NVIDIA MLNX_OFED Documentation v4.9-7.1.0.0 LTS

Table of contents

List of Figures

Overview

NVIDIA OpenFabrics Enterprise Distribution for Linux (MLNX_OFED) is a single Virtual Protocol Interconnect (VPI) software stack that operates across all NVIDIA network adapter solutions.

NVIDIA OFED (MLNX_OFED) is an NVIDIA tested and packaged version of OFED and supports two interconnect types using the same RDMA (remote DMA) and kernel bypass APIs called OFED verbs—InfiniBand and Ethernet. Up to 200Gb/s InfiniBand and RoCE (based on the RDMA over Converged Ethernet standard) over 10/25/40/50/100/200GbE are supported with OFED by NVIDIA to enable OEMs and System Integrators to meet the needs end users in the said markets.

Further information on this product can be found in the following MLNX_OFED documents:

- Release Notes
- User Manual

Software Download

Please visit nvidia.com/en-us/networking Products Software InfiniBand/VPI Drivers NVIDIA MLNX_OFED

Document Revision History

For the list of changes made to the User Manual, refer to User Manual Revision History.

For the list of older release notes, refer to Release Notes Revision History.

Release Notes

Release Notes Update History

These are the release notes for MLNX_OFED LTS. This version provides long-term support (LTS) for customers who wish to utilize the following:

- ConnectX-3
- ConnectX-3 Pro
- Connect-IB
- RDMA experimental verbs library (mlnx_lib)

For other use-cases, it is recommended to use the latest MLNX_OFED version 5.x.

Version	Date	Description
4.9-7.1.0.0	June 29, 2023	Initial release of this LTS document version.

Supported NICs Speeds

This document provides instructions on how to install the driver on NVIDIA ConnectX® network adapter solutions supporting the following uplinks to servers.

Uplink/NICs	Driver Name	Uplink Speed
ConnectX®-3/ConnectX-3 Pro	mlx4	• InfiniBand: SDR, QDR, FDR10, FDR • Ethernet: 10GbE, 40GbE 56GbE[1]
ConnectX-4	mlx5	• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE[1], 100GbE

Uplink/NICs	Driver Name	Uplink Speed
ConnectX-4 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE
ConnectX-5/ConnectX-5 Ex		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-6		• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-6 Dx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[1], 100GbE[1], 200GbE[2]
Innova™ IPsec EN		• Ethernet: 10GbE, 40GbE
Connect-IB®		• InfiniBand: SDR, QDR, FDR10, FDR

1. 56GbE is a NVIDIA propriety link speed can be achieved while connecting a NVIDIA adapter card to
   NVIDIA SX10XX switch series, or connecting a NVIDIA adapter card to another NVIDIA adapter card.

2. Supports both NRZ and PAM4 modes.

Package Contents

Package	Revision	Licenses
ar_mgr	1.0-0.2.MLNX20201014.g8577618.49710	Mellanox Confidential and Proprietary
dapl	2.1.10.1.mlnx-OFED.4.9.0.1.4.49710	Dual GPL/BSD/CPL
dump_pr	1.0-0.2.MLNX20201014.g8577618.49710	GPLv2 or BSD
fabric-collector	1.1.0.MLNX20170103.89bb2aa-0.1.49710	GPLv2 or BSD

Package	Revision	Licenses
gpio-mlxbf	1.0-0.g6d44a8a	GPLv2
hcoll	4.4.2970-1.49710	Proprietary
i2c-mlx	1.0-0.g422740c	GPLv2
ibacm	41mlnx1-OFED.4.3.3.0.0.49710	GPLv2 or BSD
ibdump	6.0.0-1.49710	BSD2+GPL2
ibsim	0.10-1.49710	GPLv2 or BSD
ibutils	1.5.7.1-0.12.gdcaeae2.49710	GPL/BSD
ibutils2	2.1.1-0.121.MLNX20200324.g061a520.49710	OpenIB.org BSD license.
infiniband-diags	5.6.0.MLNX20200211.354e4b7-0.1.49710	GPLv2 or BSD
iser	4.9-OFED.4.9.7.1.0.1	GPLv2
isert	4.9-OFED.4.9.7.1.0.1	GPLv2
kernel-mft	4.15.1-9	Dual BSD/GPL
knem	1.1.4.90mlnx2-OFED.23.04.0.5.2.1	BSD and GPLv2
libibcm	41mlnx1-OFED.4.1.0.1.0.49710	GPL/BSD
libibmad	5.4.0.MLNX20190423.1d917ae-0.1.49710	GPLv2 or BSD
libibumad	43.1.1.MLNX20200211.078947f-0.1.49710	GPLv2 or BSD
libibverbs	41mlnx1-OFED.4.9.3.0.0.49710	GPLv2 or BSD
libmlx4	41mlnx1-OFED.4.7.3.0.3.49710	GPLv2 or BSD
libmlx5	41mlnx1-OFED.4.9.0.1.2.49710	GPLv2 or BSD
libpka	1.0-1.g6cc68a2.49710	BSD
librdmacm	41mlnx1-OFED.4.7.3.0.6.49710	GPLv2 or BSD
libvma	9.0.2-1	GPLv2
mlnx-en	4.9-7.1.0.0.g382c630	GPLv2

Package	Revision	Licenses
mlnx-ethtool	5.4-1.49710	GPL
mlnx-iproute2	5.4.0-1.49710	GPL
mlnx-nfsrdma	4.9-OFED.4.9.7.1.0.1	GPLv2
mlnx-nvme	4.9-OFED.4.9.7.1.0.1	GPLv2
mlnx-ofa_kernel	4.9-OFED.4.9.7.1.0.1	GPLv2
mlxbf-livefish	1.0-0.gec08328	GPLv2
mlx-bootctl	1.3-0.g2aa74b7	GPLv2
mlx-l3cache	0.1-1.gebb0728	GPLv2
mlx-pmc	1.1-0.g1141c2e	GPLv2
mlx-trio	0.1-1.g9d13513	GPLv2
mpi-selector	1.0.3-1.49710	BSD
mpitests	3.2.20-e1a0676.49710	BSD
mstflint	4.14.0-3.49710	GPL/BSD
multiperf	3.0-0.14.g5f0fd0e.49710	BSD 3-Clause, GPL v2 or later
multiperf	3.0.0.mlnxlibs-0.13.gcdaa426.49017.49417.49710	BSD 3-Clause, GPL v2 or later
mxm	3.7.3112-1.49710	Proprietary
nvme-snap	2.1.0-126.mlnx	Proprietary
ofed-docs	4.9-OFED.4.9.7.1.0	GPL/BSD
ofed-scripts	4.9-OFED.4.9.7.1.0	GPL/BSD
openmpi	4.0.3rc4-1.49710	BSD
opensm	5.7.2.MLNX20201014.9378048-0.1.49710	GPLv2 or BSD
openvswitch	2.12.1-1.49710	ASL 2.0 and LGPLv2+ and SISSL
perftest	4.5-0.1.g23b8f9c.49710	BSD 3-Clause, GPL v2 or later

Package	Revision	Licenses
perftest	4.5.0.mlnxlibs-0.3.g1121951.49417.49710	BSD 3-Clause, GPL v2 or later
pka-mlxbf	1.0-0.g963f663	GPLv2
qperf	0.4.11-1.49710	BSD 3-Clause, GPL v2
rdma-core	50mlnx1-1.49710	GPLv2 or BSD
rshim	1.18-0.gb99e894	GPLv2
sharp	2.1.2.MLNX20200428.ddda184-1.49710	Proprietary
sockperf	3.7-0.gita1e8e835a689.49710	BSD
srp	4.9-OFED.4.9.7.1.0.1	GPLv2
srptools	41mlnx1-5.49710	GPL/BSD
tmfifo	1.5-0.g31e8a6e	GPLv2
ucx	1.8.0-1.49710	BSD

Release Notes contain the following sections:

General Support in MLNX_OFED

MLNX_OFED Supported Operating Systems

Operating System	Platform	Default Kernel Version
ALIOS7.2	AArch64	4.19.48-006.ali4000.alios7.aarch64
BCLINUX7.3	x86_64	3.10.0-514.el7.x86_64

Operating System	Platform	Default Kernel Version
BCLINUX7.4	x86_64	3.10.0-693.el7.x86_64
BCLINUX7.5	x86_64	3.10.0-862.el7.x86_64
BCLINUX7.6	x86_64	3.10.0-957.el7.x86_64
BCLINUX7.7	x86_64	3.10.0-1062.el7.bclinux.x86_64
BCLINUX8.1	x86_64	4.19.0-193.1.3.el8.bclinux.x86_64
Debian10.0	x86_64	4.19.0-5-arm64
	AArch64	4.19.0-5-amd64
Debian8.11	x86_64	3.16.0-6-amd64
Debian8.9	x86_64	3.16.0-4-amd64
Debian9.11	x86_64	4.9.0-11-amd64
Debian9.6	x86_64	4.9.0-8-amd64
Debian9.9	x86_64	4.9.0-9-amd64
EulerOS2.0sp9	AArch64	4.19.90-vhulk2006.2.0.h171.eulerosv2r9.aarch64
	x86_64	4.18.0-147.5.1.0.h269.eulerosv2r9.x86_64
Fedora30	x86_64	5.0.9-301.fc30.x86_64
Oracle Linux 6.10	x86_64	4.1.12-124.16.4.el6uek.x86_64
Oracle Linux 7.4	x86_64	4.1.12-94.3.9.el7uek.x86_64
Oracle Linux 7.7	x86_64	4.14.35-1902.3.2.el7uek.x86_64
Oracle Linux 7.8	x86_64	4.14.35-1902.300.11.el7uek.x86_64
Oracle Linux 7.9	x86_64	5.4.17-2011.6.2.el7uek.x86_64
Oracle Linux 8.0	x86_64	4.18.0-80.7.2.el8_0.x86_64
Oracle Linux 8.1	x86_64	4.18.0-147.el8.x86_64
Oracle Linux 8.2	x86_64	5.4.17-2011.1.2.el8uek.x86_64
Oracle Linux 8.3	x86_64	5.4.17-2011.7.4.el8uek.x86_64
RHEL/CentOS6.10	x86_64	2.6.32-754.el6.x86_64

Operating System	Platform	Default Kernel Version
RHEL/CentOS6.3	x86_64	2.6.32-279.el6.x86_64
RHEL/CentOS7.2	ppc64	3.10.0-327.el7.ppc64
	ppc64le	3.10.0-327.el7.ppc64le
	x86_64	3.10.0-327.el7.x86_64
RHEL/CentOS7.3	x86_64	3.10.0-514.el7.x86_64
RHEL/CentOS7.4	ppc64	3.10.0-693.el7.ppc64
	ppc64le	3.10.0-693.el7.ppc64le
	x86_64	3.10.0-693.el7.x86_64
RHEL/CentOS7.4alternate	AArch64	4.11.0-44.el7a.aarch64
RHEL/CentOS7.5	ppc64	3.10.0-862.el7.ppc64
	ppc64le	3.10.0-862.el7.ppc64le
	x86_64	3.10.0-862.el7.x86_64
RHEL/CentOS7.5alternate	AArch64	4.14.0-49.el7a.aarch64
RHEL/CentOS7.6	x86_64	3.10.0-957.el7.ppc64
	ppc64le	3.10.0-957.el7.ppc64le
	ppc64	3.10.0-957.el7.x86_64
RHEL/CentOS7.6alternate	AArch64	4.14.0-115.el7a.aarch64
	ppc64le	4.14.0-115.el7a.ppc64le
RHEL/CentOS7.7	ppc64	3.10.0-1062.el7.ppc64
	ppc64le	3.10.0-1062.el7.ppc64le
	x86_64	3.10.0-1062.el7.x86_64
RHEL/CentOS7.8	ppc64	3.10.0-1127.el7.ppc64
	ppc64le	3.10.0-1127.el7.ppc64le
	x86_64	3.10.0-1127.el7.x86_64
RHEL/CentOS7.9	ppc64	3.10.0-1160.el7.ppc64

Operating System	Platform	Default Kernel Version
	ppc64le	3.10.0-1160.el7.ppc64le
	x86_64	3.10.0-1160.el7.x86_64
RHEL/CentOS8.0	AArch64	4.18.0-80.el8.aarch64
	ppc64le	4.18.0-80.el8.ppc64le
	x86_64	4.18.0-80.el8.x86_64
RHEL/CentOS8.1	AArch64	4.18.0-147.el8.aarch64
	ppc64le	4.18.0-147.el8.ppc64le
	x86_64	4.18.0-147.el8.x86_64
RHEL/CentOS8.2	AArch64	4.18.0-193.el8.aarch64
	ppc64le	4.18.0-193.el8.ppc64le
	x86_64	4.18.0-193.el8.x86_64
RHEL/CentOS8.3	AArch64	4.18.0-240.el8.aarch64
	ppc64le	4.18.0-240.el8.ppc64le
	x86_64	4.18.0-240.el8.x86_64
RHEL/CentOS8.4	AArch64	4.18.0-305.el8.aarch64
	ppc64le	4.18.0-305.el8.ppc64le
	x86_64	4.18.0-305.el8.x86_64
RHEL/CentOS8.5	AArch64	4.18.0-348.el8.aarch64
	ppc64le	4.18.0-348.el8.ppc64le
	x86_64	4.18.0-348.el8.x86_64
RHEL/CentOS8.6	AArch64	4.18.0-372.9.1.el8.aarch64
	ppc64le	4.18.0-372.9.1.el8.ppc64le
	x86_64	4.18.0-372.9.1.el8.x86_64
RHEL/CentOS8.7	AArch64	4.18.0-425.3.1.el8.aarch64
	ppc64le	4.18.0-425.3.1.el8.ppc64le
	x86_64	4.18.0-425.3.1.el8.x86_64

Operating System	Platform	Default Kernel Version
RHEL/Rocky 8.8	AArch64	4.18.0-477.10.1.el8_8.aarch64
	ppc64le	4.18.0-477.10.1.el8_8.ppc64le
	x86_64	4.18.0-477.10.1.el8_8.x86_64
SLES11SP3	x86_64	3.0.76-0.11-default
SLES11SP4	ppc64	3.0.101-63-ppc64
	x86_64	3.0.101-63-default
SLES12SP2	x86_64	4.4.21-69-default
SLES12SP3	x86_64	4.4.73-5-default
	ppc64le	4.4.73-5-default
SLES12SP4	x86_64	4.12.14-94.41-default
	ppc64le	4.12.14-94.41-default
	AArch64	4.12.14-94.41-default
SLES12SP5	x86_64	4.12.14-120-default
	ppc64le	4.12.14-120-default
	AArch64	4.12.14-120-default
SLES15SP0	x86_64	4.12.14-23-default
SLES15SP1	x86_64	4.12.14-195-default
	ppc64le	4.12.14-195-default
	AArch64	4.12.14-195-default
SLES15SP2	x86_64	5.3.18-22-default
	ppc64le	5.3.18-22-default
	AArch64	5.3.18-22-default
SLES15SP3	x86_64	5.3.18-57-default
	ppc64le	5.3.18-57-default
	AArch64	5.3.18-57-default
Ubuntu14.04	x86_64	3.13.0-27-generic

Operating System	Platform	Default Kernel Version
Ubuntu16.04	ppc64le	4.4.0-21-generic
	x86_64	4.4.0-21-generic
Ubuntu18.04	x86_64	4.15.0-20-generic
	ppc64le	4.15.0-20-generic
	AArch64	4.15.0-20-generic
Ubuntu19.04	x86_64	5.0.0-13-generic
Ubuntu19.10	x86_64	5.3.0-19-generic
Ubuntu20.04	x86_64	5.4.0-26-generic
	ppc64le	5.4.0-26-generic
	AArch64	5.4.0-26-generic
Kernel 5.5	x86_64	5.5

Notes:

- 32 bit platforms are no longer supported in MLNX_OFED.
- For RPM based distributions, if you wish to install OFED on a different kernel, you need to create a new ISO image, using mlnx_add_kernel_support.sh script. See the MLNX_OFED User Manual for instructions.
- Upgrading MLNX_OFED on your cluster requires upgrading all of its nodes to the newest version as well.
- All OSs listed above are fully supported in Paravirtualized and SR-IOV Environments with Linux KVM Hypervisor.

Supported Non-Linux Virtual Machines

The following are the supported non-Linux Virtual Machines in this current MLNX_OFED version:

NIC	Windows Virtual Machine Type	WinOF version	Protocol
ConnectX-3	Windows 2012 R2 DC	MLNX_VPI 5.50	IPoIB, ETH

NIC	Windows Virtual Machine Type	WinOF version	Protocol
ConnectX-3 Pro	Windows 2016 DC	MLNX_VPI 5.50	IPoIB, ETH
ConnectX-4	Windows 2012 R2 DC	MLNX_WinOF2 2.40	IB, IPoIB, ETH
ConnectX-4 Lx	Windows 2016 DC	MLNX_WinOF2 2.40	IB, IPoIB, ETH

Support in ASAP2™

> ⚠ **Warning**
>
> Accelerated Switch and Packet Processing (ASAP^2) is not supported in this MLNX_OFED version.

NFS over RDMA (NFSoRDMA) Supported Operating Systems

Below is a list of all the OSs on which NFSoRDMA is supported.

- SLES12 SP4
- SLES12 SP5
- SLES15 SP1
- Ubuntu 18.04.3
- RedHat 7.5
- RedHat 7.6
- RedHat 7.7
- RedHat 7.8
- RedHat 8.0

- RedHat 8.1
- RedHat 8.6

Lustre Versions Supported by MLNX_OFED

- Lustre 2.12.3
- Lustre 2.13.0

Hardware and Software Requirements

The following are the hardware and software requirements of the current MLNX_OFED version.

- Linux operating system
- Administrator privileges on your machine(s)
- Disk Space: 1GB

For the OFED Distribution to compile on your machine, some software packages of your operating
system (OS) distribution are required.
To install the additional packages, run the following commands per OS:

Operating System	**Required Packages Installation Command**
RHEL/Oracle Linux/Fedora	yum install perl pciutils python gcc-gfortran libxml2-python tcsh libnl.i686 libnl expat glib2 tcl libstdc++ bc tk gtk2 atk cairo numactl pkgconfig ethtool lsof
XenServer	yum install perl pciutils python libxml2-python libnl expat glib2 tcl bc libstdc++ tk pkgconfig ethtool
SLES 12	zypper install pkg-config expat libstdc++6 libglib-2_0-0 lib- gtk-2_0-0 tcl libcairo2 tcsh python bc pciutils libatk-1_0-0 tk python-libxml2 lsof libnl3-200 ethtool lsof
SLESS 15	python ethtool libatk-1_0-0 python2-libxml2-python tcsh lib- stdc++6-devel-gcc7 libgtk-2_0-0 tcl libopenssl1_1 libnl3-200 make libcairo2 expat libmnl0 insserv-compat pciutils lsof lib-glib-2_0-0 pkg-config tk

Operating System	Required Packages Installation Command
Ubuntu/Debian	apt-get install perl dpkg autotools-dev autoconf libtool auto- make1.10 automake m4 dkms debhelper tcl tcl8.4 chrpath swig graphviz tcl-dev tcl8.4-dev tk-dev tk8.4-dev bison flex dpatch zlib1g-dev curl libcurl4-gnutls-dev python-libxml2 libvirt-bin libvirt0 libnl-dev libglib2.0-dev libgfortran3 automake m4 pkg-config libnuma logrotate ethtool lsof

Supported NICs Firmware Versions

> ⚠ **Warning**
>
> This MLNX_OFED version provides long term support (LTS) for customers who wish to utilize ConnectX-3, ConnectX-3 Pro and Connect-IB, as well as RDMA experimental verbs library (mlnx_lib). Any MLNX_OFED version starting from v5.1 and above does not support any of the adapter cards mentioned.

This current MLNX_OFED version supports the following Mellanox network adapter cards firmware versions:

NIC	Recommended Firmware Rev.	Additional Firmware Rev. Supported
ConnectX®-3/ConnectX-3 Pro	2.42.5000	2.40.7000
ConnectX-4	12.28.2006	12.27.4000
ConnectX-4 Lx	14.28.2006	14.27.1016
ConnectX-5/ConnectX-5 Ex	16.28.2006	16.27.2008
ConnectX-6	20.28.2006	20.27.2008
ConnectX-6 Dx	22.28.2006	N/A
Innova IPsec EN	16.28.2006	16.27.2008

NIC	Recommended Firmware Rev.	Additional Firmware Rev. Supported
Connect-IB	10.16.1200	10.16.1020

For the official firmware versions, please visit the following site:
https://network.nvidia.com/support/firmware/firmware-downloads/

RDMA CM and RoCE Modes

RoCE Modes Matrix

Software Stack / Inbox Distribution	RoCEv1 (IP Based GIDs) Supported as of Version		RoCEv2 Supported as of Version		RoCEv1 & RoCEv2 (RoCE per GID) Supported as of Version
	ConnectX-3/ ConnectX-3 Pro	ConnectX-4/ ConnectX-4 Lx/ ConnectX-5/ ConnectX-5 Ex	ConnectX-3 Pro	ConnectX-4/ ConnectX-4 Lx/ ConnectX-5/ ConnectX-5 Ex	ConnectX-3 Pro/ConnectX-4/ConnectX-4 Lx/ ConnectX-5/ConnectX-5 Ex
MLNX_OFED	2.1-x.x.x	3.0-x.x.x	2.3-x.x.x	3.0-x.x.x	3.0-x.x.x
Kernel.org	3.14	4.4	4.4	4.4	4.4
RHEL	6.6, 7.0	-	-	-	-
SLES	12	-	-	-	-
Ubuntu	14.04.4, 16.04, 15.10	-	-	-	-

Note: Support for ConnectX-5 and ConnectX-5 Ex adapter cards in MLNX_OFED starts from v4.0.

RDMA CM Default RoCE Mode

The default RoCE mode on which RDMA CM runs is RoCEv2 instead of RoCEv1, starting from MLNX_OFED v4.1. RDMA_CM session requires both the client and server sides to support the same RoCE mode. Otherwise, the client will fail to connect to the server. For further information, refer to RDMA CM and RoCE Version Defaults Community post.

MLNX_OFED Unsupported Functionalities/Features/NICs

The following are the unsupported functionalities/features/NICs in MLNX_OFED:

- ConnectX®-2 Adapter Card
- Relational Database Service (RDS)
- Ethernet over InfiniBand (EoIB) - mlx4_vnic
- mthca InfiniBand driver
- Ethernet IPoIB (eIPoIB)
- Soft-RoCE

Changes and New Features

New Features

The following are the new features that have been added to this version of MLNX_OFED.

Feature	Description
NFSoRDMA on RHEL8.6	Added support for NFSoRDMA support on RHEL8.6 OS.
Bug Fixes	See Bug Fixes section.

For additional information on the new features, please refer to MLNX_OFED User Manual.

Customer Affecting Changes

Change	Description
Installation, NVMe SNAP	Support for NVMe SNAP is discontinued on RHEL7.6-alternate.
Installation, ibutils	Support for ibutils is discontinued on RHEL7.6-alternate.

Bug Fixes

This table lists the bugs fixed in this release.
For the list of old bug fixes, please refer to MLNX_OFED Archived Bug Fixes file at:
http://www.mellanox.com/pdf/prod_software/MLNX_OFED_Archived_Bug_Fixes.pdf

Internal Reference Number	Description
3419761	Memory allocation issue may lead to OOM.
	Memory, OOM
	Discovered in Release: 4.9-2.2.4.0
	Fixed in Release: 4.9-7.1.0.0
3383081	**Description:** When using SLES15SP3 with updated kernel, the OFED build fails.
	Keywords: Installation, SLES15SP3
	Discovered in Release: 4.9-6.0.6.0
	Fixed in Release: 4.9-7.1.0.0
3427527	**Description:** Set coalesce parameters was not supported over old LTS branches.
	Keywords: NetDev, Coalesce Parameters

Internal Reference Number	Description
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-7.1.0.0
3201693	**Description:** In the flow of MR destruction, use-after-free was found.
	Keywords: RDMA
	Fixed in Release: 4.9-6.0.6.0
3203796/3251496	**Description:** Using RHEL8.6 with an updated kernel, causes issues with installation.
	Keywords: Installation
	Fixed in Release: 4.9-6.0.6.0
2944030	**Description:** An issue with the Udev script caused non-NVIDIA devices to be renamed.
	Keywords: ASAP^2, Udev, Naming
	Fixed in Release: 4.9-5.1.0.0
3037901	**Description:** RDMA traffic may fail due to incorrect tracking of outstanding work requests.
	Keywords: RDMA
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-5.1.0.0
2976200	**Description:** On the passive side, when the RDMACM disconnectReq event arrives, if the current state is MRA_REP_RCVD, it needs to cancel the MAD before entering the DREQ_RCVD and TIMEWAIT state, otherwise the destroy_id may block the request until this MAD reaches timeout.
	Keywords: RDMACM, MRA, destroy_id
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-5.1.0.0

Internal Reference Number	Description
2753944	**Description:** On rare occasion, registering a device (ib_register_device()) and loading modules in parallel (in this case (ib_cm)), ay cause a race condition to occur which would stop ib_cm from loading properly.
	Keywords: RDMA, ib_core, Racing Condition
	Fixed in Release: 4.9-5.1.0.0
2802401	**Description:** When removing the nvmet port from configfs caused a use-after-free condition.
	Keywords: nvmet-rdma Module
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-4.1.7.0
2162639	**Description:** MLNX_OFED includes several python tools, such as mlnx_qos, which rely on python modules included in the same package. On Ubuntu 20.04 OS, those are installed into a directory that is not in python modules search path.
	Keywords: mlnx_qos, Ubuntu
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-4.0.8.0
2635628	**Description:** openibd does not load automatically after reboot on Suler2sp9 OS.
	Keywords: openibd, Suler2sp9
	Discovered in Release: 4.9-3.1.5.0
	Fixed in Release: 4.9-4.0.8.0
2748862	**Description:** add-kernel-support flag was not supported on Oracle Linux 7.9 causing an installation issue.
	Keywords: openibd, Euleros2u0sp9
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-4.0.8.0

Internal Reference Number	Description
2748862	**Description:** add-kernel-support flag was not supported on Oracle Linux 7.9 causing an installation issue.
	Keywords: Installation, Oracle Linux 7.9
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-4.0.8.0
2396956	**Description:** Fixed an issue were device under massive load may hit iommu allocation failures. For more information see "RX Page Cache Size Limit" section in the user manual.
	Keywords: Legacy libibverbs
	Discovered in Release: 4.9-2.2.4.0
	Fixed in Release: 4.9-3.1.5.0
2434638	**Description:** Fixed an issue where "ibv_devinfo -v" command did not print some of the MEM_WINDOW capabilities, even though they were supported.
	Keywords: Legacy libibverbs
	Discovered in Release: 4.9-2.2.4.0
	Fixed in Release: 4.9-3.1.5.0
2292762	**Description:** Fixed a kernel panic scenario that may have taken place when using sysfs to cancel the probing of VFs and performing reboot while the VFs are still managed by the mlx5 driver.
	Keywords: Proved VFs
	Discovered in Release: 5.1-2.3.7.1
	Fixed in Release: 5.1-2.5.8.0
2265055	**Description:** Added missing release of the lock held in the traffic class error flow.
	Keywords: mutex_unlock
	Discovered in Release: 4.9-0.1.7.0

Internal Reference Number	Description
	Fixed in Release: 4.9-2.2.4.0
2245228	**Description:** Fixed an issue of a crash when attempting to access roce_enable sysfs in unprobed VFs.
	Keywords: roce_enable, unprobed VFs
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2211311	**Description:** Fixed an issue where Rx port buffers cell size was wrong, leading to wrong buffers size reported by mlnx_qos/netdev qos/buffer_size sysfs.
	Keywords: mlx5e, RX buffers, mlnx_qos
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2143067	**Description:** If Openibd was configured to enable the SRP daemon, it now also enables srp_daemon from rdma-core.
	Keywords: Openibd, SRP daemon, srp_daemon, rdma-core
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2192791	**Description:** Fixed the issue where packages neohost-backend and neohost-sdk were not properly removed by the uninstallation procedure and may have required manual removal before re-installing or upgrading the MLNX_OFED driver.
	Keywords: NEO-Host, SDK
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2226715	**Description:** Fixed an issue where bringing up PF interface failed when using SR-IOV and configuring RoCE mode for v2 only.
	Keywords: PF, SR-IOV, RoCE v2

Internal Reference Number	Description
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2242041	**Description:** Fixed leak of memory pages when using ODP.
	Keywords: ODP
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2249090	**Description:** Fixed an issue where IBV_EXP_ACCESS_TUNNELED_ATOMIC capability did not work for ibv_exp_reg_mr experimental verb.
	Keywords: ibv_exp_reg_mr
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2178677	**Description:** Marked "ibacm-devel" as a package name explicitly to avoid accidentally including a symlink to it in the UPSTREAM_LIBS version of MLNX_OFED.
	Keywords: ibacm-devel, UPSTREAM_LIBS
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2255829	**Description:** Fixed an issue with metadata packages generation in the eth-only directory. This allows using the directory as a repository for package managers.
	Keywords: Metadata packages
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2328754	**Description:** Fixed an issue that may have caused panic during peer memory invalidation flow.
	Keywords: GPUDirect

Internal Reference Number	Description
	Discovered in Release: 4.9-0.1.7.0
	Fixed in Release: 4.9-2.2.4.0
2119017	**Description:** Fixed the issue where injecting EEH may cause extra Kernel prints, such as: “EEH: Might be infinite loop in mlx5_core driver”.
	Keywords: EEH, kernel
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2076546	**Description:** Fixed the issue where in RPM-based OSs with non-default kernels, using repositories after re-creating the installer (using --add-kernel-support) would result in improper installation of the drivers.
	Keywords: Installation, OS
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2114957	**Description:** Fixed the issue where MLNX_OFED installation may have depended on python2 package even when attempting to install it on OSs whose default package is python3.
	Keywords: Installation, python
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2143258	**Description:** Fixed a typo in perftest package where help messages wrongly displayed the conversion result between Gb/s and MB/s (20^2 instead of 2^20).
	Keywords: perftest
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0

Internal Reference Number	Description
2094216	**Description:** Fixed the issue of when one of the LAG slaves went down, LAG deactivation failed, ultimately causing bandwidth degradation.
	Keywords: RoCE LAG
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2133778	**Description:** The mlx5 driver maintains a subdirectory for every open eth port in /sys/kernel/debug/. For the default network namespace, the sub-directory name is the name of the interface, like "eth8". The new convention for the network interfaces moved to the non-default network namespaces is the interfaces name followed by "@" and the port's PCI ID. For example: "eth8@0000:af:00.3".
	Keywords: Namespace
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2122684	**Description:** Fixed the issue where OFED uninstallation resulted in the removal of dependency packages, such as qemu-system-* (qemu-system-x86).
	Keywords: Uninstallation, dependency, qemu-system-x86
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2135476	**Description:** Added KMP ability to install MLNX_OFED Kernel modules on SLES12 SP5 and SLES15 kernel maintenance updates.
	Keywords: KMP, SLES, kernel
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2149577	**Description:** Fixed the issue where openibd script load used to fail when esp6_offload module did not load successfully.

Internal Reference Number	Description
	Keywords: openibd, esp6_offload
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2163879	**Description:** Added dependency of package mpi-selectors on perl-Getopt-Long system package. On minimal installs of RPM-based OSs, installing mpi-selectros will also install the required system package perl-Getopt-Long.
	Keywords: Dependency, perl-Getopt-Long
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2107532	**Description:** Fixed the issue where in certain rare scenarios, due to Rx page not being replenished, the same page fragment mistakenly became assigned to two different Rx descriptors.
	Keywords: Memory corruption, Rx page recycle
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2116234	**Description:** Fixed the issue where ibsim was missing after OFED installation.
	Keywords: ibsim, installation
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2116233	**Description:** Fixed an issue where ucx-kmem was missing after OFED installation.
	Keywords: ucx-kmem, installation
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0

Internal Reference Number	Description
2107776	**Description:** Fixed a driver load issue with Errata-kernel on SLES15 SP1.
	Keywords: Load, SLES, Errata
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2105536	**Description:** Fixed an issue in the Hairpin feature which prevented adding hairpin flows using TC tool.
	Keywords: Hairpin, TC
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2090321	**Description:** Fixed the issue where WQ queue flushing was not handled properly in the event of EEH.
	Keywords: WQ, EEH
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2076311	**Description:** Fixed a rare kernel crash scenario when exiting an application that uses RMPP mads intensively.
	Keywords: MAD RMPP
	Discovered in Release: 4.0-1.0.1.0
	Fixed in Release: 4.9-0.1.7.0
2096998	**Description:** Fixed the issue where NEO-Host could not be installed from the MLNX_OFED package when working on Ubuntu and Debian OSs.
	Keywords: NEO-Host, Ubuntu, Debian
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0

Internal Reference Number	Description
2057076	**Description:** Fixed the issue where installing MLNX_OFED using --add-kernel-support option did not work over RHEL 8 OSs.
	Keywords: --add-kernel-support, installation, RHEL
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2090186	**Description:** Fixed a possible kernel crash scenario when AER/slot reset in done in parallel to user space commands execution.
	Keywords: mlx5_core, AER, slot reset
	Discovered in Release: 4.3-1.0.1.0
	Fixed in Release: 4.9-0.1.7.0
2093410	**Description:** Added missing ECN configuration under sysfs for PFs in SwitchDev mode.
	Keywords: sysfs, ASAP, SwitchDev, ECN
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
1916029	**Description:** Fixed the issue of when firmware response time to commands became very long, some commands failed upon timeout. The driver may have then triggered a timeout completion on the wrong entry, leading to a NULL pointer call trace.
	Keywords: Firmware, timeout, NULL
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2036394	**Description:** Added driver support for kernels with the old XDP_REDIRECT infrastructure that uses the following NetDev operations: .ndo_xdp_flush and .ndo_xdp_xmit.
	Keywords: XDP_REDIRECT, Soft lockup
	Discovered in Release: 4.7-3.2.9.0

Internal Reference Number	Description
	Fixed in Release: 4.9-0.1.7.0
2072871	**Description:** Fixed an issue where the usage of --excludedocs Open MPI RPM option resulted in the removal of non-documentation related files.
	Keywords: --excludedocs, Open MPI, RPM
	Discovered in Release: 4.5-1.0.1.0
	Fixed in Release: 4.9-0.1.7.0
2072884	**Description:** Removed all cases of automated loading of MLNX_OFED kernel modules outside of openibd to preserve the startup process of previous MLNX_OFED versions. These loads conflict with openibd, which has its own logic to overcome issues. Such issues can be inbox driver load instead of MLNX_OFED, or module load with wrong parameter value. They might also load modules while openibd is trying to unload the driver stack.
	Keywords: Installation, openibd
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2052037	**Description:** Disabled automated loading of some modules through udev triggers to preserve the startup process of previous MLNX_OFED versions.
	Keywords: Installation, udev
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2022634	**Description:** Fixed a typo in the packages build command line which could cause the installation of MLNX_OFED on SLES OSs to fail when using the option --without-depcheck.
	Keywords: Installation, SLES
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0

Internal Reference Number	Description
2022619	**Description:** Fixed the issue where uninstallation of MLNX_OFED would hang due to a bug in the package dependency check.
	Keywords: Uninstallation, dependency
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2047221	**Description:** Reference count (refcount) for RDMA connection ID (cm_id) was not incremented in rdma_resolve_addr() function, resulting in a cm_id use-after-free access. A fix was applied to increment the cm_id refcount.
	Keywords: rdma_resolve_addr(), cm_id
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.9-0.1.7.0
2045181	**Description:** Fixed a race condition which caused kernel panic when moving two ports to SwitchDev mode at the same time.
	Keywords: ASAP, SwitchDev, race
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2004488	**Description:** Allowed accessing sysfs hardware counters in SwitchDev mode.
	Keywords: ASAP, hardware counters, sysfs, SwitchDev
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2030943	**Description:** Function smp_processor_id() is called in the RX page recycle flow to determine the core to run on. This is intended to run in NAPI context. However, due to a bug in backporting, the RX page recycle was mistakenly called also in the RQ close flow when not needed.
	Keywords: Rx page recycle, smp_processor_id

Internal Reference Number	Description
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.9-0.1.7.0
2074487	**Description:** Fixed an issue where port link state was automatically changed (without admin state involvement) to "UP" after reboot.
	Keywords: Link state, UP
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2022618	**Description:** Fixed a hang with ConnectX-3 adapter cards when running over SLES 11 OS.
	Keywords: OS, SLES, ConnectX-3
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.9-0.1.7.0
2064711	**Description:** Fixed an issue where RDMA CM connection failed when port space was small.
	Keywords: RDMA CM
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
2076424	**Description:** Traffic mirroring with OVS offload and non-offload over VxLAN interface is now supported. **Note**: For kernel 4.9, make sure to use a dedicated OVS version.
	Keywords: VxLAN, OVS
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
1828321	**Description:** Fixed the issue of when working with VF LAG while the bond device is in active-active mode, running fwreset would result in unequal traffic on both PFs, and PFs would not reach line rate.
	Keywords: VF LAG, bonding, PF

Internal Reference Number	Description
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.9-0.1.7.0
1975293	**Description:** Installing OFED with --with-openvswitch flag no longer requires manual removal of the existing Open vSwitch.
	Keywords: OVS, Open vSwitch, openvswitch
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
1939719	**Description:** Fixed an issue of when running openibd restart after the installation of MLNX_OFED on SLES12 SP5 and SLES15 SP1 OSs with the latest Kernel (v4.12.14) resulted in an error that the modules did not belong to that Kernel. This was due to the fact that the module installed by MLNX_OFED was incompatible with new Kernel's module.
	Keywords: SLES, operating system, OS, installation, Kernel, module
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
2001966	**Description:** Fixed an issue of when bond was created over VF netdevices in SwitchDev mode, the VF netdevice would be treated as representor netdevice. This caused the mlx5_core driver to crash in case it received netdevice events related to bond device.
	Keywords: PF, VF, SwitchDev, netdevice, bonding
	Discovered in Release: 4.7-3.2.9.0
	Fixed in Release: 4.9-0.1.7.0
1816629	**Description:** Fixed an issue where following a bad affinity occurrence in VF LAG mode, traffic was sent after the port went up/down in the switch.
	Keywords: Traffic, VF LAG
	Discovered in Release: 4.6-1.0.1.1

Internal Reference Number	Description
	Fixed in Release: 4.9-0.1.7.0
1718531	**Description:** Added support for VLAN header rewrite on CentOS 7.2 OS.
	Keywords: VLAN, ASAP, switchdev, CentOS 7.2
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.9-0.1.7.0
1556337	**Description:** Fixed the issue where adding VxLAN decapsulation rule with enc_tos and enc_ttl failed.
	Keywords: VxLAN, decapsulation
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.9-0.1.7.0
1949260	**Description:** Fixed a race condition that resulted in kernel panic when running IPoIB traffic in Connected mode.
	Keywords: IPoIB
	Discovered in Release: 4.5-1.0.1.0
	Fixed in Release: 4.7-3.2.9.0
1973828	**Description:** Fixed wrong EEPROM length for small form factor (SFF) 8472 from 256 to 512 bytes.
	Keywords: EEPROM, SFF
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.7-3.2.9.0
1915553	**Description:** Fixed the issue where errno field was not sent in all error flows of ibv_reg_mr API.
	Keywords: ibv_reg_mr
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-3.2.9.0

Internal Reference Number	Description
1970901	**Description:** Fixed the issue where mlx5 IRQ name did not change to express the state of the interface.
	Keywords: Ethernet, PCIe, IRQ
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.7-3.2.9.0
1915587	**Description:** Udaddy application is now functional in Legacy mode.
	Keywords: Udaddy, MLNX_OFED legacy, RDMA-CM
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.7-3.2.9.0
1931421	**Description:** Added support for E-Switch (SR-IOV Legacy) mode in RHEL 7.7 OSs.
	Keywords: E-Switch, SR-IOV, RHEL, RedHat
	Discovered in Release: 4.7-1.0.0.1
	Fixed in Release: 4.7-3.2.9.0
1945411/1839353	**Description:** Fixed the issue of when XDP_REDIRECT fails, pages got double-freed due to a bug in the refcnt_bias feature.
	Keywords: XDP, XDP_REDIRECT, refcnt_bias
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-3.2.9.0
1547200	**Description**: Fixed an issue where IPoIB Tx queue may get stuck, leading to timeout warnings in dmesg.
	Keywords: IPoIB
	Discovered in Release: 4.5-1.0.1.0
	Fixed in Release: 4.7-1.0.0.1
1817636	**Description:** Fixed the issue of when disabling one port on the Server side, VF-LAG Tx Affinity would not work on the Client side.

Internal Reference Number	Description
	Keywords: VF-LAG, Tx Affinity
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-1.0.0.1
1800525	**Description:** When configuring the Time-stamping feature, CQE compression will be disabled. This fix entails the removal of a warning message that appeared upon attempting to disable CQE compression when it has already been disabled.
	Keywords: Time-stamping, CQE compression
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-1.0.0.1
1431282	**Description**: Fixed the issue where software reset may have resulted in an order inversion of interface names.
	Keywords: Software reset
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.7-1.0.0.1
1843020	**Description:** Server reboot may result in a system crash.
	Keywords: reboot, crash
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.7-1.0.0.1
1734102	**Description:** Fixed the issue where Ubuntu v16.04.05 and v16.04.05 OSs could not be used with their native kernels.
	Keywords: Ubuntu, Kernel, OS
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-1.0.0.1
1811973	**Description:** VF mirroring offload is now supported.
	Keywords: ASAP^2, VF mirroring

Internal Reference Number	Description
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-1.0.0.1
1841634	**Description:** The number of guaranteed counters per VF is now calculated based on the number of ports mapped to that VF. This allows more VFs to have counters allocated.
	Keywords: Counters, VF
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.7-1.0.0.1
1758983	**Description:** Installing MLNX_OFED on RHEL 7.6 OSs platform x86_64 and RHEL 7.6 ALT OSs platform PPCLE using YUM is now supported.
	Keywords: RHEL, RedHat, YUM, OS, operating system
	Discovered in Release: 4.6-1.0.1.1
	Fixed in Release: 4.7-1.0.0.1
1523548	**Description:** Fixed the issue where RDMA connection persisted even after dropping the network interface.
	Keywords: Network interface, RDMA
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.6-1.0.1.1
1712870	**Description:** Fixed the issue where small packets with non-zero padding were wrongly reported as "checksum complete" even though the padding was not covered by the csum calculation. These packets now report "checksum unnecessary". In addition, an ethtool private flag has been introduced to control the "checksum complete" feature: ethtool --set-priv-flags eth1 rx_no_csum_complete on/off
	Keywords: csum error, checksum, mlx5_core
	Discovered in Release: 4.5-1.0.1.0

Internal Reference Number	Description
	Fixed in Release: 4.6-1.0.1.1
1648597	**Description**: Fixed the wrong wording in the FW tracer ownership startup message (from "FW Tracer Owner" to "FWTracer: Ownership granted and active").
	Keywords: FW Tracer
	Discovered in Release: 4.5-1.0.1.0
	Fixed in Release: 4.6-1.0.1.1
1581631	**Description**: Fixed the issue where GID entries referenced to by a certain user application could not be deleted while that user application was running.
	Keywords: RoCE, GID
	Discovered in Release: 4.5-1.0.1.0
	Fixed in Release: 4.6-1.0.1.1
1368390	**Description**: Fixed the issue where MLNX_OFED could not be installed on RHEL 7.x Alt OSs using YUM repository.
	Keywords: Installation, YUM, RHEL
	Discovered in Release: 4.3-3.0.2.1
	Fixed in Release: 4.6-1.0.1.1
1531817	**Description**: Fixed an issue of when the number of channels configured was less than the number of CPUs available, part of the CPUs would not be used by Tx queues.
	Keywords: Performance, Tx, CPU
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.5-1.0.1.0
1571977	**Description**: Fixed an issue of when the same CQ is connected to some QPs with SRQ and some without, wrong *wr_id* might be reported by *ibv_poll_cq* .
	Keywords: libmlx5, wr_id

Internal Reference Number	Description
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.5-1.0.1.0
1380135	**Description**: Fixed the issue where IB port link used to flap due to MAD heartbeat response delay when using new CQ API.
	Keywords: IB port link, CQ API, MAD heartbeat
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.5-1.0.1.0
1498931	**Description**: Fixed the issue where establishing TCP connection took too long due to failure of SA PathRecord query callback handler.
	Keywords: TCP, SA PathRecord
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.5-1.0.1.0
1514096	**Description**: Fixed the issue where lack of high order allocations caused driver load failure. All high order allocations are now changed to order-0 allocations.
	Keywords: mlx5, high order allocation
	Discovered in Release: 4.0-2.0.2.0
	Fixed in Release: 4.5-1.0.1.0
1524932	**Description**: Fixed a backport issue on some OSs, such as RHEL v7.x, where mlx5 driver would support *ip link set DEVICE vf NUM rate TXRATE* old command, instead of *ip link set DEVICE vf NUM max_tx_rate TXRATE min_tx_rate TXRATE* new command.
	Keywords: mlx5 driver
	Discovered in Release: 4.0-2.0.2.0
	Fixed in Release: 4.5-1.0.1.0
1498585	**Description**: Fixed the issue of when performing configuration changes, mlx5e counters values were reset.

Internal Reference Number	Description
	Keywords: Ethernet counters
	Discovered in Release: 4.0-2.0.2.0
	Fixed in Release: 4.5-1.0.1.0
1484603	**Description**: Fixed the issue of when using ibv_exp_cqe_ts_to_ns verb to convert a packet's hardware timestamp to UTC time in nanoseconds, the result may appear backwards compared to the converted time of a previous packet.
	Keywords: libibverbs
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.5-1.0.1.0
1425027	**Description**: Fixed the issue where attempting to establish a RoCE connection on the default GID or on IPv6 link-local address might have failed when two or more netdevices that belong to HCA ports were slaves under a bonding master. This might also have resulted in the following error message in the kernel log: "*__ib_cache_gid_add: unable to add gid fe80:0000:0000:0000:f652:14ff:fe46:7391 error=-28* ".
	Keywords: RoCE, bonding
	Discovered in Release: 4.4-1.0.0.0
	Fixed in Release: 4.5-1.0.1.0
1480206	**Description**: Modified mlx5_ib SRQs behavior. Now the SRQs are allocated to "order 1" pages instead of contiguous ones to lower the probability of out-of-memory scenarios.
	Keywords: SRQ, mlx5_ib
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.4-2.0.7.0
1363375	**Description**: Modified mlx5_ib QPs behavior. Now the QPs are allocated to "order 1" pages instead of contiguous ones to lower the probability of out-of-memory scenarios.

Internal Reference Number	Description
	Keywords: IPoIB, mlx5_ib
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.4-2.0.7.0
1332080	**Description**: Modified mlx4_ib QPs behavior. Now the QPs are allocated to “order 1” pages instead of contiguous ones to lower the probability of out-of-memory scenarios.
	Keywords: IPoIB, mlx4_ib
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.4-1.0.0.0
1412468	**Description**: Added support for multi-host connection on mstflint’s mstfwreset.
	Keywords: mstfwreset, mstflint, MFT, multi-host
	Discovered in Release: 4.3-1.0.1.0
	Fixed in Release: 4.4-1.0.0.0
1423319	**Description**: Removed the following prints on server shutdown: mlx5_core 0005:81:00.1: mlx5_enter_error_state:96:(pid1): start mlx5_core 0005:81:00.1: mlx5_enter_error_state:109:(pid1): end
	Keywords: mlx5, fast shutdown
	Discovered in Release: 4.3-1.0.1.0
	Fixed in Release: 4.4-1.0.0.0
1433092	**Description**: Fixed an issue of when querying for IBV_EXP_VALUES_HW_CLOCK_NS (using ibv_exp_query_values function) without querying for IBV_EXP_VALUES_HW_CLOCK, 0 value was returned.
	Keywords: mlx5, CQE time-stamping
	Discovered in Release: 4.3-1.0.1.0
	Fixed in Release: 4.4-1.0.0.0

Internal Reference Number	Description
1318251	**Description**: Fixed the issue of when bringing mlx4/mlx5 devices up or down, a call trace in *nvme_rdma_remove_one* or *nvmet_rdma_remove_one* may occur.
	Keywords: NVMEoF, mlx4, mlx5, call trace
	Discovered in Release: 4.3-1.0.1.0
	Fixed in Release: 4.4-1.0.0.0
1181815	**Description**: Fixed an issue where 4K UD packets were dropped when working with 4K MTU on mlx4 devices.
	Keywords: mlx4, 4K MTU, UD
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.3-1.0.1.0
1247458	**Description**: Added support for VLAN Tag (VST) creation on RedHat v7.4 with new iproute2 packages (iptool).
	Keywords: SR-IOV, VST, RedHat
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.3-1.0.1.0
1229554	**Description**: Enabled RDMA CM to honor incoming requests coming from ports of different devices.
	Keywords: RDMA CM
	Discovered in Release: 4.2-1.0.0.0
	Fixed in Release: 4.3-1.0.1.0
1262257	**Description**: Fixed an issue where sending Work Requests (WRs) with multiple entries where the first entry is less than 18 bytes used to fail.
	Keywords: ConnectX-5; libibverbs; Raw QP
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.3-1.0.1.0

Internal Reference Number	Description
1249358/1261023	**Description**: Fixed the issue of when the interface was down, ethtool counters ceased to increase. As a result, RoCE traffic counters were not always counted.
	Keywords: Ethtool counters, mlx5
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.3-1.0.1.0
1244509	**Description**: Fixed compilation errors of MLNX_OFED over kernel when CONFIG_PTP_1588_CLOCK parameter was not set.
	Keywords: PTP, mlx5e
	Discovered in Release: 4.2-1.2.0.0
	Fixed in Release: 4.3-1.0.1.0
1266802	**Description**: Fixed an issue where the system used to hang when trying to allocate multiple device memory buffers from different processes simultaneously.
	Keywords: Device memory programming
	Discovered in Release: 4.2-1.0.0.0
	Fixed in Release: 4.3-1.0.1.0
1120424	**Description**: Fixed incorrect SGE number of RSS QP.
	Keywords: RSS, SGE
	Discovered in Release: 4.1-1.0.2.0
	Fixed in Release: 4.2-1.0.0.0
1078887	**Description**: Fixed an issue where post_list and CQ_mod features in perftest did not function when running the *--run_infinitely* flag.
	Keywords: perftest, --run_infinitely
	Discovered in Release: 4.2-1.0.1.0
	Fixed in Release: 4.2-1.2.0.0

Internal Reference Number	Description
1186260	**Description**: Fixed the issue where CNP counters exposed under /sys/class/infiniband/mlx5_bond_0/ports/1/hw_counters/ did not aggregate both physical functions when working in RoCE LAG mode.
	Keywords: RoCE, LAG, ECN, Congestion Counters
	Discovered in Release: 4.2-1.0.1.0
	Fixed in Release: 4.2-1.2.0.0
1178129	**Description**: Fixed an issue that prevented Windows virtual machines running over MLNX_OFED Linux hypervisors from operating ConnectX-3 IB ports. When such failures occurred, the following message (or similar) appeared in the Linux HV message log when users attempted to start up a Windows VM running a ConnectX-3 VF: "mlx4_core 0000:81:00.0: vhcr command 0x1a slave:1 in_param 0x793000 in_mod=0x210 op_mod=0x0 failed with error:0, status -22"
	Keywords: SR-IOV, RDMA, VM, KVM, Windows
	Discovered in Release: 4.2-1.0.1.0
	Fixed in Release: 4.2-1.2.0.0
1192374	**Description**: Fixed wrong calculation of *max_device_ctx* capability in ConnectX-4, ConnectX-4 Lx, and ConnectX-5 HCAs.
	Keywords: ibv_exp_query_device, max_device_ctx mlx5
	Discovered in Release: 4.2-1.0.1.0
	Fixed in Release: 4.2-1.2.0.0
1084791	**Description**: Fixed the issue where occasionally, after reboot, rpm commands used to fail and create a core file, with messages such as "Bus error (core dumped)", causing the openibd service to fail to start.
	Keywords: rpm, openibd
	Discovered in Release: 3.4-2.0.0.0
	Fixed in Release: 4.2-1.0.0.0

Internal Reference Number	Description
960642/960653	**Description**: Added support for *min_tx_rate* and *max_tx_rate* limit per virtual function ConnectX-5 and ConnectX-5 Ex adapter cards.
	Keywords: SR-IOV, mlx5
	Discovered in Release: 4.0-1.0.1.0
	Fixed in Release: 4.2-1.0.0.0
866072/869183	**Description**: Fixed the issue where RoCE v2 multicast traffic using RDMA-CM with IPv4 address was not received.
	Keywords: RoCE
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.2-1.0.0.0
1163835	**Description**: Fixed an issue where *ethtool -P* output was 00:00:00:00:00:00 when using old kernels.
	Keywords: ethtool, Permanent MAC address, mlx4, mlx5
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.2-1.0.0.0
1067158	**Description**: Replaced a few “GPL only” legacy libibverbs functions with upstream implementation that conforms with libibverbs GPL/BSD dual license model.
	Keywords: libibverbs, license
	Discovered in Release: 4.1-1.0.2.0
	Fixed in Release: 4.2-1.0.0.0
1119377	**Description**: Fixed an issue where ACCESS_REG command failure used to appear upon RoCE Multihost driver restart in dmesg. Such an error message looked as follows: *mlx5_core 0000:01:00.0: mlx5_cmd_check:705:(pid 20037): ACCESS_REG(0x805) op_mod(0x0) failed, status bad parameter(0x3), syndrome (0x15c356)*
	Keywords: RoCE, multihost, mlx5

Internal Reference Number	Description
	Discovered in Release: 4.1-1.0.2.0
	Fixed in Release: 4.2-1.0.0.0
1122937	**Description**: Fixed an issue where concurrent client requests got corrupted when working in persistent server mode due to a race condition on the server side.
	Keywords: librdmacm, rping
	Discovered in Release: 4.1-1.0.2.0
	Fixed in Release: 4.2-1.0.0.0
1102158	**Description**: Fixed an issue where client side did not exit gracefully in RTT mode when the server side was not reachable.
	Keywords: librdmacm, rping
	Discovered in Release: 4.1-1.0.2.0
	Fixed in Release: 4.2-1.0.0.0
1038933	**Description**: Fixed a backport issue where IPv6 procedures were called while they were not supported in the underlying kernel.
	Keywords: iw_cm
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1064722	**Description**: Added log debug prints when changing HW configuration via DCB. To enable log debug prints, run: *ethtool -s <devname> msglvl hw on/off*
	Keywords: DCB, msglvl
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1013076	**Description**: Fixed the issue where reassembly of packets larger than 64k might have failed when ipfrag threshold was low. This issue was present only on RHEL 6.3, 6.4, 6.5, and Ubuntu 12.04.

Internal Reference Number	Description
	This packet drop could be seen from the netstat tool, indicated by the “packet reassembles failed” counter.
	Keywords: IPoIB, Packet Fragmentation
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1022251	**Description**: Fixed SKB memory leak issue that was introduced in kernel 4.11, and added warning messages to the Soft RoCE driver for easy detection of future SKB leaks.
	Keywords: Soft RoCE
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1044546	**Description**: Fixed the issue where a kernel crash used to occur when RXe device was coupled with a virtual (dummy) device.
	Keywords: Soft RoCE
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1047617	**Description**: Fixed the issue where a race condition in the RoCE GID cache used to cause for the loss of IP-based GIDs.
	Keywords: RoCE, GID
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1006768	**Description**: Fixed the issue where an rdma_cm connection between a client and a server that were on the same host was not possible when working over VLAN interfaces.
	Keywords: RDMACM
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0

Internal Reference Number	Description
801807	**Description**: Fixed an issue where RDMACM connection used to fail upon high connection rate accompanied with the error message: *RDMA_CM_EVENT_UNREACHABLE* .
	Keywords: RDMACM
	Discovered in Release: 3.0-2.0.1
	Fixed in Release: 4.1-1.0.2.0
869768	**Description**: Fixed the issue where SR-IOV was not supported in systems with a page size greater than 16KB.
	Keywords: SR-IOV, mlx5, PPC
	Discovered in Release: 4.0-2.0.0.1
	Fixed in Release: 4.1-1.0.2.0
1155972	**Description**: Fixed mlx4 kernel crash upon server shutdown due to NULL pointer dereference.
	Keywords: mlx4, shutdown
	Discovered in Release: 3.3-1.0.4.0
	Fixed in Release: 4.0-2.0.0.1
919545	**Description**: Fixed the issue of when the Kernel becomes out of memory upon driver start, it could crash on SLES 12 SP2.
	Keywords: mlx_5 Eth Driver
	Discovered in Release: 3.4-2.0.0.0
	Fixed in Release: 4.0-2.0.0.1
970668	**Description**: Fixed the issue where very high stress on DC QP transport might have triggered NMI messages on specific servers.
	Keywords: mlx5 Driver
	Discovered in Release: 4.0-1.0.1.0
	Fixed in Release: 4.0-2.0.0.1

Internal Reference Number	Description
966134	**Description**: Allowed Ethernet VFs to open Raw Ethernet QPs even if RoCE is not supported for the VF.
	Keywords: mlx4_ib
	Discovered in Release: 3.0-1.0.1
	Fixed in Release: 4.0-2.0.0.1
864063	**Description**: Fixed the issue of when Spoof-check may have been turned on for MAC address 00:00:00:00:00:00.
	Keywords: mlx4
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-2.0.0.1
869209	**Description**: Fixed an issue that caused TCP packets to be received in an out of order manner when Large Receive Offload (LRO) is on.
	Keywords: mlx5_en
	Discovered in Release: 3.3-1.0.0.0
	Fixed in Release: 4.0-2.0.0.1
913319	**Description**: Fixed the issue of low performance when creating many address handles.
	Keywords: libibverbs
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-1.0.1.0
912897	**Description**: Added debug prints to *ib_umem_get* function to fix lack of error indication when this function fails.
	Keywords: InfiniBand
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-1.0.1.0
945887	**Description**: [ConnectX-3] Fixed the issue where multicast traffic over Raw Ethernet QP on virtual functions were received on the

Internal Reference Number	Description
	same QP (loopback).
	Keywords: SR-IOV
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-1.0.1.0
920292	**Description**: Fixed three issues in libmlx5 that were found by NVIDIA in the patches that are part of MLNX_OFED v3.4: 1. mlx5_exp_peer_commit_qp returns number of entries = 4 instead of 3. 2. Peer capability check is wrong - should fail the check when there is neither NOR nor GEQ support. 3. Missing break in mlx5_exp_peer_peek_cq. There is now fallthrough in the IBV_EXP_PEER_PEEK_ABSOLUTE case.
	Keywords: libmlx5
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-1.0.1.0
890285	**Description**: Fixed the issue where memory allocation for CQ buffers used to fail when increasing the RX ring size.
	Keywords: mlx5_core
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-1.0.1.0
867094	**Description**: Fixed the issue where MLNX_OFED used to fail to load on 4K page Arm architecture.
	Keywords: Arm
	Discovered in Release: 3.4-1.0.0.0
	Fixed in Release: 4.0-1.0.1.0

Known Issues

The following is a list of general limitations and known issues of the various components of this Mellanox OFED for Linux release.

For the list of old known issues, please refer to Mellanox OFED Archived Known Issues file at:
http://www.mellanox.com/pdf/prod_software/MLNX_OFED_Archived_Known_Issues.pdf

Internal Ref. Number	Issue
2894838	**Description:** Running 'ip link show' command over RHEL8.5 using ConnectX-3 with VFs will print "Truncated VFs" to the screen.
	Workaround: Use the following OFED IP link command: /opt/mellanox/iproute2/sbin/ip link show
	Keywords: IP Link, Virtual Functions, ConnectX-3
	Discovered in Release: 4.9-4.1.7.0
2793596	**Description:** On Sles15Sp3, MFT restart does not work.
	Workaround: Install MFT manually from https://www.mellanox.com/products/adapter-software/firmware-tools.
	Keywords: MFT
	Discovered in Release: 4.9-4.0.8.0
2794326	**Description:** When upgrading MLNX_OFED from 4.9-4 to 5.4-2 GA using Yum installation, the installation fails due to ibutils.
	Workaround: Before the upgrade, remove ibutils manually (and the metapackage with it) using the following command: yum remove ibutils
	Keywords: Installation, ibutils
	Discovered in Release: 4.9-4.0.8.0
2753944	**Description:** On rare occasion, registering a device (ib_register_device()) and loading modules in parallel in this case (ib_cm) , a racing condition may occur which would stop ib_cm from loading properly.

Internal Ref. Number	Issue
	Workaround: Add modprobe.d rules to force the ib_cm driver to load before the mlx4_ib and mlx5_ib drivers: install mlx4_ib { /sbin/modprobe ib_cm; /sbin/modprobe -ignore-install mlx4_ib $CMDLINE_OPTS; } install mlx5_ib { /sbin/modprobe ib_cm; /sbin/modprobe —ignore-install mlx5_ib $CMDLINE_OPTS; }
	Keywords: ib_core, Racing Condition
	Discovered in Release: 4.9-4.0.8.0
2636998	**Description:** When using Debian or Ubuntu operating systems, installing MLNX_OFED with mlnxofedinstall and then proceeding to upgrade with a package manager (apt), the mlnx-rdma-core-dkms package remains installed and fails to rebuild.
	Workaround: Before upgrade, remove mlnx-rdma-rxe-dkms: dpkg --purge mlnx-rdma-rxe-dkms
	Keywords: Upgrade, Debian, Ubuntu, mlnx-rdma-core-dkms
	Discovered in Release: 4.9-3.1.5.0
2338121	**Description:** UCX will not work while running with upstream-libs if librdmacm is not installed.
	Workaround: Install rdmacm or disable VMC (-x HCOLL_MCAST=^vmc).
	Keywords: RDMA
	Discovered in Release: 4.9-2.2.4.0
2440042	**Description:** Using ODP on specific hardware may cause intermittent failures (issue only reproduced on IBM POWER8 S822LC).
	Workaround: If the program fail is seen, disable ODP. Or, to use ODP with ConnectX-4 and above, it is recommended to use MLNX_OFED version 5.2 and above.
	Keywords: ar_mgr; dump_pr; upgrade; installation
	Discovered in Release: 4.9-2.2.4.0
2432304	**Description:** ar_mgr and dump_pr plugin versions are not updated when upgrading the MLNX_OFED version.

Internal Ref. Number	Issue
	Workaround: Prior to upgrading your MLNX_OFED version, make sure to uninstall ar_mgr and dump_pr subnet manager plugins. For example, on Ubuntu systems, run: dpkg --remove mlnx-ofed-all ar-mgr dump-pr
	Keywords: ar_mgr; dump_pr; upgrade; installation
	Discovered in Release: 4.9-2.2.4.0
2339456	**Description:** OFED installation requires the --add-kernel-support flag on some of the Errata kernels of RedHat 7.8.
	Workaround: N/A
	Keywords: Installation, Errata, RedHat, OS
	Discovered in Release: 4.9-2.2.4.0
2328653	**Description:** Dependency between qemu and libibverbs may cause qemu failures after OFED installation on Ubuntu v20.04 or SLES 15.2 KVM systems.
	Workaround: N/A
	Keywords: qemu, libibverbs, installation, OS, Ubuntu, SLES, SUSE
	Discovered in Release: 4.9-2.2.4.0
2345669	**Description:** AliOS installation with add-kernel-support may require the installation of additional packages.
	Workaround: Install the required packages.
	Keywords: Installation
	Discovered in Release: 4.9-2.2.4.0
2312063	**Description:** MKEY_BY_NAME is not supported.
	Workaround: N/A
	Keywords: MKEY_BY_NAME
	Discovered in Release: 4.9-2.2.4.0
2046307	**Description:** Excessive toggling between modes (Connected and Datagram) and interface up and down may cause a crash.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: System crash, mode change
	Discovered in Release: 4.9-0.1.7.0
1550266	**Description:** XDP is not supported over ConnectX-3 and ConnectX-3 Pro adapter cards.
	Workaround: N/A
	Keywords: XDP, ConnectX-3
	Discovered in Release: 4.9-0.1.7.0
2117822	**Description:** On ConnectX-3 and ConnectX-3 Pro adapter cards, no traffic runs between VLANs of any type over VLAN of type ctag (protocol 802.1Q).
	Workaround: N/A
	Keywords: ConnectX-3, VLAN
	Discovered in Release: 4.9-0.1.7.0
2142218	**Description:** On ConnectX-3 and ConnectX-3 Pro adapter cards, driver might hang when found under the following conditions, collectively: • OS kernel is older than 4.10 • Interface is down • CONFIG_NET_RX_BUSY_POLL parameter is set • netdev_ops.ndo_busy_poll is defined
	Workaround: N/A
	Keywords: ConnectX-3
	Discovered in Release: 4.9-0.1.7.0
2156645	**Description:** MLNX_LIBS provider packages, such as libmlx5, cannot be installed simultaneously with ibverbs-providers distribution package when working with Ubuntu and Debian OSs.
	Workaround: Before installing MLNX_OFED of type MLNX_LIBS, make sure that ibverbs-providers package is not installed.

Internal Ref. Number	Issue
	Keywords: MLNX_LIBS, libmlx5, ibverbs-providers, Debian, Ubuntu
	Discovered in Release: 4.9-0.1.7.0
2105447	**Description:** hns_roce warning messages will appear in the dmesg after reboot on Euler2 SP3 OSs.
	Workaround: N/A
	Keywords: hns_roce, dmesg, Euler
	Discovered in Release: 4.9-0.1.7.0
2110321	**Description:** Multiple driver restarts may cause IPoIB soft lockup.
	Workaround: N/A
	Keywords: Driver restart, IPoIB
	Discovered in Release: 4.9-0.1.7.0
2112251	**Description:** On kernels 4.10-4.14, when Geneve tunnel's remote endpoint is defined using IPv6, packets larger than MTU are not fragmented, resulting in no traffic sent.
	Workaround: Define geneve tunnel's remote endpoint using IPv4.
	Keywords: Kernel, Geneve, IPv4, IPv6, MTU, fragmentation
	Discovered in Release: 4.9-0.1.7.0
2119210	**Description:** Multiple driver restarts may cause a stress and result in mlx5 commands check error message in the log.
	Workaround: N/A
	Keywords: Driver restart, syndrome, error message
	Discovered in Release: 4.9-0.1.7.0
2111349	**Description:** Ethtool --show-fec/--get-fec are not supported over ConnectX-6 and ConnectX-6 Dx adapter cards.
	Workaround: N/A
	Keywords: Ethtool, ConnectX-6 Dx
	Discovered in Release: 4.9-0.1.7.0

Internal Ref. Number	Issue
2119984	**Description:** IPsec crypto offloads does not work when ESN is enabled.
	Workaround: N/A
	Keywords: IPsec, ESN
	Discovered in Release: 4.9-0.1.7.0
2102902	**Description:** A kernel panic may occur over RH8.0-4.18.0-80.el8.x86_64 OS when opening kTLS offload connection due to a bug in kernel TLS stack.
	Workaround: N/A
	Keywords: TLS offload, mlx5e
	Discovered in Release: 4.9-0.1.7.0
2111534	**Description:** A Kernel panic may occur over Ubuntu19.04-5.0.0-38-generic OS when opening kTLS offload connection due to a bug in the Kernel TLS stack.
	Workaround: N/A
	Keywords: TLS offload, mlx5e
	Discovered in Release: 4.9-0.1.7.0
2117845	**Description:** Relaxed ordering memory regions are not supported when working with CAPI. Registering memory region with relaxed ordering while CAPI enabled will result in a registration failure.
	Workaround: N/A
	Keywords: Relaxed ordering, memory region, MR, CAPI
	Discovered in Release: 4.9-0.1.7.0
2083942	**Description**: The content of file /sys/class/net/ /statistics/multicast may be out of date and may display values lower than the real values.
	Workaround: Run ethtool -S <NETIF> to show the actual multicast counters and to update the content of file /sys/class/net/ /statistics/multicast.
	Keywords: Multicast counters
	Discovered in Release: 4.9-0.1.7.0

Internal Ref. Number	Issue
2035950	**Description:** An internal error might take place in the firmware when performing any of the following in VF LAG mode, when at least one VF of either PF is still bound/attached to a VM. 1. Removing PF from the bond (using ifdown, ip link or any other function) 2. Attempting to disable SR-IOV
	Workaround: N/A
	Keywords: VF LAG, binding, firmware, FW, PF, SR-IOV
	Discovered in Release: 4.9-0.1.7.0
2094176	**Description:** When running in a large scale in VF-LAG mode, bandwidth may be unstable.
	Workaround: N/A
	Keywords: VF LAG
	Discovered in Release: 4.9-0.1.7.0
2044544	**Description:** When working with OSs with Kernel v4.10, bonding module does not allow setting MTUs larger than 1500 on a bonding interface.
	Workaround: Upgrade your Kernel version to v4.11 or above.
	Keywords: Bonding, MTU, Kernel
	Discovered in Release: 4.9-0.1.7.0
1882932	**Description:** Libibverbs dependencies are removed during OFED installation, requiring manual installation of libraries that OFED does not reinstall.
	Workaround: Manually install missing packages.
	Keywords: libibverbs, installation
	Discovered in Release: 4.9-0.1.7.0
2093746	**Description:** Devlink health dumps are not supported on kernels lower than v5.3.
	Workaround: N/A

Internal Ref. Number	Issue
	Keywords: Devlink, health report, dump
	Discovered in Release: 4.9-0.1.7.0
2020260	**Description:** When changing the Trust mode to DSCP, there is an interval between the change taking effect in the hardware and updating the inline mode of the SQ in the driver. If any traffic is transmitted during this interval, the driver will not inline enough headers, resulting in a CQE error in the NIC.

	Workaround: Set the interface down, change the trust mode, then bring the interface back up.
	`ip link set eth0 down`
	`mlnx_qos -i eth0 --trust dscp`
	`ip link set eth0 up`

	Keywords: DSCP, inline, SQ, CQE
	Discovered in Release: 4.9-0.1.7.0
2083427	**Description:** For kernels with connection tracking support, neigh update events are not supported, requiring users to have static ARPs to work with OVS and VxLAN.
	Workaround: N/A
	Keywords: VxLAN, VF LAG, neigh, ARP
	Discovered in Release: 4.9-0.1.7.0
2043739	**Description:** Userspace RoCE UD QPs are not supported over distributions such as SLES11 SP4 and RedHat 6.10 for which the netlink 3 libraries (libnl-3 and libnl-route3) are not available.
	Workaround: N/A
	Keywords: RoCE UD, QP, SLES, RedHat, RHEL, netlink
	Discovered in Release: 4.9-0.1.7.0
2067746	**Description**: When attaching a second slave to a bond, some bond interface GIDs might disappear.
	Workaround: Re-create and re-configure the bond device.

Internal Ref. Number	Issue
	Keywords: Bond, GID
	Discovered in Release: 4.9-0.1.7.0
-	**Description:** The argparse module is installed by default in Python versions =>2.7 and >=3.2. In case an older Python version is used, the argparse module is not installed by default.
	Workaround: Install the argparse module manually.
	Keywords: Python, MFT, argparse, installation
	Discovered in Release: 4.7-3.2.9.0
1979834	**Description:** When running MLNX_OFED on Kernel 4.10 with ConnectX-3/ConnectX-3 Pro NICs, deleting VxLAN may result in a crash.
	Workaround: Upgrade the Kernel version to v4.14 to avoid the crash.
	Keywords: Kernel, OS, ConnectX-3, VxLAN
	Discovered in Release: 4.7-3.2.9.0
1973238	**Description:** ib_core unload may fail on Ubuntu 18.04.2 OS with the following error message: "Module ib_core is in use"
	Workaround: Stop ibacm.socket using the following commands: systemctl stop ibacm.socket systemctl disable ibacm.socket
	Keywords: ib_core, Ubuntu, ibacm
	Discovered in Release: 4.7-3.2.9.0
1970429	**Description:** With HW offloading in SR-IOV SwitchDev mode, the fragmented ICMP echo request/reply packets (with length larger than MTU) do not function properly. The correct behavior is for the fragments to miss the offloading flow and go to the slow path. However, the current behavior is as follows. • Ingress (to the VM): All echo request fragments miss the corresponding offloading flow, but all echo reply fragments hit the corresponding offloading flow • Egress (from the VM): The first fragment still hits the corresponding offloading flow, and the subsequent fragments miss the

Internal Ref. Number	Issue
	corresponding offloading flow
	Workaround: N/A
	Keywords: HW offloading, SR-IOV, SwitchDev, ICMP, VM, virtualization
	Discovered in Release: 4.7-3.2.9.0
1969580	**Description:** RHEL 6.10 OS is not supported in SR-IOV mode.
	Workaround: N/A
	Keywords: RHEL, RedHat, OS, operating system, SR-IOV, virtualization
	Discovered in Release: 4.7-3.2.9.0
1919335	**Description:** On SLES 11 SP4, RedHat 6.9 and 6.10 OSs, on hosts where OpenSM is running, the low-level driver's internal error reset flow will cause a kernel crash when OpenSM is killed (after the reset occurs). This is due to a bug in these kernels where opening the umad device (by OpenSM) does not take a reference count on the underlying device.
	Workaround: Run OpenSM on a host with a more recent Kernel.
	Keywords: SLES, RedHat, CR-Dump, OpenSM
	Discovered in Release: 4.7-3.2.9.0
1893464	**Description:** ibacm is not tested with MLNX_OFED or its components.
	Workaround: N/A
	Keywords: ibacm, component
	Discovered in Release: 4.7-1.0.0.1
1921981	**Description:** On Ubuntu, Debian and RedHat 8 and above OSS, parsing the mfa2 file using the mstarchive might result in a segmentation fault.
	Workaround: Use mlxarchive to parse the mfa2 file instead.
	Keywords: MFT, mfa2, mstarchive, mlxarchive, Ubuntu, Debian, RedHat, operating system
	Discovered in Release: 4.7-1.0.0.1

Internal Ref. Number	Issue
1921799	**Description:** MLNX_OFED installation over SLES15 SP1 ARM OSs fails unless --add-kernel-support flag is added to the installation command.
	Workaround: N/A
	Keywords: SLES, installation
	Discovered in Release: 4.7-1.0.0.1
1840288	**Description:** MLNX_OFED does not support XDP features on RedHat 7 OS, despite the declared support by RedHat.
	Workaround: N/A
	Keywords: XDP, RedHat
	Discovered in Release: 4.7-1.0.0.1
1919335	**Description:** On SLES 11 SP4, RedHat 6.9 and 6.10 OSs, bringing the OpenSM down after CR-Dump results in a panic.
	Workaround: N/A
	Keywords: SLES, RedHat, CR-Dump, OpenSM
	Discovered in Release: 4.7-1.0.0.1
1821235	**Description:** When using mlx5dv_dr API for flow creation, for flows which execute the "encapsulation" action or "push vlan" action, metadata C registers will be reset to zero.
	Workaround: Use the both actions at the end of the flow process.
	Keywords: Flow steering
	Discovered in Release: 4.7-1.0.0.1
1911130	**Description:** When Offloaded Traffic Sniffer feature is on, the usage of "all default" flow steering rule could cause a deadlock.
	Workaround: N/A
	Keywords: Offloaded Traffic Sniffer, steering, deadlock
	Discovered in Release: 4.7-1.0.0.1
1897199	**Description:** When using the RDMA statistics feature and attempting to unbind a QP from a counter, not including the counter-id as an argument

Internal Ref. Number	Issue
	in the CLI will result in a segmentation fault.
	Workaround: N/A
	Keywords: RDMA, QP, segfault, unbinding
	Discovered in Release: 4.7-1.0.0.1
1869219	**Description**: On Fedora 27 OSs, reboot/shutdown operations may fail after uninstalling the MLNX_OFED package.
	Workaround: N/A
	Keywords: Fedora 27, uninstall, reboot, shutdown
	Discovered in Release: 4.7-1.0.0.1
1892663	**Description:** mlnx_tune script does not support python3 interpreter.
	Workaround: Run mlnx_tune with python2 interpreter only.
	Keywords: mlnx_tune, python3, python2
	Discovered in Release: 4.7-1.0.0.1
1341833	**Description:** On CoreOS, assigning a static IP address to PKeys using ifcfg configuration file option fails after restarting the driver.
	Workaround: Manually run “ifdown” and then “ifup”.
	Keywords: CoresOS, PKey, restart_driver
	Discovered in Release: 4.6-1.0.1.1
1504785	**Description:** A lost interrupt issue in pass-through virtual machines may prevent the driver from loading, followed by printing managed pages errors to the dmesg.
	Workaround: Restart the driver.
	Keywords: VM, virtual machine
	Discovered in Release: 4.6-1.0.1.1
1630228	**Description:** Tunnel stateless offloads are wrongly forbidden for E-Switch manager function.
	Workaround: Set the stateless offloads cap to be permanently '1'.

Internal Ref. Number	Issue
	Keywords: Stateless offloads cap
	Discovered in Release: 4.6-1.0.1.1
1764415	**Description:** Unbinding PFs on LAG devices results in a "Failed to modify QP to RESET" error message.
	Workaround: N/A
	Keywords: RoCE LAG, unbind, PF, RDMA
	Discovered in Release: 4.6-1.0.1.1
1769208	**Description:** Contrary to the standard DSCP mode setting procedure in SR-IOV mode, now, in order for this configuration to take effect, the DSCP trust mode has to be set before the VF is created, and not the other way around.
	Workaround: Make sure to set the DSCP trust mode before creating the VF.
	Keywords: DSCP, trust mode, VF
	Discovered in Release: 4.6-1.0.1.1
1779150	**Description:** Upgrading the MLNX_OFED version over SLES 15 SP0 and SP1 OSs on PPCLE platforms might fail due to an isert-kmp-default issue.
	Workaround: Remove the isert-kmp-default package manually
	Keywords: Installation, SLES, PPCLE
	Discovered in Release: 4.6-1.0.1.1
1806565	**Description:** RoCE default GIDs v1 and v2 are derived from the MAC address of the corresponding netdevice's PCI function, and they resemble the IPv6 address. However, in systems where the IPv6 link local address generated does not depend on the MAC address, RoCEv2 default GID should not be used.
	Workaround: Use RoCEv2 default GID.
	Keywords: RoCE
	Discovered in Release: 4.6-1.0.1.1

Internal Ref. Number	Issue
1834997	**Description:** When working with VF Lag while the bond device is in active-active mode, traffic on both physical ports may not reach line rate.
	Workaround: N/A
	Keywords: VF LAG, bonding, bandwidth degradation, fairness
	Discovered in Release: 4.6-1.0.1.1
1839907	**Description:** In mlx4 devices, enabling RX-FCS offload does not disable LRO, and vice-versa.
	Workaround: Disable the RX-FCS or LRO separately.
	Keywords: Frame Check Sequence (FCS), Large Receive Offload (LRO)
	Discovered in Release: 4.6-1.0.1.1
1735161	**Description:** Innova cards do no support InfiniBand mode.
	Workaround: N/A
	Keywords: Innova, IB, InfiniBand
	Discovered in Release: 4.6-1.0.1.1
1787667	**Description:** NVMe-oF driver of MLNX OFED v4.6-x.x.x.x does not function on SLES12 SP4 and SLES15 SP1 OSs, as they have a built-in NVME driver in the Linux image. Therefore, Mellanox NVME and NVME-oF drivers cannot be loaded. For tracking purposes of this bug, see Bugzilla issue #1150850 and Bugzilla issue #1150846.
	Workaround: Change the kernel configuration of NVMe-oF driver to be "=m" and recompile the kernel.
	Keywords: NVME-oF, NVME, SLES
	Discovered in Release: 4.6-1.0.1.1
1759593	**Description:** OFED installation on XenServer OSs requires using the -u flag.
	Workaround: N/A
	Keywords: Installation, XenServer, OS, operating system

Internal Ref. Number	Issue
	Discovered in Release: 4.6-1.0.1.1
1753629	**Description:** A bonding bug found in Kernels 4.12 and 4.13 may cause a slave to become permanently stuck in BOND_LINK_FAIL state. As a result, the following message may appear in dmesg: bond: link status down for interface eth1, disabling it in 100 ms
	Workaround: N/A
	Keywords: Bonding, slave
	Discovered in Release: 4.6-1.0.1.1
1734102	**Description:** Ubuntu v16.04.05 and v16.04.05 OSs can only be used with Kernels of version 4.4.0-143 or below.
	Workaround: N/A
	Keywords: Ubuntu, Kernel, OS
	Discovered in Release: 4.6-1.0.1.1
1712068	**Description:** Uninstalling MLNX_OFED automatically results in the uninstallation of several libraries that are included in the MLNX_OFED package, such as InfiniBand-related libraries.
	Workaround: If these libraries are required, reinstall them using the local package manager (yum/dnf).
	Keywords: MLNX_OFED libraries
	Discovered in Release: 4.6-1.0.1.1
-	**Description:** Due to changes in libraries, MFT v4.11.0 and below are not forward compatible with MLNX_OFED v4.6-1.0.0.0 and above. Therefore, with MLNX_OFED v4.6-1.0.0.0 and above, it is recommended to use MFT v4.12.0 and above.
	Workaround: N/A
	Keywords: MFT compatible
	Discovered in Release: 4.6-1.0.1.1
1730840	**Description:** On ConnectX-4 HCAs, GID index for RoCE v2 is inconsistent when toggling between enabled and disabled interface modes.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: RoCE v2, GID
	Discovered in Release: 4.6-1.0.1.1
1731005	**Description:** MLNX_OFED v4.6 YUM and Zypper installations fail on RHEL8.0, SLES15.0 and PPCLE OSs.
	Workaround: N/A
	Keywords: YUM, Zypper, installation, RHEL, RedHat, SLES, PPCLE
	Discovered in Release: 4.6-1.0.1.1
1717428	**Description**: On kernels 4.10-4.14, MTUs larger than 1500 cannot be set for a GRE interface with any driver (IPv4 or IPv6).
	Workaround: Upgrade your kernel to any version higher than v4.14.
	Keywords: Fedora 27, gretap, ip_gre, ip_tunnel, ip6_gre, ip6_tunnel
	Discovered in Release: 4.6-1.0.1.1
1748343	**Description**: Driver reload takes several minutes when a large number of VFs exists.
	Workaround: N/A
	Keywords: VF, SR-IOV
	Discovered in Release: 4.6-1.0.1.1
1748537	**Description**: Cannot set max Tx rate for VFs from the ARM.
	Workaround: N/A
	Keywords: Host control, max Tx rate
	Discovered in Release: 4.6-1.0.1.1
1732940	**Description**: Software counters not working for representor net devices.
	Workaround: N/A
	Keywords: mlx5, counters, representors
	Discovered in Release: 4.6-1.0.1.1

Internal Ref. Number	Issue
1733974	**Description**: Running heavy traffic (such as 'ping flood') while bringing up and down other mlx5 interfaces may result in "INFO: rcu_preempt dectected stalls on CPUS/tasks:" call traces.
	Workaround: N/A
	Keywords: mlx5
	Discovered in Release: 4.6-1.0.1.1
1731939	**Description:** Get/Set Forward Error Correction FEC configuration is not supported on ConnectX-6 HCAs with 200Gbps speed rate.
	Workaround: N/A
	Keywords: Forward Error Correction, FEC, 200Gbps
	Discovered in Release: 4.6-1.0.1.1
1715789	**Description:** Mellanox Firmware Tools (MFT) package is missing from Ubuntu v18.04.2 OS.
	Workaround: Manually install MFT.
	Keywords: MFT, Ubuntu, operating system
	Discovered in Release: 4.6-1.0.1.1
1652864	**Description:** On ConnectX-3 and ConnectX-3 Pro HCAs, CR-Dump poll is not supported using sysfs commands.
	Workaround: If supported in your Kernel, use the devlink tool as an alternative to sysfs to achieve CR-Dump support.
	Keywords: mlx4, devlink, CR-Dump
	Discovered in Release: 4.6-1.0.1.1
1699031	**Description:** When attempting to destroy IPoIB bonding interface on PPCLE setups, a leak of resources might occur.
	Workaround: N/A
	Keywords: IPoIB, bonding, PPCLE
	Discovered in Release: 4.6-1.0.1.1

Internal Ref. Number	Issue
-	**Description**: On ConnectX-6 HCAs and above, an attempt to configure advertisement (any bitmap) will result in advertising the whole capabilities.
	Workaround: N/A
	Keywords: 200Gmbps, advertisement, Ethtool
	Discovered in Release: 4.6-1.0.1.1
1699289	**Description:** HW LRO feature is disabled OOB, which results in increased CPU utilization on the Receive side. On ConnectX-5 adapter cards and above, this causes a bandwidth drop for a few streams.
	Workaround: Make sure to enable HW LRO in the driver: ethtool -k <intf> lro ethtool --set-priv-flag <intf> hw_lro on
	Keywords: HW LRO, ConnectX-5 and above
	Discovered in Release: 4.5-1.0.1.0
1583487	**Description**: MPI package is not part of MLNX_OFED package in Fedora 28 OS.
	Workaround: Manually install MPI package.
	Keywords: MPI package, Fedora
	Discovered in Release: 4.5-1.0.1.0
1403313	**Description**: Attempting to allocate an excessive number of VFs per PF in operating systems with kernel versions below v4.15 might fail due to a known issue in the Kernel.
	Workaround: Make sure to update the Kernel version to v4.15 or above.
	Keywords: VF, PF, IOMMU, Kernel, OS
	Discovered in Release: 4.5-1.0.1.0
-	**Description**: NEO-Host is not supported on the following OSs: • SLES12 SP3 • SLES12 SP4 • SLES15 • Fedora 28

Internal Ref. Number	Issue
	• RHEL7.1 • RHEL7.4 ALT (Pegas1.0) • REL 7.5 • RHEL7.6 • XenServer 4.9
	Workaround: N/A
	Keywords: NEO-Host, operating systems
	Discovered in Release: 4.5-1.0.1.0
1521877	**Description**: On SLES 12 SP1 OSs, a kernel tracepoint issue may cause undefined behavior when inserting a kernel module with a wrong parameter.
	Workaround: N/A
	Keywords: mlx5 driver, SLES 12 SP1
	Discovered in Release: 4.5-1.0.1.0
1547200	**Description:** When running IPoIB connected traffic with multicasts in parallel, SKB crashes.
	Workaround: N/A
	Keywords: IPoIB, SKB
	Discovered in Release: 4.5-1.0.1.0
1504073	**Description**: When using ConnectX-5 with LRO over PPC systems, the HCA might experience back pressure due to delayed PCI Write operations. In this case, bandwidth might drop from line-rate to ~35Gb/s. Packet loss or pause frames might also be observed.
	Workaround: Look for an indication of PCI back pressure (“outbound_pci_stalled_wr” counter in ethtools advancing). Disabling LRO helps reduce the back pressure and its effects.
	Keywords: Flow Control, LRO
	Discovered in Release: 4.4-1.0.0.0

Internal Ref. Number	Issue
1424233	**Description**: On RHEL v7.3, 7.4 and 7.5 OSs, setting IPv4-IP-forwarding will turn off LRO on existing interfaces. Turning LRO back on manually using ethtool and adding a VLAN interface may cause a warning call trace.
	Workaround: Make sure IPv4-IP-forwarding and LRO are not turned on at the same time.
	Keywords: IPv4 forwarding, LRO
	Discovered in Release: 4.4-1.0.0.0
1418447	**Description**: When working in IPoIB ULP (non-enhanced) mode, IPv6 may disappear in case ring size is changed dynamically (while the driver is running).
	Workaround: There are three workarounds for this issue: • Perform static configuration of ring size instead of dynamic configuration • In case you have run dynamic configuration, run *ifdown ifup* afterwards • On supported kernels, enable *keep_addr_on_down* IPv6 sysfs parameter before configuring the ring size dynamically
	Keywords: IPoIB, ULP mode, ring size
	Discovered in Release: 4.4-1.0.0.0
1442507	**Description**: Retpoline support in GCC causes an increase in CPU utilization, which results in IP forwarding's 15% performance drop.
	Workaround: N/A
	Keywords: Retpoline, GCC, CPU, IP forwarding, Spectre attack
	Discovered in Release: 4.4-1.0.0.0
1417414	**Description**: When working with old kernel versions that do not include the unregister_netdevice_notifier function fix (introduced in "net: In unregister_netdevice_notifier unregister the netdevices" commit), reloading ib_ipoib module using modprobe will fail with the following error message: " *Cannot allocate memory* ".

Internal Ref. Number	Issue
	Workaround: Reload the driver instead of modprobe by running: */etc/init.d/openibd restart*
	Keywords: IPoIB
	Discovered in Release: 4.4-1.0.0.0
1400381	**Description**: On SLES 11 SP3 PPC64 OSs, a memory allocation issue may prevent the interface from loading after reboot, resulting in a call trace in the message log.
	Workaround: Restart the driver.
	Keywords: SLES11 SP3
	Discovered in Release: 4.4-1.0.0.0
1425129	**Description**: MLNX_OFED cannot be installed on SLES 15 OSs using Zypper repository.
	Workaround: Install MLNX_OFED using the standard installation script instead of Zypper repository.
	Keywords: Installation, SLES, Zypper
	Discovered in Release: 4.4-1.0.0.0
1241056	**Description**: When working with ConnectX-4/ConnectX-5 HCAs on PPC systems with Hardware LRO and Adaptive Rx support, bandwidth drops from full wire speed (FWS) to ~60Gb/s.
	Workaround: Make sure to disable Adaptive Rx when enabling Hardware LRO: *ethtool -C <interface> adaptive-rx off* *ethtool -C <interface> rx-usecs 8 rx-frames 128*
	Keywords: Hardware LRO, Adaptive Rx, PPC
	Discovered in Release: 4.3-1.0.1.0
1090612	**Description**: NVMEoF protocol does not support LBA format with non-zero metadata size. Therefore, NVMe namespace configured to LBA format with metadata size bigger than 0 will cause Enhanced Error Handling (EEH) in PowerPC systems.
	Workaround: Configure the NVMe namespace to use LBA format with zero sized metadata.

Internal Ref. Number	Issue
	Keywords: NVMEoF, PowerPC, EEH
	Discovered in Release: 4.3-1.0.1.0
1243581	**Description**: In switchdev mode, the IB device exposed does not support MADs. As a result, tools such as ibstat that work with MADs will not function properly.
	Workaround: N/A
	Keywords: switchdev, IB representors, mlx5, MADs
	Discovered in Release: 4.3-1.0.1.0
1309621	**Description**: In switchdev mode default configuration, stateless offloads/steering based on inner headers is not supported.
	Workaround: To enable stateless offloads/steering based on inner headers, disable encap by running: devlink dev eswitch show pci/0000:83:00.1 encap disable Or, in case devlink is not supported by the kernel, run: echo none > /sys/kernel/debug/mlx5/<BDF>/compat/encap **Note**: This is a hardware-related limitation.
	Keywords: switchdev, stateless offload, steering
	Discovered in Release: 4.3-1.0.1.0
1268718	**Description**: ConnectX-5 supports up to 62 IB representors. When attempting to move to switchdev mode where more than 62 VFs are initialized, the call will fail with the following error message: " *devlink answers: Invalid argument* "
	Workaround: N/A
	Keywords: ConnectX-5, IB representors
	Discovered in Release: 4.3-1.0.1.0
1275082	**Description**: When setting a non-default IPv6 link local address or an address that is not based on the device MAC, connection establishments over RoCEv2 might fail.
	Workaround: N/A
	Keywords: IPV6, RoCE, link local address

Internal Ref. Number	Issue
	Discovered in Release: 4.3-1.0.1.0
1307336	**Description**: In RoCE LAG mode, when running *ibdev2netdev -v* , the port state of the second port of the mlx4_0 IB device will read “NA” since this IB device does not have a second port.
	Workaround: N/A
	Keywords: mlx4, RoCE LAG, ibdev2netdev, bonding
	Discovered in Release: 4.3-1.0.1.0
1316654	**Description**: PKEY interface receives PTP delay requests without a time-stamp.
	Workaround: Run ptp4l over the parent interface.
	Keywords: PKEY, PTP
	Discovered in Release: 4.3-1.0.1.0
1296355	**Description**: Number of MSI-X that can be allocated for VFs and PFs in total is limited to 2300 on Power9 platforms.
	Workaround: N/A
	Keywords: MSI-X, VF, PF, PPC, SR-IOV
	Discovered in Release: 4.3-1.0.1.0
1294934	**Description**: Firmware reset might cause Enhanced Error Handling (EEH) on Power7 platforms.
	Workaround: N/A
	Keywords: EEH, PPC
	Discovered in Release: 4.3-1.0.1.0
1259293	**Description**: On Fedora 20 operating systems, driver load fails with an error message such as: “ *[185.262460] kmem_cache_sanity_check (fs_ftes_0000:00:06.0): Cache name already exists.* ” This is caused by SLUB allocators grouping multiple slab kmem_cache_create into one slab cache alias to save memory and increase cache hotness. This results in the slab name to be considered stale.

Internal Ref. Number	Issue
	Workaround: Upgrade the kernel version to kernel-3.19.8-100.fc20.x86_64. Note that after rebooting to the new kernel, you will need to rebuild MLNX_OFED against the new kernel version.
	Keywords: Fedora, driver load
	Discovered in Release: 4.3-1.0.1.0
1264359	**Description**: When running perftest (ib_send_bw, ib_write_bw, etc.) in rdma-cm mode, the resp_cqe_error counter under /sys/class/infiniband/mlx5_0/ports/1/hw_counters/resp_cqe_error might increase. This behavior is expected and it is a result of receive WQEs that were not consumed.
	Workaround: N/A
	Keywords: perftest, RDMA CM, mlx5
	Discovered in Release: 4.3-1.0.1.0
1294575	**Description**: Traffic may hang while working in IPoIB SR-IOV environment.
	Workaround: N/A
	Keywords: IPoIB, SR-IOV
	Discovered in Release: 4.3-1.0.1.0
1227577	**Description**: Due to Enhanced IPoIB's lack of priority-based flow control, PTP accuracy may adversely be affected by heavy TCP traffic.
	Workaround: N/A
	Keywords: Enhanced IPoIB, PTP
	Discovered in Release: 4.3-1.0.1.0
1264956	**Description**: Configuring SR-IOV after disabling RoCE LAG using sysfs (/sys/bus/pci/drivers/mlx5_core/ /roce_lag_enable) might result in RoCE LAG being enabled again in case SR-IOV configuration fails.
	Workaround: Make sure to disable RoCE LAG once again.
	Keywords: RoCE LAG, SR-IOV

Internal Ref. Number	Issue
	Discovered in Release: 4.3-1.0.1.0
1263043	**Description**: On RHEL7.4, due to an OS issue introduced in kmod package version 20-15.el7_4.6, parsing the depmod configuration files will fail, resulting in either of the following issues: • Driver restart failure prompting an error message, such as: " *ERROR: Module mlx5_core belong to kernel which is not a part of MLNX_OFED, skipping... "* • nvmet_rdma kernel module dysfunction, despite installing MLNX_OFED using the "--with-nvmf " option. An error message, such as: " *nvmet_rdma: unknown parameter 'offload_mem_start' ignored "* will be seen in *dmesg* output
	Workaround: Go to *RedHat webpage* to upgrade the kmod package version.
	Keywords: driver restart, kmod, kmp, nvmf, nvmet_rdma
	Discovered in Release: 4.2-1.2.0.0
1229160	**Description**: Changing IPoIB Tx/Rx ring size dynamically using ethtool is not permitted.
	Workaround: Use the send_queue_size/recv_queue_size module parameters to change the Tx/Rx ring size.
	Keywords: IPoIB, queue size
	Discovered in Release: 4.2-1.2.0.0
1214477	**Description**: On vRedHat 7.2 operating systems, when Network Manager is enabled, IPoIB interfaces may not get an IPv6 address due to an issue in the Network Manager.
	Workaround: Disable Network Manager or upgrade its version.
	Keywords: Network Manager, IPoIB, IPv6
	Discovered in Release: 4.2-1.2.0.0
-	**Description**: Packet Size (Actual Packet MTU) limitation for IPsec offload on Innova IPsec adapter cards: The current offload implementation does not support IP fragmentation. The original packet size should be such

Internal Ref. Number	Issue
	that it does not exceed the interface's MTU size after the ESP transformation (encryption of the original IP packet which increases its length) and the headers (outer IP header) are added: • Inner IP packet size <= I/F MTU - ESP additions (20) - outer_IP (20) - fragmentation issue reserved length (56) • Inner IP packet size <= I/F MTU - 96 This mostly affects forwarded traffic into smaller MTU, as well as UDP traffic. TCP does PMTU discovery by default and clamps the MSS accordingly.
	Workaround: N/A
	Keywords: Innova IPsec, MTU
	Discovered in Release: 4.2-1.0.0.0
-	**Description**: No LLC/SNAP support on Innova IPsec adapter cards.
	Workaround: N/A
	Keywords: Innova IPsec, LLC/SNAP
	Discovered in Release: 4.2-1.0.0.0
-	**Description**: No support for FEC on Innova IPsec adapter cards. When using switches, there may be a need to change its configuration.
	Workaround: N/A
	Keywords: Innova IPsec, FEC
	Discovered in Release: 4.2-1.0.0.0
955929	**Description**: Heavy traffic may cause SYN flooding when using Innova IPsec adapter cards.
	Workaround: N/A
	Keywords: Innova IPsec, SYN flooding
	Discovered in Release: 4.2-1.0.0.0
-	**Description**: Priority Based Flow Control is not supported on Innova IPsec adapter cards.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: Innova IPsec, Priority Based Flow Control
	Discovered in Release: 4.2-1.0.0.0
-	**Description**: Pause configuration is not supported when using Innova IPsec adapter cards. Default pause is global pause (enabled).
	Workaround: N/A
	Keywords: Innova IPsec, Global pause
	Discovered in Release: 4.2-1.0.0.0
1045097	**Description**: Connecting and disconnecting a cable several times may cause a link up failure when using Innova IPsec adapter cards.
	Workaround: N/A
	Keywords: Innova IPsec, Cable, link up
	Discovered in Release: 4.2-1.0.0.0
-	**Description**: On Innova IPsec adapter cards, supported MTU is between 512 and 2012 bytes. Setting MTU values outside this range might fail or might cause traffic loss.
	Workaround: Set MTU between 512 and 2012 bytes.
	Keywords: Innova IPsec, MTU
	Discovered in Release: 4.2-1.0.0.0
1177196	**Description**: If OpenSM version is 4.8.1 and below, the IB interfaces link remains Down while the "SRIOV_IB_ROUTING_MODE_P1=1" and "SRIOV_IB_ROUTING_MODE_P2=1" flags are enabled in the HCA.
	Workaround: N/A
	Keywords: OpenSM, SR-IOV, IB link
	Discovered in Release: 4.2-1.0.0.0
1118530	**Description**: On kernel versions 4.10-4.13, when resetting sriov_numvfs to 0 on PowerPC systems, the following dmesg warning will appear: mlx5_core <BDF>: can't update enabled VF BAR0

Internal Ref. Number	Issue
	Workaround: Reboot the system to reset *sriov_numvfs* value.
	Keywords: SR-IOV, numvfs
	Discovered in Release: 4.2-1.0.0.0
1125184	**Description**: In old kernel versions, such as Ubuntu 14.04 and RedHat 7.1, VXLAN interface does not reply to ARP requests for a MAC address that exists in its own ARP table. This issue was fixed in the following newer kernel versions: Ubuntu 16.04 and RedHat 7.3.
	Workaround: N/A
	Keywords: ARP, VXLAN
	Discovered in Release: 4.2-1.0.0.0
1171764	**Description**: Connecting multiple ports on the same server to the same subnet (IP/IB) will cause all interfaces connected to that subnet to respond to ARP requests. As a result, wrong ARP replies might be received when trying to resolve IP addresses.
	Workaround: Run the following to make sure only the interface with the requested IP address responds to the ARP request: *sysctl -w net.ipv4.conf.all.arp_ignore=1*
	Keywords: IPoIB, librdmacm, ARP
	Discovered in Release: 4.2-1.0.0.0
1134323	**Description**: When using kernel versions older than version 4.7 with IOMMU enabled, performance degradations and logical issues (such as soft lockup) might occur upon high load of traffic. This is caused due to the fact that IOMMU IOVA allocations are centralized, requiring many synchronization operations and high locking overhead amongst CPUs.

Internal Ref. Number	Issue
	Workaround: Use kernel v4.7 or above, or a backported kernel that includes the following patches: • 2aac630429d9 iommu/vt-d: change intel-iommu to use IOVA frame numbers • 9257b4a206fc iommu/iova: introduce per-cpu caching to iova allocation • 22e2f9fa63b0 iommu/vt-d: Use per-cpu IOVA caching
	Keywords: IOMMU, soft lockup
	Discovered in Release: 4.2-1.0.0.0
1135738	**Description**: On 64k page size setups, DMA memory might run out when trying to increase the ring size/number of channels.
	Workaround: Reduce the ring size/number of channels.
	Keywords: DMA, 64K page
	Discovered in Release: 4.2-1.0.0.0
1159650	**Description**: When configuring VF VST, VLAN-tagged outgoing packets will be dropped in case of ConnectX-4 HCAs. In case of ConnectX-5 HCAs, VLAN-tagged outgoing packets will have another VLAN tag inserted.
	Workaround: N/A
	Keywords: VST
	Discovered in Release: 4.2-1.0.0.0
1157770	**Description**: On Passthrough/VM machines with relatively old QEMU and libvirtd, CMD timeout might occur upon driver load. After timeout, no other commands will be completed and all driver operations will be stuck.
	Workaround: Upgrade the QEMU and libvirtd on the KVM server. Tested with (Ubuntu 16.10) are the following versions: • libvirt 2.1.0 • QEMU 2.6.1

Internal Ref. Number	Issue
	Keywords: QEMU
	Discovered in Release: 4.2-1.0.0.0
1147703	**Description**: Using dm-multipath for High Availability on top of NVMEoF block devices must be done with “directio” path checker.
	Workaround: N/A
	Keywords: NVMEoF
	Discovered in Release: 4.2-1.0.0.0
1152408	**Description**: RedHat v7.3 PPCLE and v7.4 PPCLE operating systems do not support KVM qemu out of the box. The following error message will appear when attempting to run *virt-install* to create new VMs: *Cant find qemu-kvm packge to install*
	Workaround: Acquire the following rpms from the beta version of 7.4ALT to 7.3/7.4 PPCLE (in the same order): • qemu-img-.el7a.ppc64le.rpm • qemu-kvm-common-.el7a.ppc64le.rpm • qemu-kvm-.el7a.ppc64le.rpm
	Keywords: Virtualization, PPC, Power8, KVM, RedHat, PPC64LE
	Discovered in Release: 4.2-1.0.0.0
1012719	**Description**: A soft lockup in the CQ polling flow might occur when running very high stress on the GSI QP (RDMA-CM applications). This is a transient situation from which the driver will later recover.
	Workaround: N/A
	Keywords: RDMA-CM, GSI QP, CQ
	Discovered in Release: 4.2-1.0.0.0
1062940	**Description**: When running Network Manger on devices on which Enhanced IPoIB is enabled, CONNECTED_MODE can only be set to NO/AUTO. Setting it to YES will prevent the interface from being configured.
	Workaround: N/A

Internal Ref. Number	Issue
	Keywords: Enhanced IPoIB, network manager, connected_mode
	Discovered in Release: 4.2-1.0.0.0
1078630	**Description**: When working in RoCE LAG over kernel v3.10, a kernel crash might occur when unloading the driver as the Network Manager is running.
	Workaround: Stop the Network Manager before unloading the driver and start it back once the driver unload is complete.
	Keywords: RoCE LAG, network manager
	Discovered in Release: 4.2-1.0.0.0
1149557	**Description**: When setting VGT+, the maximal number of allowed VLAN IDs presented in the sysfs is 813 (up to the first 813).
	Workaround: N/A
	Keywords: VGT+
	Discovered in Release: 4.2-1.0.0.0
1122619	**Description**: On Arm setups, DMA memory resource is limited due to a default CMA limitation.
	Workaround: Increase the CMA limitation or cancel its use, using the kernel's CMD line parameters: • Add the parameter cma=256M to increase the CMA limit to 256MB • Add the parameter cma=0 to disable the use of CMA
	Keywords: IPoIB, CMA
	Discovered in Release: 4.2-1.0.0.0
1146837	**Description**: On SLES11 SP1 operating system, IPoIB interface renaming process may fail due to a broken udev rule, leaving interfaces with names like ib0_rename.
	Workaround: 1. Open the udev conf file "/etc/udev/rules.d/70-persistent-net.rules", and remove such lines as SUBSYSTEM=="net", ACTION=="add",

Internal Ref. Number	Issue
	DRIVERS=="?*", ==""" , NAME="eth0". 2. Reload the driver stack.
	Keywords: IPoIB
	Discovered in Release: 4.2-1.0.0.0
-	**Description**: NVMEoF support is available for the following: • SLES 12.3 and above • RHEL 7.2 and above (Host side only) • RHEL 7.4 and above (Host and Target side) • OS with distribution/custom kernel >= 4.8.x
	Workaround: N/A
	Keywords: NVMEoF Host/Target
995665/1165919	**Description**: In kernels below v4.13, connection between NVMEoF host and target cannot be established in a hyper-threaded system with more than 1 socket.
	Workaround: On the host side, connect to NVMEoF subsystem using --*nr-io-queues* <*num_queues*> flag. Note that *num_queues* must be lower or equal to *num_sockets* multiplied with num_cores_per_socket.
	Keywords: NVMEoF
1039346	**Description**: Enabling multiple namespaces per subsystem while using NVMEoF target offload is not supported on ConnectX-5 adapter cards.
	Workaround: To enable more than one namespace, create a subsystem for each one.
	Keywords: NVMEoF Target Offload, namespace
1072347	**Description**: Ethtool -i displays incorrect driver name for devices with enhanced IPoIB support.
	Workaround: N/A
	Keywords: Enhanced IPoIB, Ethtool

Internal Ref. Number	Issue
1071457	**Description**: PKEY-related limitations in enhanced IPoIB: • Since the parent interface ib<x> and the child interface ib<x>.yyyy share the same receive resources, the parent interface's MTU cannot be less than the child interface's MTU • Interface counters and Ethtool control are not supported on child interfaces • Parent interface should be in UP state to enable child interface to receive traffic
	Workaround: N/A
	Keywords: PKEY, Enhanced IPoIB, MTU, Ethtool, Interface Counters
1059451	**Description**: When Enhanced IPoIB is enabled, the following module parameters will not be functional: • send_queue_size • recv_queue_size • max_nonsrq_conn_qp
	Workaround: N/A
	Keywords: Enhance IPoIB
1030301	**Description**: Creating virtual functions on a device that is in LAG mode will destroy the LAG configuration. The boding device over the Ethernet NICs will continue to work as expected.
	Workaround: N/A
	Keywords: LAG, SR-IOV
1047616	**Description**: When node GUID of a device is set to zero (0000:0000:0000:0000), RDMA_CM user space application may crash.
	Workaround: Set node GUID to a nonzero value.
	Keywords: RDMA_CM
1061298	**Description**: Since enhanced IPoIB does not support connected mode on RedHat operating systems, when using network manger and

Internal Ref. Number	Issue
	enhanced IPoIB capable devices, *CONNECTED_MODE* must be set to NO/AUTO. Setting *CONNECTED_MODE* to yes will cause the interface to not be configured.
	Workaround: N/A
	Keywords: Enhanced IPoIB
1068215	**Description**: When enhanced IPoIB mode is enabled, ring size limit is 8k. When it is disabled, ring size limit is decreased to 4k.
	Workaround: N/A
	Keywords: Enhanced IPoIB
1051701	**Description**: New versions of iproute which support new kernel features may misbehave on old kernels that do not support these new features.
	Workaround: N/A
	Keywords: iproute
1007830	**Description**: When working on Xenserver hypervisor with SR-IOV enabled on it, make sure the following instructions are applied: 1. Right after enabling SR-IOV, unbind all driver instances of the virtual functions from their PCI slots. 2. It is not allowed to unbind PF driver instance while having active VFs.
	Workaround: N/A
	Keywords: SR-IOV
1008583	**Description**: A soft lockup in the CQ polling flow might occur when running very high stress on the GSI QP (RDMA-CM applications). This is a transient situation and the driver recovers from it after a while.
	Workaround: N/A
	Keywords: RDMA-CM
1007356	**Description**: Creating a PKEY interface using " *ip link* " is not supported.

Internal Ref. Number	Issue
	Workaround: Use sysfs to create a PKEY interface.
	Keywords: IPoIB, PKEY
1000197	**Description**: Displaying multicast groups using sysfs may not show all the entries on Fedora 23 OS.
	Workaround: N/A
	Keywords: IPoIB
1010148	**Description**: Upgrading from MLNX_OFED v3.x to v4.x using yum and apt-get repositories fails.
	Workaround: Remove MLNX_OFED v3.x using the *ofed_uninstall.sh* script, and only then install MLNX_OFED v4.x as usual.
	Keywords: Installation
1005786	**Description**: When using ConnectX-5 adapter cards, the following error might be printed to dmesg, indicating temporary lack of DMA pages: *"mlx5_core ... give_pages:289:(pid x): Y pages alloc time exceeded the max permitted duration* *mlx5_core ... page_notify_fail:263:(pid x): Page allocation failure notification on func_id(z) sent to fw* *mlx5_core ... pages_work_handler:471:(pid x): give fail -12"* **Example**: This might happen when trying to open more than 64 VFs per port.
	Workaround: N/A
	Keywords: mlx5_core, DMA
1008066/1009004	**Description**: Performing some operations on the user end during reboot might cause call trace/panic, due to bugs found in the Linux kernel. For example: Running *get_vf_stats* (via iptool) during reboot.
	Workaround: N/A
	Keywords: mlx5_core, reboot
1009488	**Description**: Mounting MLNX_OFED to a path that contains special characters, such as parenthesis or spaces is not supported. For example,

Internal Ref. Number	Issue
	when mounting MLNX_OFED to “/media/CDROM(vcd)/”, installation will fail and the following error message will be displayed: # cd /media/CDROM\(vcd\)/ # ./mlnxofedinstall sh: 1: Syntax error: "(" unexpected
	Workaround: N/A
	Keywords: Installation
982144	**Description**: When offload traffic sniffer is on, the bandwidth could decrease up to 50%.
	Workaround: N/A
	Keywords: Offload Traffic Sniffer
981045	**Description**: On kernels below v4.2, when removing a bonding module with devices different from ARPHRD_ETHER, a call trace may be received.
	Workaround: Remove the bond in the following order: Remove the slaves, delete the bond, and only then remove the bonding module.
	Keywords: Bonding
980066/981314	**Description**: Soft RoCE does not support Extended Reliable Connection (XRC).
	Workaround: N/A
	Keywords: Soft RoCE, XRC
982534	**Description**: In ConnectX-3, when using a server with page size of 64K, the UAR BAR will become too small. This may cause one of the following issues: 1. mlx4_core driver does not load. 2. The mlx4_core driver does load, but calls to *ibv_open_device* may return ENOMEM errors.
	Workaround: 1. Add the following parameter in the firmware's ini file under [HCA] section: log2_uar_bar_megabytes = 7

Internal Ref. Number	Issue
	2. Re-burn the firmware with the new ini file.
	Keywords: PPC
981362	**Description**: On several OSs, setting a number of TC is not supported via the tc tool.
	Workaround: Set the number of TC via the /sys/class/net/ /qos/tc_num sysfs file.
	Keywords: Ethernet, TC
980257	**Description**: An issue in InfiniBand bond interfaces may cause memory corruption in Ubuntu v14.04 and v14.10 OSs. The memory corruption happens when attempting to reload the driver while the bond is up with InfiniBand salves.
	Workaround: Delete the bond before restarting the driver.
	Keywords: Bonding, IPoIB
980034/981311	**Description**: Soft RoCE counters located under /sys/class/infiniband/ /ports/1/counters/ directory are not supported.
	Workaround: N/A
	Keywords: Soft RoCE
979907	**Description**: Only the following two experimental verbs are supported for Soft RoCE: • ibv_exp_query_device • ibv_exp_poll_cq.
	Workaround: N/A
	Keywords: Soft RoCE
979457	**Description**: When setting IOMMU=ON, a severe performance degradation may occur due to a bug in IOMMU.
	Workaround: Make sure the following patches are found in your kernel: • iommu/vt-d: Fix PASID table allocation • iommu/vt-d: Fix IOMMU lookup for SR-IOV Virtual Functions

Internal Ref. Number	Issue
	Note: These patches are already available in Ubuntu 16.04.02 and 17.04 OSs.
	Keywords: Performance, IOMMU
977852	**Description**: *rdma_cm* running over IB ports does not support UD QPs on ConnectX-3 adapter cards.
	Workaround: N/A
	Keywords: SR-IOV, RDMA CM
955113/977990	**Description**: In RoCE LAG over ConnectX-4 adapter cards, the script ibdev2netdev may show a wrong port state for the bonded device. This means that although the IB device/port mlx5_bond_0/1 is up (as seen in ibstat), ibdev2netdev may report that it is down.
	Workaround: N/A
	Keywords: RoCE, LAG, bonding

User Manual

Introduction

This manual is intended for system administrators responsible for the installation, configuration, management and maintenance of the software and hardware of VPI (InfiniBand, Ethernet) adapter cards. It is also intended for application developers.

This document provides instructions on how to install the driver on NVIDIA ConnectX® network adapter solutions supporting the following uplinks to servers.

Uplink/NICs	Driver Name	Uplink Speed
ConnectX®-3/ConnectX-3 Pro	mlx4	• InfiniBand: SDR, QDR, FDR10, FDR • Ethernet: 10GbE, 40GbE 56GbE[1]
ConnectX-4	mlx5	• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE[1], 100GbE

Uplink/NICs	Driver Name	Uplink Speed
ConnectX-4 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE
ConnectX-5/ConnectX-5 Ex		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-6		• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-6 Dx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[1], 100GbE[1], 200GbE[2]
Innova™ IPsec EN		• Ethernet: 10GbE, 40GbE
Connect-IB®		• InfiniBand: SDR, QDR, FDR10, FDR

1. 56GbE is a NVIDIA propriety link speed can be achieved while connecting a NVIDIA adapter card to
   NVIDIA SX10XX switch series, or connecting a NVIDIA adapter card to another NVIDIA adapter card.

2. Supports both NRZ and PAM4 modes.

All NVIDIA network adapter cards are compatible with OpenFabrics-based RDMA protocols and software and are supported by major operating system distributions.

NVIDIA OFED is certified with the following products:

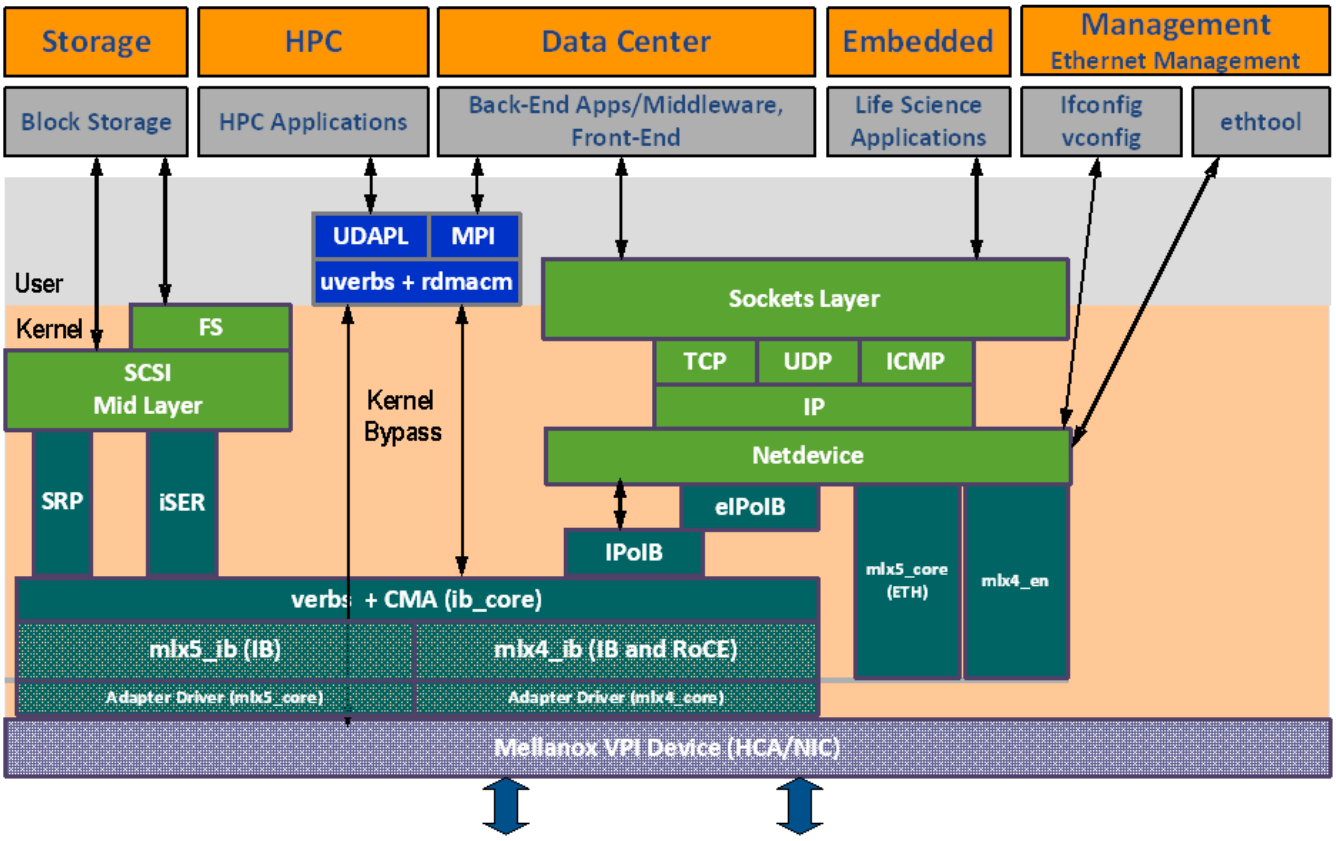
- NVIDIA Messaging Accelerator (VMA™) software: Socket acceleration library that performs OS bypass for standard socket-based applications.
  Please note, VMA support is provided separately from NVIDIA OFED support. For further information, please refer to the VMA documentation (https://docs.nvidia.com/networking/category/vma).

- NVIDIA Unified Fabric Manager (UFM®) software: Powerful platform for managing demanding scale-out computing fabric environments, built on top of the OpenSM industry standard routing engine.

- Fabric Collective Accelerator (FCA)—FCA is a NVIDIA MPI-integrated software package that utilizes CORE-Direct technology for implementing the MPI collectives communications.

Stack Architecture

The figure below shows a diagram of the NVIDIA OFED stack, and how upper layer protocols (ULPs) interface with the hardware and with the kernel and userspace. The application level also shows the versatility of markets that NVIDIA OFED applies to.



The following subsections briefly describe the various components of the NVIDIA OFED stack.

mlx4 VPI Driver

mlx4 is the low-level driver implementation for the ConnectX® family adapters designed by NVIDIA. ConnectX-3 adapters can operate as an InfiniBand adapter, or as an Ethernet

NIC. The OFED driver supports InfiniBand and Ethernet NIC configurations. To accommodate the supported configurations, the driver is split into the following modules:

mlx4_core
Handles low-level functions like device initialization and firmware commands processing. Also controls resource allocation so that the InfiniBand and Ethernet functions can share the device without interfering with each other.
mlx4_ib
Handles InfiniBand-specific functions and plugs into the InfiniBand mid layer.
mlx4_en
A 10/40GigE driver under drivers/net/ethernet/mellanox/mlx4 that handles Ethernet specific functions and plugs into the netdev mid layer.

mlx5 Driver
mlx5 is the low-level driver implementation for the Connect-IB® and ConnectX-4 adapters designed by NVIDIA. Connect-IB® operates as an InfiniBand adapter whereas and ConnectX®-4 operates as a VPI adapter (Infiniband and Ethernet). The mlx5 driver is comprised of the following kernel modules:

mlx5_core
Acts as a library of common functions (e.g. initializing the device after reset) required by Connect-IB® and ConnectX®-4 adapter cards. mlx5_core driver also implements the Ethernet interfaces for ConnectX®-4. Unlike mlx4_en/core, mlx5 drivers do not require the mlx5_en module as the Ethernet functionalities are built-in in the mlx5_core module.
mlx5_ib
Handles InfiniBand-specific functions and plugs into the InfiniBand mid layer.

libmlx5

libmlx5 is the provider library that implements hardware specific user-space functionality. If there is no compatibility between the firmware and the driver, the driver will not load and a message will be printed in the dmesg.

The following are the libmlx5 environment variables:

- MLX5_FREEZE_ON_ERROR_CQE
    - Causes the process to hang in a loop of completion with error, which is not flushed with error or retry exceeded occurs/
    - Otherwise disabled

- MLX5_POST_SEND_PREFER_BF
    - Configures every work request that can use blue flame will use blue flame
- Otherwise - blue flame depends on the size of the message and inline indication in the packet
- MLX5_SHUT_UP_BF
    - Disables blue flame feature
    - Otherwise - do not disable
- MLX5_SINGLE_THREADED
    - All spinlocks are disabled
    - Otherwise - spinlocks enabled
    - Used by applications that are single threaded and would like to save the overhead of taking spinlocks.
- MLX5_CQE_SIZE
    - 64 - completion queue entry size is 64 bytes (default)
    - 128 - completion queue entry size is 128 bytes
- MLX5_SCATTER_TO_CQE
    - Small buffers are scattered to the completion queue entry and manipulated by the driver. Valid for RC transport.
    - Default is 1, otherwise disabled
- MLX5_ENABLE_CQE_COMPRESSION
    - Saves PCIe bandwidth by compressing a few CQEs into a smaller amount of bytes on PCIe. Setting this variable 1enables CQE compression.
    - Default value 0 (disabled)

- MLX5_RELAXED_PACKET_ORDERING_ON
  See "Out-of-Order (OOO) Data Placement Experimental Verbs" section.

Mid-layer Core

Core services include management interface (MAD), connection manager (CM) interface, and Subnet Administrator (SA) interface. The stack includes components for both user-mode and kernel applications. The core services run in the kernel and expose an interface to user-mode for verbs, CM and management.

Upper Layer Protocols (ULPs)

IP over IB (IPoIB)

The IP over IB (IPoIB) driver is a network interface implementation over InfiniBand. IPoIB encapsulates IP datagrams over an InfiniBand connected or datagram transport service. IPoIB pre-appends the IP datagrams with an encapsulation header and sends the outcome over the InfiniBand transport service. The transport service is Unreliable Datagram (UD) by default, but it may also be configured to be Reliable Connected (RC), in case RC is supported. The interface supports unicast, multicast and broadcast. For details, see "IP over InfiniBand (IPoIB)" section.

iSCSI Extensions for RDMA (iSER)

iSCSI Extensions for RDMA (iSER) extends the iSCSI protocol to RDMA. It permits data to be transferred directly into and out of SCSI buffers without intermediate data copies. For further information, please refer to "iSCSI Extensions for RDMA (iSER)" section.

SCSI RDMA Protocol (SRP)

SCSI RDMA Protocol (SRP) is designed to take full advantage of the protocol offload and RDMA features provided by the InfiniBand architecture. SRP allows a large body of SCSI software to be readily used on InfiniBand architecture. The SRP driver—known as the SRP Initiator—differs from traditional low-level SCSI drivers in Linux. The SRP Initiator does not control a local HBA; instead, it controls a connection to an I/O controller—known as the SRP Target—to provide access to remote storage devices across an InfiniBand fabric. The SRP Target resides in an I/O unit and provides storage services. See "SRP - SCSI RDMA Protocol" section.

User Direct Access Programming Library (uDAPL)

User Direct Access Programming Library (uDAPL) is a standard API that promotes data center application data messaging performance, scalability, and reliability over RDMA interconnects InfiniBand and RoCE. The uDAPL interface is defined by the DAT collaborative. This release of the uDAPL reference implementation package for both DAT 1.2 and 2.0 specification is timed to coincide with OFED release of the Open Fabrics (www.openfabrics.org) software stack.

MPI

Message Passing Interface (MPI) is a library specification that enables the development of parallel software libraries to utilize parallel computers, clusters, and heterogeneous networks. OFED includes the following MPI implementation over InfiniBand:

- Open MPI – an open source MPI-2 implementation by the Open MPI Project

OFED also includes MPI benchmark tests such as OSU BW/LAT, Intel MPI BeBenchmarkand Presta.

InfiniBand Subnet Manager

All InfiniBand-compliant ULPs require a proper operation of a Subnet Manager (SM) running on the InfiniBand fabric, at all times. An SM can run on any node or on an IB switch. OpenSM is an InfiniBand-compliant Subnet Manager, and it is installed as part of OFED[1].

1. OpenSM is disabled by default. See "OpenSM" section for details on enabling it.

Diagnostic Utilities

OFED includes the following two diagnostic packages for use by network and data center managers:

- ibutils – NVIDIA diagnostic utilities
- infiniband-diags – OpenFabrics Alliance InfiniBand diagnostic tools

NVIDIA Firmware Tools

The NVIDIA Firmware Tools (MFT) package is a set of firmware management tools for a single InfiniBand node. MFT can be used for:

- Generating a standard or customized NVIDIA firmware image

- Burning a firmware image to a single InfiniBand node

MFT includes a set of tools used for performing firmware update and configuration, as well as debug and diagnostics, and provides MST service. For the full list of available tools within MFT, please refer to MFT documentation (docs.nvidia.com/networking/category/mft).

NVIDIA OFED Package

ISO Image

OFED for Linux (MLNX_OFED_LINUX) is provided as ISO images or as a tarball, one per supported Linux distribution and CPU architecture, that includes *source code* and *binary* RPMs, firmware, utilities, and documentation. The ISO image contains an installation script (called mlnxofedinstall) that performs the necessary steps to accomplish the following:

- Discover the currently installed kernel
- Uninstall any InfiniBand stacks that are part of the standard operating system distribution or another vendor's commercial stack
- Install the MLNX_OFED_LINUX binary RPMs (if they are available for the current kernel)
- Identify the currently installed InfiniBand HCAs and perform the required firmware updates

Software Components

MLNX_OFED_LINUX contains the following software components:

- NVIDIA Host Channel Adapter Drivers
    - mlx5, mlx4 (VPI), which is split into multiple modules:
        - mlx4_core (low-level helper)
        - mlx4_ib (IB)
        - mlx5_ib

        - mlx5_core (includes Ethernet)
        - mlx4_en (Ethernet)
- Mid-layer core
    - Verbs, MADs, SA, CM, CMA, uVerbs, uMADs
- Upper Layer Protocols (ULPs)
    - IPoIB, SRP Initiator and SRP
- MPI
    - Open MPI stack supporting the InfiniBand, RoCE and Ethernet interfaces
    - MPI benchmark tests (OSU BW/LAT, Intel MPI Benchmark, Presta)
- OpenSM: InfiniBand Subnet Manager
- Utilities
    - Diagnostic tools
    - Performance tests
    - Sysinfo (see Sysinfo User Manual)
- Firmware tools (MFT)
- Source code for all the OFED software modules (for use under the conditions mentioned in the modules' LICENSE files)
- Documentation

Firmware

The ISO image includes the following firmware item:

- mlnx-fw-updater RPM/DEB package, which contains firmware binaries for supported devices (using mlxfwmanager tool).

Directory Structure

The ISO image of MLNX_OFED_LINUX contains the following files and directories:

- mlnxofedinstall - This is the MLNX_OFED_LINUX installation script.
- ofed_uninstall.sh - This is the MLNX_OFED_LINUX un-installation script.
- <RPMS folders> - Directory of binary RPMs for a specific CPU architecture.
- src/ - Directory of the OFED source tarball.

> ⚠ **Warning**
>
> MLNX_OFED includes the OFED source RPM packages used as a build platform for kernel code but does not include the sources of NVIDIA proprietary packages.

- mlnx_add_kernel_support.sh - Script required to rebuild MLNX_OFED_LINUX for customized kernel version on supported Linux Distribution
- RPM based - A script required to rebuild MLNX_OFED_LINUX for customized kernel version on supported RPM-based Linux Distribution
- docs/ - Directory of NVIDIA OFED related documentation

Module Parameters

mlx4 Module Parameters

In order to set **mlx4** parameters, add the following line(s) to **/etc/modprobe.d/mlx4.conf**:

```
options mlx4_core parameter=<value>
```

and/or

```
options mlx4_ib parameter=<value>
```

and/or

```
options mlx4_en parameter=<value>
```

The following sections list the available mlx4 parameters.

mlx4_ib Parameters

sm_guid_assign:	Enable SM alias_GUID assignment if sm_guid_assign > 0 (Default: 0) (int)
dev_assign_str:[1]	Map device function numbers to IB device numbers (e.g.'0000:04:00.0-0,002b:1c:0b.a-1,...'). Hexadecimal digits for the device function (e.g. 002b:1c:0b.a) and decimal for IB device numbers (e.g. 1). Max supported devices - 32 (string)
en_ecn	Enable q/ecn [enable = 1, disable = 0 (default)] (bool)

1. In the current version, this parameter is using a decimal number to describe the InfiniBand device, and not a hexadecimal number as in previous versions.
   The purpose is to uniform the mapping of device function numbers with InfiniBand device numbers, as defined for other module parameters (e.g. num_vfs and probe_vf). For example, to map mlx4_15 to device function number 04:00.0 in the current version, we use “options mlx4_ib dev_assign_str=04:00.0-15”, as opposed to the previous version where we used “options mlx4_ib dev_assign_str=04:00.0-f”

mlx4_core Parameters

debug_level	Enable debug tracing if > 0 (int)
msi_x	0 - don't use MSI-X, 1 - use MSI-X, >1 - limit number of MSI-X irqs to msi_x (non-SRIOV only) (int)
enable_sys_tune	Tune the cpu's for better performance (default 0) (int)

block_loopback	Block multicast loopback packets if > 0 (default: 1) (int)
num_vfs	Either a single value (e.g. '5') to define uniform num_vfs value for all devices functions or a string to map device function numbers to their num_vfs values (e.g. '0000:04:00.0-5,002b:1c:0b.a-15'). Hexadecimal digits for the device function (e.g. 002b:1c:0b.a) and decimal for num_vfs value (e.g. 15). (string)
probe_vf	Either a single value (e.g. '3') to indicate that the Hypervisor driver itself should activate this number of VFs for each HCA on the host, or a string to map device function numbers to their probe_vf values (e.g. '0000:04:00.0-3,002b:1c:0b.a-13'). Hexadecimal digits for the device function (e.g. 002b:1c:0b.a) and decimal for probe_vf value (e.g. 13). (string)
log_num_mgm_entry_size	log mgm size, that defines the num of qp per mcg, for example: 10 gives 248.range: 7 <= log_num_mgm_entry_size <= 12. To activate device managed flow steering when available, set to -1 (int)
high_rate_steer	Enable steering mode for higher packet rate (obsolete, set "Enable optimized steering" option in log_num_mgm_entry_size to use this mode). (int)
fast_drop	Enable fast packet drop when no recieve WQEs are posted (int)
enable_64b_cqe_eqe	Enable 64 byte CQEs/EQEs when the FW supports this if non-zero (default: 1) (int)
log_num_mac	Log2 max number of MACs per ETH port (1-7) (int)
log_num_vlan	(Obsolete) Log2 max number of VLANs per ETH port (0-7) (int)
log_mtts_per_seg	Log2 number of MTT entries per segment (0-7) (default: 0) (int)
port_type_array	Either pair of values (e.g. '1,2') to define uniform port1/port2 types configuration for all devices functions or a string to map device function numbers to their pair of port types values (e.g. '0000:04:00.0-1;2,002b:1c:0b.a-1;1'). 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').

log_num_qp	log maximum number of QPs per HCA (default: 19) (int)
log_num_srq	log maximum number of SRQs per HCA (default: 16) (int)
log_rdmarc_per_qp	log number of RDMARC buffers per QP (default: 4) (int)
log_num_cq	log maximum number of CQs per HCA (default: 16) (int)
log_num_mcg	log maximum number of multicast groups per HCA (default: 13) (int)
log_num_mpt	log maximum number of memory protection table entries per HCA (default: 19) (int)
log_num_mtt	log maximum number of memory translation table segments per HCA (default: max(20, 2*MTTs for register all of the host memory limited to 30)) (int)
enable_qos	Enable Quality of Service support in the HCA (default: off) (bool)
internal_err_reset	Reset device on internal errors if non-zero (default is 1) (int)
ingress_parser_mode	Mode of ingress parser for ConnectX3-Pro. 0 - standard. 1 - checksum for non TCP/UDP. (default: standard) (int)
roce_mode	Set RoCE modes supported by the port
ud_gid_type	Set gid type for UD QPs
log_num_mgm_entry_size	log mgm size, that defines the num of qp per mcg, for example: 10 gives 248.range: 7 <= log_num_mgm_entry_size <= 12 (default = -10).
use_prio	Enable steering by VLAN priority on ETH ports (deprecated) (bool)
enable_vfs_qos	Enable Virtual VFs QoS (default: off) (bool)

mlx4_en_only_mode	Load in Ethernet only mode (int)
enable_4k_uar	Enable using 4K UAR. Should not be enabled if have VFs which do not support 4K UARs (default: true) (bool)
mlx4_en_only_mode	Load in Ethernet only mode (int)
rr_proto	IP next protocol for RoCEv1.5 or destination port for RoCEv2. Setting 0 means using driver default values (deprecated) (int)

mlx4_en Parameters

inline_thold	The threshold for using inline data (int) Default and max value is 104 bytes. Saves PCI read operation transaction, packet less then threshold size will be copied to hw buffer directly. (range: 17-104)
udp_rss:	Enable RSS for incoming UDP traffic (uint) On by default. Once disabled no RSS for incoming UDP traffic will be done.
pfctx	Priority-based Flow Control policy on TX[7:0]. Per priority bit mask (uint)
pfcrx	Priority-based Flow Control policy on RX[7:0]. Per priority bit mask (uint)
udev_dev_port_dev_id	Work with dev_id or dev_port when supported by the kernel. Range: 0 <= udev_dev_port_dev_id <= 2 (default = 0).
udev_dev_port_dev_id:	Work with dev_id or dev_port when supported by the kernel. Range: 0 <= udev_dev_port_dev_id <= 2 (default = 0). • 0: Work with dev_port if supported by the kernel, otherwise work with dev_id. • 1: Work only with dev_id regardless of dev_port support. • 2: Work with both of dev_id and dev_port (if dev_port is supported by the kernel). (int)

mlx5_core Module Parameters

The mlx5_core module supports a single parameter used to select the profile which defines the number of resources supported.

prof_sel	The parameter name for selecting the profile. The supported values for profiles are: • 0 - for medium resources, medium performance • 1 - for low resources • 2 - for high performance (int) (default)
guids	charp
node_guid	guids configuration. This module parameter will be obsolete!
debug_mask	debug_mask: 1 = dump cmd data, 2 = dump cmd exec time, 3 = both. Default=0 (uint)
probe_vf	probe VFs or not, 0 = not probe, 1 = probe. Default = 1 (bool)
num_of_groups	Controls the number of large groups in the FDB flow table. Default=4; Range=1-1024

ib_core Parameters

send_queue_size	Size of send queue in number of work requests (int)
recv_queue_size	Size of receive queue in number of work requests (int)
force_mr	Force usage of MRs for RDMA READ/WRITE operations (bool)
roce_v1_noncompat_gid	Default GID auto configuration (Default: yes) (bool)

ib_ipoib Parameters

max_nonsrq_conn_qp	Max number of connected-mode QPs per interface (applied only if shared receive queue is not available) (int)
mcast_debug_level	Enable multicast debug tracing if > 0 (int)
send_queue_size	Number of descriptors in send queue (int)
recv_queue_size	Number of descriptors in receive queue (int)
debug_level	Enable debug tracing if > 0 (int)

ipoib_enhanced	Enable IPoIB enhanced for capable devices (default = 1) (0-1) (int)

Devlink Parameters

The following parameters, supported in mlx4 driver only, can be changed using the Devlink user interface:

Parameter	Description	Parameter Type
internal_error_reset	Enables resetting the device on internal errors	Generic
max_macs	Max number of MACs per ETH port	Generic
region_snapshot_enable	Enables capturing region snapshots	Generic
enable_64b_cqe_eqe	Enables 64 byte CQEs/EQEs when supported by FW	Driver-specific
enable_4k_uar	Enables using 4K UAR	Driver-specific

Device Capabilities

Normally, an application needs to query the device capabilities before attempting to create a resource. It is essential for the application to be able to operate over different devices with different capabilities.

Specifically, when creating a QP, the user needs to specify the maximum number of outstanding work requests that the QP supports. This value should not exceed the queried capabilities. However, even when you specify a number that does not exceed the queried capability, the verbs can still fail since some other factors such as the number of scatter/gather entries requested, or the size of the inline data required, affect the maximum possible work requests. Hence an application should try to decrease this size (halving is a good new value) and retry until it succeeds.

Installation

This chapter describes how to install and test the Mellanox OFED for Linux package on a single host machine with Mellanox InfiniBand and/or Ethernet adapter hardware installed.

The chapter contains the following sections:

- Hardware and Software Requirements
- Downloading MLNX_OFED
- Installing MLNX_OFED
- Uninstalling MLNX_OFED
- Updating Firmware After Installation
- UEFI Secure Boot
- Performance Tuning

Features Overview and Configuration

The chapter contains the following sections:

- Ethernet Network
- InfiniBand Network
- Storage Protocols
- Virtualization
- Resiliency
- Docker Containers
- HPC-X™
- Fast Driver Unload

- OVS Offload Using ASAP² Direct

Programming

> **Warning**
>
> This chapter is aimed for application developers and expert users that wish to develop applications over MLNX_OFED.

Raw Ethernet Programming

Raw Ethernet programming enables writing an application that bypasses the kernel stack. To achieve this, packet headers and offload options need to be provided by the application.
For a basic example on how to use Raw Ethernet programming, refer to the Raw Ethernet Programming: Basic Introduction—Code Example Community post.

Packet Pacing

Packet pacing is a raw Ethernet sender feature that enables controlling the rate of each QP, per send queue.
For a basic example on how to use packet pacing per flow over libibverbs, refer to Raw Ethernet Programming: Packet Pacing—Code Example Community post.

TCP Segmentation Offload (TSO)

TCP Segmentation Offload (TSO) enables the adapter cards to accept a large amount of data with a size greater than the MTU size. The TSO engine splits the data into separate packets and inserts the user-specified L2/L3/L4 headers automatically per packet. With the usage of TSO, CPU is offloaded from dealing with a large throughput of data.
To be able to program that on the sender side, refer to the Raw Ethernet Programming: TSO—Code Example Community post.

ToS Based Steering

ToS/DSCP is an 8-bit field in the IP packet that enables different service levels to be assigned to network traffic. This is achieved by marking each packet in the network with a DSCP code and appropriating the corresponding level of service to it.
To be able to steer packets according to the ToS field on the receiver side, refer to the Raw Ethernet Programming: ToS—Code Example Community post.

Flow ID Based Steering

Flow ID based steering enables developing a code that will steer packets using flow ID when developing Raw Ethernet over verbs. For more information on flow ID based steering, refer to the Raw Ethernet Programming: Flow ID Steering—Code Example Community post.

VXLAN Based Steering

VXLAN based steering enables developing a code that will steer packets using the VXLAN tunnel ID when developing Raw Ethernet over verbs. For more information on VXLAN based steering, refer to the Raw Ethernet Programming: VXLAN Steering—Code Example Community post.

Device Memory Programming

Warning

This feature is supported on ConnectX-5/ConnectX-5 Ex adapter cards and above only.

Device Memory is an API that allows using on-chip memory located on the device as a data buffer for send/receive and RDMA operations. The device memory can be mapped and accessed directly by user and kernel applications, and can be allocated in various sizes, registered as memory regions with local and remote access keys for performing the send/receive and RDMA operations.
Using the device memory to store packets for transmission can significantly reduce transmission latency compared to the host memory.

Device Memory Programming Model

The new API introduces a similar procedure to the host memory for sending packets from the buffer:

- ibv_alloc_dm()/ibv_free_dm() - to allocate/free device memory
- ibv_reg_dm_mr - to register the allocated device memory buffer as a memory region and get a memory key for local/remote access by the device
- ibv_memcpy_to_dm - to copy data to a device memory buffer
- ibv_memcpy_from_dm - to copy data from a device memory buffer
- ibv_post_send/ibv_post_receive - to request the device to perform a send/receive operation using the memory key

For examples, see Device Memory.

RDMA-CM QP Timeout Control

RDMA-CM QP Timeout Control feature enables users to control the QP timeout for QPs created with RDMA-CM.

A new option in 'rdma_set_option' function has been added to enable overriding calculated QP timeout, in order to provide QP attributes for QP modification. To achieve that, rdma_set_option() should be called with the new flag RDMA_OPTION_ID_ACK_TIMEOUT.
Example:

```
rdma_set_option(cma_id, RDMA_OPTION_ID, RDMA_OPTION_ID_ACK_TIMEOUT,
&timeout, sizeof(timeout));
```

RDMA-CM Application Managed QP

Applications which do not create a QP through rdma_create_qp() may want to postpone the ESTABLISHED event on the passive side, to let the active side complete an application-specific connection establishment phase. For example, modifying the init state of the QP created by the application to RTR state, or make some preparations for receiving messages from the passive side. The feature returns a new event on the active side: CONNECT_RESPONSE, instead of ESTABLISHED, if id->qp==NULL. This gives the application a chance to perform the extra connection setup. Afterwards, the new rdma_establish() API should be called to complete the connection and generate an ESTABLISHED event on the passive side.

In addition, this feature exposes the 'rdma_init_qp_attr' function in librdmacm API, which enables applications to get the parameters for creating Address Handler (AH) or control QP attributes after its creation.

InfiniBand Fabric Utilities

This section first describes common configuration, interface, and addressing for all the tools in the package.

Common Configuration, Interface and Addressing

Topology File (Optional)

An InfiniBand fabric is composed of switches and channel adapter (HCA/TCA) devices. To identify devices in a fabric (or even in one switch system), each device is given a GUID (a MAC equivalent). Since a GUID is a non-user-friendly string of characters, it is better to alias it to a meaningful, user-given name. For this objective, the IB Diagnostic Tools can be provided with a “topology file”, which is an optional configuration file specifying the IB fabric topology in user-given names.
For diagnostic tools to fully support the topology file, the user may need to provide the local system name (if the local hostname is not used in the topology file).
To specify a topology file to a diagnostic tool use one of the following two options:

1. On the command line, specify the file name using the option '-t <topology file name>'

2. Define the environment variable IBDIAG_TOPO_FILE

To specify the local system name to an diagnostic tool use one of the following two options:

1. On the command line, specify the system name using the option '-s <local system name>'

2. Define the environment variable IBDIAG_SYS_NAME

InfiniBand Interface Definition

The diagnostic tools installed on a machine connect to the IB fabric by means of an HCA port through which they send MADs. To specify this port to an IB diagnostic tool use one of the following options:

1. On the command line, specify the port number using the option '-p <local port number>' (see below)

2. Define the environment variable IBDIAG_PORT_NUM

In case more than one HCA device is installed on the local machine, it is necessary to specify the device's index to the tool as well. For this use on of the following options:

1. On the command line, specify the index of the local device using the following option: '-i <index of local device>'

2. Define the environment variable IBDIAG_DEV_IDX

Addressing

⚠ **Warning**

This section applies to the ibdiagpath tool only. A tool command may require defining the destination device or port to which it applies.

The following addressing modes can be used to define the IB ports:

- Using a Directed Route to the destination: (Tool option '-d')
  This option defines a directed route of output port numbers from the local port to the destination.

- Using port LIDs: (Tool option '-l'):
  In this mode, the source and destination ports are defined by means of their LIDs. If the fabric is configured to allow multiple LIDs per port, then using any of them is valid for defining a port.

- Using port names defined in the topology file: (Tool option '-n')
  This option refers to the source and destination ports by the names defined in the topology file. (Therefore, this option is relevant only if a topology file is specified to the tool.) In this mode, the tool uses the names to extract the port LIDs from the matched topology, then the tool operates as in the '-l' option.

Diagnostic Utilities

The diagnostic utilities described in this chapter provide means for debugging the connectivity and status of InfiniBand (IB) devices in a fabric.

Diagnostic Utilities

Utility	Description
dump_fts	Dumps tables for every switch found in an ibnetdiscover scan of the subnet. The dump file format is compatible with loading into OpenSM using the -R file -U /path/to/dump-file syntax. For further information, please refer to the tool's man page.
ibaddr	Can be used to show the LID and GID addresses of the specified port or the local port by default. This utility can be used as simple address resolver. For further information, please refer to the tool's man page.

Utility	Description
ibcacheedit	Allows users to edit an ibnetdiscover cache created through the --cache option in ibnetdiscover(8). For further information, please refer to the tool's man page.
ibccconfig	Supports the configuration of congestion control settings on switches and HCAs. For further information, please refer to the tool's man page.
ibccquery	Supports the querying of settings and other information related to congestion control. For further information, please refer to the tool's man page.
ibcongest	Provides static congestion analysis. It calculates routing for a given topology (topo-mode) or uses extracted lst/fdb files (lst-mode). Additionally, it analyzes congestion for a traffic schedule provided in a "schedule-file" or uses an automatically generated schedule of all-to-all-shift. To display a help message which details the tool's options, please run "/opt/ibutils2/bin/ibcongest -h". For further information, please refer to the tool's man page.
ibdev2netdev	Enables association between IB devices and ports and the associated net device. Additionally it reports the state of the net device link. For further information, please refer to the tool's man page.
ibdiagnet (of ibutils 2)	Scans the fabric using directed route packets and extracts all the available information regarding its connectivity and devices. An ibdiagnet run performs the following stages: • Fabric discovery • Duplicated GUIDs detection • Links in INIT state and unresponsive links detection • Counters fetch • Error counters check • Routing checks • Link width and speed checks • Alias GUIDs check • Subnet Manager check • Partition keys check • Nodes information

Utility	Description
	Note: This version of ibdiagnet is included in the ibutils2 package, and it is run by default after installing Mellanox OFED. To use this ibdiagnet version, run: ibdiagnet. For further information, either: 1. Run ibdiagnet -H Or 2. Refer to docs.nvidia.com/networking/display/ibdiagnetUserManualv10
ibdiagpath	Traces a path between two end-points and provides information regarding the nodes and ports traversed along the path. It utilizes device specific health queries for the different devices along the path. The way ibdiagpath operates depends on the addressing mode used in the command line. If directed route addressing is used (--dr_path flag), the local node is the source node and the route to the destination port is known apriori (for example: ibdiagpath --dr_path 0,1). On the other hand, if LID-route addressing is employed, --src_lid and --dest_lid, then the source and destination ports of a route are specified by their LIDs. In this case, the actual path from the local port to the source port, and from the source port to the destination port, is defined by means of Subnet Management Linear Forwarding Table queries of the switch nodes along that path. Therefore, the path cannot be predicted as it may change. Example: ibdiagpath --src_lid 1 --dest_lid 28 For further information, please refer to the tool's -help flag.
ibdump	Dump InfiniBand traffic that flows to and from Mellanox Technologies ConnectX® family adapters InfiniBand ports. Note the following: • ibdump is not supported for Virtual functions (SR-IOV). • Infiniband traffic sniffing is supported on all HCAs. • Ethernet and RoCE sniffing is supported only on Connect-X3 and Connect-X3 Pro cards. The dump file can be loaded by the Wireshark tool for graphical traffic analysis. The following describes a workflow for local HCA (adapter) sniffing: 1. Run ibdump with the desired options 2. Run the application that you wish its traffic to be analyzed 3. Stop ibdump (CTRL-C) or wait for the data buffer to fill (in --mem-mode) 4. Open Wireshark and load the generated file

Utility	Description
	To download Wireshark for a Linux or Windows environment go to www.wireshark.org. **Note:** Although ibdump is a Linux application, the generated .pcap file may be analyzed on either operating system. [mlx4] In order for ibdump to function with RoCE, Flow Steering must be enabled. To do so: 1. Add the following to /etc/modprobe.d/mlnx.conf file: options mlx4_core log_num_mgm_entry_size=-1 2. Restart the drivers. **Note:** If one of the HCA's port is configured as InfiniBand, ibdump requires IPoIB DMFS to be enabled. For further information, please refer to Flow Steering Configuration section. For further information, please refer to the tool's man page.
iblinkinfo	Reports link info for each port in an InfiniBand fabric, node by node. Option- ally, iblinkinfo can do partial scans and limit its output to parts of a fabric. For further information, please refer to the tool's man page.
ibnetdiscover	Performs InfiniBand subnet discovery and outputs a human readable topology file. GUIDs, node types, and port numbers are displayed as well as port LIDs and node descriptions. All nodes (and links) are displayed (full topology). This utility can also be used to list the current connected nodes. The output is printed to the standard output unless a topology file is specified. For further information, please refer to the tool's man page.
ibnetsplit	Automatically groups hosts and creates scripts that can be run in order to split the network into sub-networks containing one group of hosts. For further information, please refer to the tool's man page.
ibnodes	Uses the current InfiniBand subnet topology or an already saved topology file and extracts the InfiniBand nodes (CAs and switches). For further information, please refer to the tool's man page.
ibping	Uses vendor mads to validate connectivity between InfiniBand nodes. On exit, (IP) ping like output is show. ibping is run as client/server. The default is to run as client. Note also that a default ping server is implemented within the kernel. For further information, please refer to the tool's man page.
ibports	Enables querying the logical (link) and physical port states of an InfiniBand port. It also allows adjusting the link speed that is enabled on any InfiniBand port.

Utility	Description
tate	If the queried port is a switch port, then ibportstate can be used to: • disable, enable or reset the port • validate the port's link width and speed against the peer port In case of multiple channel adapters (CAs) or multiple ports without a CA/ port being specified, a port is chosen by the utility according to the following criteria: • The first ACTIVE port that is found. • If not found, the first port that is UP (physical link state is LinkUp). For further information, please refer to the tool's man page.
ibqueryerrors	The default behavior is to report the port error counters which exceed a threshold for each port in the fabric. The default threshold is zero (0). Error fields can also be suppressed entirely. In addition to reporting errors on every port, ibqueryerrors can report the port transmit and receive data as well as report full link information to the remote port if available. For further information, please refer to the tool's man page.
ibroute	Uses SMPs to display the forwarding tables—unicast (LinearForwarding- Table or LFT) or multicast (MulticastForwardingTable or MFT)—for the specified switch LID and the optional lid (mlid) range. The default range is all valid entries in the range 1 to FDBTop. For further information, please refer to the tool's man page.
ibstat	ibstat is a binary which displays basic information obtained from the local IB driver. Output includes LID, SMLID, port state, link width active, and port physical state. For further information, please refer to the tool's man page.
ibstatus	Displays basic information obtained from the local InfiniBand driver. Output includes LID, SMLID, port state, port physical state, port width and port rate. For further information, please refer to the tool's man page.
ibswitches	Traces the InfiniBand subnet topology or uses an already saved topology file to extract the InfiniBand switches. For further information, please refer to the tool's man page.

Utility	Description
ibsysstat	Uses vendor mads to validate connectivity between InfiniBand nodes and obtain other information about the InfiniBand node. ibsysstat is run as client/ server. The default is to run as client. For further information, please refer to the tool's man page.
ibtopodiff	Compares a topology file and a discovered listing ofsubnet.lst/ibdiagnet.lst and reports mismatches. Two different algorithms provided: • Using the -e option is more suitable for MANY mismatches it applies less heuristics and provide details about the match • Providing the -s, -p and -g starts a detailed heuristics that should be used when only small number of changes are expected For further information, please refer to the tool's man page.
ibtracert	Uses SMPs to trace the path from a source GID/LID to a destination GID/ LID. Each hop along the path is displayed until the destination is reached or a hop does not respond. By using the -m option, multicast path tracing can be performed between source and destination nodes. For further information, please refer to the tool's man page.
ibv_asyncwatch	Display asynchronous events forwarded to userspace for an InfiniBand device. For further information, please refer to the tool's man page.
ibv_devices	Lists InfiniBand devices available for use from userspace, including node GUIDs. For further information, please refer to the tool's man page.
ibv_devinfo	Queries InfiniBand devices and prints about them information that is available for use from userspace. For further information, please refer to the tool's man page.
mstflint	Queries and burns a binary firmware-image file on non-volatile (Flash) memories of NVIDIA InfiniBand and Ethernet network adapters. The tool requires root privileges for Flash access. To run mstflint, you must know the device location on the PCI bus.

Utility	Description
	Note: If you purchased a standard NVIDIA network adapter card, please download the firmware image from nvidia.com/en-us/networking/ Support Support Firmware Download. If you purchased a non-standard card from a vendor other than NVIDIA, please contact your vendor. For further information, please refer to the tool's man page.
perfquery	Queries InfiniBand ports' performance and error counters. Optionally, it displays aggregated counters for all ports of a node. It can also reset counters after reading them or simply reset them. For further information, please refer to the tool's man page.
saquery	Issues the selected SA query. Node records are queried by default. For further information, please refer to the tool's man page.
sminfo	Issues and dumps the output of an sminfo query in human readable format. The target SM is the one listed in the local port info or the SM specified by the optional SM LID or by the SM direct routed path. **Note:** Using sminfo for any purpose other than a simple query might result in a malfunction of the target SM. For further information, please refer to the tool's man page.
smparquery	Sends SMP query for adaptive routing and private LFT features. For further information, please refer to the tool's man page.
smpdump	A general purpose SMP utility which gets SM attributes from a specified SMA. The result is dumped in hex by default. For further information, please refer to the tool's man page.
smpquery	Provides a basic subset of standard SMP queries to query Subnet management attributes such as node info, node description, switch info, and port info. For further information, please refer to the tool's man page.

Link Level Retransmission (LLR) in FDR Links

With the introduction of FDR 56 Gbps technology, Mellanox enabled a proprietary technology called LLR (Link Level Retransmission) to improve the reliability of FDR links.

This proprietary LLR technology adds additional CRC checking to the data stream and retransmits portions of packets with CRC errors at the local link level. Customers should be aware of the following facts associated with LLR technology:

- Traditional methods of checking the link health can be masked because the LLR technology automatically fixes errors. The traditional IB symbol error counter will show no errors when LLR is active.
- Latency of the fabric can be impacted slightly due to LLR retransmissions. Traditional IB performance utilities can be used to monitor any latency impact.
- Bandwidth of links can be reduced if cable performance degrades and LLR retransmissions become too numerous. Traditional IB bandwidth performance utilities can be used to monitor any bandwidth impact.

Due to these factors, an LLR retransmission rate counter has been added to the ibdiagnet utility that can give end users an indication of the link health.

➢ *To monitor LLR retransmission rate:*

1. Run ibdiagnet, no special flags required.
2. If the LLR retransmission rate limit is exceeded it will print to the screen.
3. The default limit is set to 500 and requires further investigation if exceeded.
4. The LLR retransmission rate is reflected in the results file /var/tmp/ibdiagnet2/ibdiagnet2.pm.

The default value of 500 retransmissions/sec has been determined by Mellanox based on the extensive simulations and testing. Links exhibiting a lower LLR retransmission rate should not raise special concern.

Performance Utilities

The performance utilities described in this chapter are intended to be used as a performance micro-benchmark.

Utility	Description
ib_atomic_bw	Calculates the BW of RDMA Atomic transactions between a pair of machines. One acts as a server and the other as a client. The client RDMA sends atomic operation to the server and calculate the BW by sampling the CPU each time it receive a successful completion. The test supports features such as Bidirectional, in which they both RDMA atomic to each other at the same time, change of MTU size, tx size, number of iteration, message size and more. Using the "-a" flag provides results for all message sizes. For further information, please refer to the tool's man page.
ib_atomic_lat	Calculates the latency of RDMA Atomic transaction of message_size between a pair of machines. One acts as a server and the other as a client. The client sends RDMA atomic operation and sample the CPU clock when it receives a successful completion, in order to calculate latency. For further information, please refer to the tool's man page.
ib_read_bw	Calculates the BW of RDMA read between a pair of machines. One acts as a server and the other as a client. The client RDMA reads the server memory and calculate the BW by sampling the CPU each time it receive a successful completion. The test supports features such as Bidirectional, in which they both RDMA read from each other memory's at the same time, change of MTU size, tx size, number of iteration, message size and more. Read is available only in RC connection mode (as specified in IB spec). For further information, please refer to the tool's man page.
ib_read_lat	Calculates the latency of RDMA read operation of message_size between a pair of machines. One acts as a server and the other as a client. They perform a ping pong benchmark on which one side RDMA reads the memory of the other side only after the other side have read his memory. Each of the sides samples the CPU clock each time they read the other side memory , in order to calculate latency. Read is available only in RC connection mode (as specified in IB spec). For further information, please refer to the tool's man page.
ib_send_bw	Calculates the BW of SEND between a pair of machines. One acts as a server and the other as a client. The server receive packets from the client and they both calculate the throughput of the operation. The test supports features such as Bidirectional, on which they both send and receive at the same time, change of MTU size, tx size, number of iteration, message size and more. Using the "-a" provides results for all message sizes. For further information, please refer to the tool's man page.

Utility	Description
ib_send_lat	Calculates the latency of sending a packet in message_size between a pair of machines. One acts as a server and the other as a client. They perform a ping pong benchmark on which you send packet only if you receive one. Each of the sides samples the CPU each time they receive a packet in order to calculate the latency. Using the "-a" provides results for all message sizes. For further information, please refer to the tool's man page.
ib_write_bw	Calculates the BW of RDMA write between a pair of machines. One acts as a server and the other as a client. The client RDMA writes to the server memory and calculates the BW by sampling the CPU each time it receives a successful completion. The test supports features such as Bidirectional, in which they both RDMA write to each other at the same time, change of MTU size, tx size, number of iteration, message size and more. Using the "-a" flag provides results for all message sizes. For further information, please refer to the tool's man page.
ib_write_lat	Calculates the latency of RDMA write operation of message_size between a pair of machines. One acts as a server and the other as a client. They perform a ping pong benchmark on which one side RDMA writes to the other side memory only after the other side wrote on his memory. Each of the sides samples the CPU clock each time they write to the other side memory, in order to calculate latency. For further information, please refer to the tool's man page.
raw_ethernet_bw	Calculates the BW of SEND between a pair of machines. One acts as a server and the other as a client. The server receive packets from the client and they both calculate the throughput of the operation. The test supports features such as Bidirectional, on which they both send and receive at the same time, change of MTU size, tx size, number of iteration, message size and more. Using the "-a" provides results for all message sizes. For further information, please refer to the tool's man page.
raw_ethernet_lat	Calculates the latency of sending a packet in message_size between a pair of machines. One acts as a server and the other as a client. They perform a ping pong benchmark on which you send packet only if you receive one. Each of the sides samples the CPU each time they receive a packet in order to calculate the latency. Using the "-a" provides results for all message sizes. For further information, please refer to the tool's man page.

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.

The chapter contains the following sections:

Common Abbreviations and Related Documents

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)

Abbreviation/ Acronym	Description
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
IB	InfiniBand
iSER	iSCSI RDMA Protocol
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
IPoIB	IP over InfiniBand
PFC	Priority Flow Control
PR	Path Record
RoCE	RDMA over Converged Ethernet
SL	Service Level
SRP	SCSI RDMA Protocol
MPI	Message Passing Interface
QoS	Quality of Service
ULP	Upper Layer Protocol
VL	Virtual Lane
vHBA	Virtual SCSI Host Bus Adapter
uDAPL	User Direct Access Programming Library

Glossary

The following is a list of concepts and terms related to InfiniBand in general and to Subnet Managers in particular. It is included here for ease of reference, but the main reference remains the *InfiniBand Architecture Specification*.

Term	Description
Channel Adapter (CA), Host Channel Adapter (HCA)	An IB device that terminates an IB link and executes transport functions. This may be an HCA (Host CA) or a TCA (Target CA)
HCA Card	A network adapter card based on an InfiniBand channel adapter device
IB Devices	An integrated circuit implementing InfiniBand compliant communication
IB Cluster/Fabric/ Subnet	A set of IB devices connected by IB cables
In-Band	A term assigned to administration activities traversing the IB connectivity only
Local Identifier (ID)	An address assigned to a port (data sink or source point) by the Subnet Manager, unique within the subnet, used for directing packets within the subnet
Local Device/Node/ System	The IB Host Channel Adapter (HCA) Card installed on the machine running IBDIAG tools
Local Port	The IB port of the HCA through which IBDIAG tools connect to the IB fabric
Master Subnet Manager	The Subnet Manager that is authoritative, that has the reference configuration information for the subnet
Multicast Forwarding Tables	A table that exists in every switch providing the list of ports to forward received multicast packet. The table is organized by MLID
Network Interface Card (NIC)	A network adapter card that plugs into the PCI Express slot and provides one or more ports to an Ethernet network
Standby Subnet Manager	A Subnet Manager that is currently quiescent, and not in the role of a Master Subnet Manager, by the agency of the master SM

Term	Description
Subnet Administrator (SA)	An application (normally part of the Subnet Manager) that implements the interface for querying and manipulating subnet management data
Subnet Manager (SM)	One of several entities involved in the configuration and control of the IB fabric
Unicast Linear Forwarding Tables (LFT)	A table that exists in every switch providing the port through which packets should be sent to each LID
Virtual Protocol Interconnect (VPI)	A Mellanox Technologies technology that allows Mellanox channel adapter devices (ConnectX®) to simultaneously connect to an InfiniBand subnet and a 10GigE subnet (each subnet connects to one of the adapter ports)

Related Documentation

Document Name	Description
InfiniBand Architecture Specification, Vol. 1, Release 1.2.1	The InfiniBand Architecture Specification that is provided by IBTA
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
MFT User Manual and Release Notes	Mellanox Firmware Tools (MFT) User Manual and Release Notes documents
WinOF User Manual	Mellanox WinOF User Manual describes the installation, configuration, and operation of Mellanox Windows driver
VMA User Manual	Mellanox VMA User Manual describes the installation, configuration, and operation of Mellanox VMA driver

Documentation History

- Release Notes History
- User Manual Revision History

undefined