIC ESXi Management Tools

nmlxcli tools is a Mellanox esxcli command line extension for ConnectX®-3 onwards drivers' management for ESXi 6.0 and later.

This tool enables querying of Mellanox NIC and driver properties directly from driver / firmware.

Once the tool bundle is installed (see "Installing nmlxcli" section below), a new NameSpace named 'mellanox' will be available when executing main #esxcli command, containing additional nested NameSpaces and available commands for each NameSpace.

For general information on 'esxcli' commands usage, syntax, NameSpaces and commands, refer to the VMware vSphere Documentation Center:

https://pubs.vmware.com/vsphere-65/topic/com.vmware.vcli.getstart.doc/GUID-CDD49A32-91DB-454D-8603-3A3E4A09DC59.html

During 'nmlxcli' commands execution, most of the output is formatted using the standard esxcli formatter, thus if required, the option of overriding the standard formatter used for a given command is available, for example:

Executing 'esxcli --formatter=xml mellanox uplink list' produces XML output of given command.

For general information on esxcli generated output formatter, refer to the VMware vSphere Documentation Center:

https://pubs.vmware.com/vsphere-65/topic/com.vmware.vcli.examples.doc/GUID-227F889B-3EC0-48F2-85F5-BF5BD3946AA9.html

> The current implementation does not support private statistics output formatting.

> In case of execution failure, the utility will prompt to standard output or/and log located at '/var/log/syslog.log'.

4.3.1 Requirements

Mellanox 'nmlxcli' tool is compatible with ConnectX-4 onward driver.

4.3.2 Installing nmlxcli

nmlxcli installation is performed as standard offline bundle.

To install nmlxcli:

1. Run

```
esxcli software vib install -d <path_to_nmlxcli_extension_bun- dle.zip>
```

   For general information on updating ESXi from a zip bundle, refer to the VMware vSphere Documentation Center:

   https://pubs.vmware.com/vsphere-65/topic/com.vmware.vsphere.upgrade.doc/GUID-22A4B153-CB21-47B4-974E-2E5BB8AC6874.html

2. For the new Mellanox namespace to function:

- Restart the ESXi host daemon.

```
/etc/init.d/hostd restart
```

or

- reboot ESXi host.

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

5 Troubleshooting

You may be able to easily resolve the issues described in this section. If a problem persists and you are unable to resolve it yourself, please contact your Mellanox representative or Mellanox Support at support@mellanox.com.

5.1 Ethernet Related Issues

Issue	Cause	Solution
No link	Mis-configuration of the switch port or using a cable not supporting link rate	• Ensure the switch port is not down • Ensure the switch port rate is configured to the same rate as the adapter's port
No link with break-out cable	Misuse of the break-out cable or misconfiguration of the switch's split ports	• Use supported ports on the switch with proper configuration. For further information, please refer to the MLNX_OS User Manual • Make sure the QSFP breakout cable side is connected to the SwitchX
Physical link fails to negotiate to maximum supported rate	The adapter is running an outdated firmware	Install the latest firmware on the adapter
Physical link fails to come up	The cable is not connected to the port or the port on the other end of the cable is disabled	Ensure that the cable is connected on both ends or use a known working cable

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

5.2 Installation Related Issues

Issue	Cause	Solution
Driver installation fails.	The install script may fail for the following reasons: • Failed to uninstall the previous installation due to dependencies being used • The operating system is not supported	• Uninstall the previous driver before installing the new one • Use a supported operating system and kernel

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

6 Document Revision History

Release	Date	Description
Rev 4.17.15.16	May 12, 2019	Added the following section: • "Enhanced Network Stack (ENS)"
Rev 4.17.14.2	October 2018	Added the following sections: • Section "Explicit Congestion Notification (ECN)" • Section "Dynamic RSS" • Section "Multiple RSS Engines" Updated the following section: • Section "nmlx5_core Parameters"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

7 Change Log History

Feature/Change	Description
Rev. 4.16.13.5	
Explicit Congestion Notification (ECN)	Explicit Congestion Notification (ECN) is an extension to the Internet Protocol and to the Transmission Control Protocol. ECN allows end-to-end notification of network congestion without dropping packets. To configure ECN behavior, download the nmlxcli tool from the Mellanox site. For further information, refer to the User Manual section *Explicit Congestion Notification (ECN)*.
Bug Fixes	See "Bug Fixes History" section.
Rev. 4.16.12.12	
Packet Capture Utility	Packet Capture utility duplicates all traffic, including RDMA, in its raw Ethernet form (before stripping) to a dedicated "sniffing" QP, and then passes it to an ESX drop capture point. It allows gathering of Ethernet and RoCE bidirectional traffic via pktcap-uw and viewing it using regular Ethernet tools, e.g. Wireshark To enable/disable packet capture, download the nmlxcli tool from the Mellanox site. For further information, refer to the User Manual section Packet Capture Utility.
SR-IOV max_vfs module parameter Type Modification	Changed the type of the SR-IOV max_vfs module parameter from a single integer value to an array of unsigned integers. For further information, refer to the User Manual.
Bug Fixes	See "Bug Fixes History" section.
Rev. 4.16.10.3	
InfiniBand SR-IOV	Enables the creation of InfiniBand virtual functions, allowing the guests to operate over an InfiniBand fabric.
ESXi CLI	Added ESXi CLI support for ESXi 6.5
Rev. 4.16.7.8	
Adapter Cards	Added support for ConnectX-5/ConnectX-5 Ex adapter cards. **Note:** ConnectX-5/ConnectX-5 Ex cards are currently at beta level.
Rev. 4.16.7.8	
Geneve Stateless Offload	Geneve network protocol is encapsulated into IP frame (L2 tunneling). Encapsulation is suggested as a means to alter the normal IP routing for datagrams, by delivering them to an intermediate destination that would otherwise not be selected based on the (network part of the) IP Destination Address field in the original IP header.

Feature/Change	Description
Remote Direct Memory Access (RDMA)	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. **Note:** It is recommended to use RoCE with PFC enabled in driver and network switches. For how to enable PFC in the driver see section *Priority Flow Control (PFC)* in the User Manual.
Set Link Speed	Enables you to set the link speed to a specific link speed supported by ESXi. For further information, see the User Manual section *"Set Link Speed"*.
Priority Flow Control (PFC)	Applies pause functionality to specific classes of traffic on the Ethernet link. For further information, see the User Manual section *"Priority Flow Control (PFC)"*.
NetQ RSS	Allows the user to configure multiple hardware queues backing up the single RX queue. NetQ RSS improves vMotion performance and multiple streams of IPv4/IPv6 TCP/UDP/IPSEC bandwidth over single interface between the Virtual Machines. For further information, see the User Manual section *"NetQ RSS"*.
Default Queue RSS (DRSS)	Allows the user to configure multiple hardware queues backing up the default RX queue. DRSS improves performance for large scale multicast traffic between hypervisors and Virtual Machines interfaces. For further information, see the User Manual section *"Default Queue Receive Side Scaling (DRSS)"*.
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.
	Support for up to 8 ConnectX-4 ports and up to 16 VFs. For further information, refer to the User Manual
RX/TX Ring Resize	Allows the network administrator to set new RX\TX ring buffer size.
VXLAN Hardware Stateless Offloads for ConnectX®-4	VXLAN hardware offload enables the traditional offloads to be performed on the encapsulated traffic.
NetDump	Enables a host to transmit diagnostic information via the network to a remote netdump service, which stores it on disk. Network-based coredump collection can be configured in addition to or instead of disk-based coredump collection.
NetQueue	NetQueue is a performance technology in VMware ESXi that significantly improves performance in Ethernet virtualized environments.
Wake-on-LAN	Allows a network administrator to remotely power on a system or to wake it up from sleep mode

Feature/Change	Description
Hardware Offload	• Large Send Offload (TCP Segmentation Offload) • RSS (Device RSS)
Hardware Capabilities	• Multiple Tx/Rx rings • Fixed Pass-Through • Single/Dual port • MSI-X
Ethernet Network	• TX/RX checksum • Auto moderation and Coalescing • VLAN stripping offload

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25", Tab stops: Not at 0.5"