NVIDIA MLNX_OFED Documentation v23.10-0.5.5.0

Table of contents

Release Notes 9

General Support 13

Changes and New Features 29

Bug Fixes in This Version 33

Known Issues 36

User Manual 84

Introduction 84

Installation 96

Features Overview and Configuration 97

Programming 98

InfiniBand Fabric Utilities 101

Troubleshooting 112

Common Abbreviations and Related Documents 113

Documentation History 117

Release Notes History 117

User Manual Revision History 117

Legal Notices and 3rd Party Licenses 122

List of Figures

Figure 0. Image2019 1 10 10 55 56 Version 1 Modificationdate 1700558030970 Api V2

Figure 1. Procedure Heading Icon Version 1 Modificationdate 1700558035223 Api V2

Figure 2. Procedure Heading Icon Version 1 Modificationdate 1700558035223 Api V2

Figure 3. Worddave635fed9c99097774044df72a47e9130 Version 1 Modificationdate 1700558043550 Api V2

Figure 4. Image2019 2 12 10 26 53 Version 1 Modificationdate 1700558050843 Api V2

Figure 5. Procedure Heading Icon Version 1 Modificationdate 1700558058610 Api V2

Figure 6. Procedure Heading Icon Version 1 Modificationdate 1700558060693 Api V2

Figure 7. Procedure Heading Icon Version 1 Modificationdate 1700558060693 Api V2

Figure 8. Procedure Heading Icon Version 1 Modificationdate 1700558067833 Api V2

Figure 9. Image2022 7 26 10 29 13 Version 1 Modificationdate 1700558078533 Api V2

Figure 10. Image2022 7 28 14 58 53 Version 1 Modificationdate 1700558078123 Api V2

Figure 11. Image2022 7 28 15 0 34 Version 1 Modificationdate 1700558076633 Api V2

Figure 12. Image2022 7 28 14 59 57 Version 1 Modificationdate 1700558077223 Api V2

Figure 13. Image2022 7 28 15 1 11 Version 1 Modificationdate 1700558076203 Api V2

Figure 14. Image2022 7 28 15 1 42 Version 1 Modificationdate 1700558075790 Api V2

Figure 15. Image2022 7 28 15 2 14 Version 1 Modificationdate 1700558075360 Api V2

Figure 16. Image2022 7 28 15 4 11 Version 1 Modificationdate 1700558074960 Api V2

Figure 17. Image2022 7 28 15 6 40 Version 1 Modificationdate 1700558074120 Api V2

Figure 18. Procedure Heading Icon Version 1 Modificationdate 1700558120137 Api V2

Figure 19. Procedure Heading Icon Version 1 Modificationdate 1700558120137 Api V2

Figure 20. Procedure Heading Icon Version 1 Modificationdate 1700558120137 Api V2

Figure 21. Procedure Heading Icon Version 1 Modificationdate 1700558120137 Api V2

Figure 22. Procedure Heading Icon Version 1 Modificationdate 1700558120137 Api V2

Figure 23. Procedure Heading Icon Version 1 Modificationdate 1700558140867 Api V2

Figure 24. Procedure Heading Icon Version 1 Modificationdate 1700558140867 Api V2

Figure 25. Procedure Heading Icon Version 1 Modificationdate 1700558140867 Api V2

Figure 26. Procedure Heading Icon Version 1 Modificationdate 1700558140867 Api V2

Figure 27. Worddavb2ee67a7eb9aae5c536610e39a37dcc5 Version 1 Modificationdate 1700558149957 Api V2

Figure 28. Worddav6931c32564b3b0c166f4a26788219144 Version 1 Modificationdate 1700558148430 Api V2

Figure 29. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 30. Image2019 3 8 12 50 6 Version 1 Modificationdate 1700558147407 Api V2

Figure 31. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 32. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 33. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 34. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 35. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 36. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 37. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 38. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 39. Procedure Heading Icon Version 1 Modificationdate 1700558148107 Api V2

Figure 40. Sriov Live Migration Stage 2 Version 1 Modificationdate 1700558154100 Api V2

Figure 41. Image2020 7 26 14 25 6 Version 1 Modificationdate 1700558156647 Api V2

Figure 42. Image2020 7 26 14 26 25 Version 1 Modificationdate 1700558157113 Api V2

Figure 43. Image2020 7 26 14 27 26 Version 1 Modificationdate 1700558157527 Api V2

Figure 44. Image2020 7 26 14 28 39 Version 1 Modificationdate 1700558157997 Api V2

Figure 45. Image2020 7 26 14 30 13 Version 1 Modificationdate 1700558158687 Api V2

Figure 46. Image2020 7 26 14 33 10 Version 1 Modificationdate 1700558159303 Api V2

Figure 47. Image2020 7 26 14 34 7 Version 1 Modificationdate 1700558159930 Api V2

Figure 48. Image2020 7 26 14 34 18 Version 1 Modificationdate 1700558160323 Api V2

Figure 49. Image2020 7 26 14 38 28 Version 1 Modificationdate 1700558160810 Api V2

Figure 50. Image2020 7 26 14 39 14 Version 1 Modificationdate 1700558161683 Api V2

Figure 51. Image2020 7 26 14 39 27 Version 1 Modificationdate 1700558162377 Api V2

Figure 52. Procedure Heading Icon Version 1 Modificationdate 1700558167087 Api V2

Figure 53. Worddav336f9b6791fd85e08c8e6897697cd75b Version 1 Modificationdate 1700558168697 Api V2

Figure 54. Procedure Heading Icon Version 1 Modificationdate 1700558170943 Api V2

Figure 55. Procedure Heading Icon Version 1 Modificationdate 1700558170943 Api V2

Figure 56. Procedure Heading Icon Version 1 Modificationdate 1700558170943 Api V2

Figure 57. Procedure Heading Icon Version 1 Modificationdate 1700558180877 Api V2

For a PDF version of this manual, click here.

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 400Gb/s InfiniBand and RoCE (based on the RDMA over Converged Ethernet standard) over 10/25/40/50/100/200/400Gb/s are supported with OFED to enable OEMs and System Integrators to meet the needs of 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 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 changes made to the Release Notes, refer to Release Notes History.

Release Notes

Note

This is a long-term support (LTS) release. LTS is the practice of maintaining a software product for an extended period of time (up to three years) to help increase product stability. LTS releases include bug fixes and security patches.

Release Notes Update History

Version	Date	Description
23.10-0.5.5.0	November 7, 2023	Initial release of this document version. This release introduces Changes and New Features and Bug Fixes.

Warning

As of MLNX_OFED version v5.1-0.6.6.0, the following are no longer supported.

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

To utilize the above devices/libraries, refer to version 4.9 long-term support (LTS).

Release Notes contain the following sections:

- General Support
- Changes and New Features
- Bug Fixes in This Version
- Known Issues

Supported NIC Speeds

The Linux Driver operates across all NVIDIA network adapter solutions supporting the following uplinks to servers:

Uplink/Adapter Card	Driver Name	Uplink Speed
BlueField-2	mlx5	• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2]
BlueField		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-7		• InfiniBand: EDR, HDR100, HDR, NDR200, NDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE [3], 400GbE
ConnectX-6 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2]
ConnectX-6 Dx		• Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]

Uplink/Adapter Card	Driver Name	Uplink Speed
ConnectX-6		• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-5/ConnectX-5 Ex		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-4 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE
ConnectX-4		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE[1], 100GbE

1. 56GbE is an NVIDIA proprietary link speed and can be achieved while connecting an NVIDIA adapter card to NVIDIA SX10XX switch series or when connecting an NVIDIA adapter card to another NVIDIA adapter card.

2. Speed that supports both NRZ and PAM4 modes in Force mode and Auto-Negotiation mode.

3. Speed that supports PAM4 mode only.

Package Contents

Package	Revision	Licenses
clusterkit	1.11.442-1.2310055	BSD
dpcp	1.1.43-1.2310055	BSD-3-Clause
hcoll	4.8.3223-1.2310055	Proprietary
ibarr	0.1.3-1.2310055	(GPL-2.0 WITH Linux-syscall-note) OR BSD-2-Clause
ibdump	6.0.0-1.2310055	BSD2+GPL2

Package	Revision	Licenses
ibsim	0.12-1.2310055	GPLv2 or BSD
ibutils2	2.1.1-0.1.MLNX20231105.g79770a56.2310055	Mellanox Confidential and Proprietary
iser	23.10-OFED.23.10.0.5.3.1	GPLv2
isert	23.10-OFED.23.10.0.5.3.1	GPLv2
kernel-mft	4.26.0-93	Dual BSD/GPL
knem	1.1.4.90mlnx3-OFED.23.10.0.2.1.1	BSD and GPLv2
libvma	9.8.40-1	GPLv2 or BSD
libxlio	3.20.7-1	GPLv2 or BSD
mlnx-en	23.10-0.5.5.0.g97aff11	GPLv2
mlnx-ethtool	6.4-1.2310055	GPL
mlnx-iproute2	6.4.0-1.2310055	GPL
mlnx-nfsrdma	23.10-OFED.23.10.0.5.3.1	GPLv2
mlnx-nvme	23.10-OFED.23.10.0.5.3.1	GPLv2
mlnx-ofa_kernel	23.10-OFED.23.10.0.5.5.1	GPLv2
mlnx-tools	23.10-0.2310055	GPLv2 or BSD
mlx-steering-dump	1.0.0-0.2310055	GPLv2
mpitests	3.2.21-8418f75.2310055	BSD
mstflint	4.16.1-2.2310055	GPL/BSD
multiperf	3.0-3.0.2310055	BSD 3-Clause, GPL v2 or later
ofed-docs	23.10-OFED.23.10.0.5.5	GPL/BSD
ofed-scripts	23.10-OFED.23.10.0.5.5	GPL/BSD
openmpi	4.1.7a1-1.2310055	BSD
opensm	5.17.0.MLNX20231105.d437ae0a-0.1.2310055	GPLv2 or BSD

Package	Revision	Licenses
openvswitch	2.17.8-1.2310055	ASL 2.0 and LGPLv2+ and SISSL
perftest	23.10.0-0.29.g0705c22.2310055	BSD 3-Clause, GPL v2 or later
rdma-core	2307mlnx47-1.2310055	GPLv2 or BSD
rshim	2.0.12-0.g65537f7	GPLv2
sharp	3.5.0.MLNX20231105.d6571e71-1.2310055	Proprietary
sockperf	3.10-0.git5ebd327da983.2310055	BSD
srp	23.10-OFED.23.10.0.5.3.1	GPLv2
ucx	1.16.0-1.2310055	BSD
xpmem	2.7.3-1.2310055	GPLv2 and LGPLv2.1
xpmem-lib	2.7-0.2310055	LGPLv2.1

General Support

Supported Operating Systems

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
Alma 8.5	x86_64	4.18.0-348.12.2.EL8_5.X86_64	Community	✗	✗	✗	✗	✗	✗
Anolis OS 8.4	AArch64	4.18.0-348.2.1.AN8_4.AARCH64	Community	✗	✗	✗	✗	✗	✗
	x86_64	4.18.0-305.AN8.X86_64	Community	✗	✗	✗	✗	✗	✗

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
			y						
Anolis OS 8.6	AArch64	5.10.134+	Primary	✗	✗	✗	✗	✗	✗
	x86_64	5.10.134+	Primary	✗	✗	✗	✗	✗	✗
BCLINUX 21.10SP2	AArch64	4.19.90-2107.6.0.0098.oe1.bclinux.aarch64	Primary	✗	✗	✗	✓	✗	✗
	x86_64	4.19.90-2107.6.0.0100.oe1.bclinux.x86_64	Primary	✗	✗	✓	✓	✗	✗
CentOS Stream v8	AArch64	4.18.0-517.EL8.AARCH64	Community	✗	✗	✗	✗	✗	✗
	x86_64	4.18.0-517.EL8.X86_64	Community	✗	✗	✗	✗	✗	✗
CentOS Stream v9	AArch64	5.14.0-378.EL9.AARCH64	Community	✗	✗	✗	✗	✗	✗
	x86_64	5.14.0-378.el9.x86_64	Community	✗	✗	✗	✗	✗	✗
CTYUNOS2.0	AArch64	4.19.90-2102.2.0.0062.ctl2.aarch64	Primary	✗	✗	✗	✗	✗	✗

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
	x86_64	4.19.90-2102.2.0.0062.ctl2.x86_64	Primary	✗	✗	✗	✗	✗	✗
CTYUNOS23.01	AArch64	5.10.0-136.12.0.86.ctl3.aarch64	Primary	✗	✗	✗	✗	✗	✗
	x86_64	5.10.0-136.12.0.86.ctl3.x86_64	Primary	✗	✗	✗	✗	✗	✗
Debian9.13	AArch64	4.9.0-13-arm64	Primary	✗	✗	✗	✗	✗	✗
	x86_64	4.9.0-13-amd64	Primary	✗	✗	✗	✗	✗	✗
Debian10.8	AArch64	4.19.0-14-arm64	Primary	✗	✗	✗	✗	✗	✗
Debian10.9	x86_64	4.19.0-14-amd64	Primary	✗	✗	✗	✗	✗	✗
	x86_64	4.19.0-16-amd64	Primary	✓	✗	✗	✓	✗	✗
Debian11.3	AArch64	5.10.0-13-arm64	Primary	✓	✗	✗	✓	✗	✗
	x86_64	5.10.0-13-amd64	Primary	✓	✗	✗	✓	✗	✗
Debian12	AArch64	6.1.0-10-arm64	Primary	✓	✗	✗	✓	✗	✗

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
	x86_64	6.1.0-10-amd64	Primary	✔	✘	✘	✔	✘	✘
EulerOS2.0sp9	AArch64	4.19.90-vhulk2006.2.0.h171.eulerosv2r9.aarch64	Community	✘	✘	✘	✘	✘	✘
	x86_64	4.18.0-147.5.1.0.h269.eulerosv2r9.x86_64	Community	✘	✘	✘	✘	✘	✘
EulerOS2.0sp10	AArch64	4.19.90-vhulk2110.1.0.h860.eulerosv2r10.aarch64	Community	✘	✘	✘	✘	✘	✘
	x86_64	4.18.0-147.5.2.4.h694.eulerosv2r10.x86_64	Community	✘	✘	✘	✘	✘	✘
EulerOS2.0sp11	AArch64	5.10.0-60.18.0.50.h323.eulerosv2r11.aarch64	Primary	✘	✘	✘	✘	✘	✘
	x86_64	5.10.0-60.18.0.50.h323.eulerosv2r11.x86_64	Primary	✘	✘	✘	✘	✘	✘
EulerOS2.0sp12	AArch64	5.10.0-136.12.0.86.h1032.eulerosv2r12.aarch64	Primary	✘	✘	✘	✘	✘	✘

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
	x86_64	5.10.0-136.12.0.86.h1032.eulerosv2r12.x86_64	Primary	✗	✗	✗	✗	✗	✗
KYLIN10 SP2	AArch64	4.19.90-24.4.v2101.ky10.aarch64	Primary	✗	✗	✗	✗	✗	✗
	x86_64	4.19.90-24.4.v2101.ky10.x86_64	Primary	✓	✗	✗	✗	✗	✗
KYLIN10 SP3	AArch64	4.19.90-52.15.v2207.ky10.aarch64	Primary	✓	✗	✗	✗	✗	✗
	x86_64	4.19.90-52.15.v2207.ky10.x86_64	Primary	✓	✗	✗	✗	✗	✗
Mariner 2.0	x86_64	5.15.118.1-1.cm2.x86_64	Community	✗	✗	✗	✗	✗	✗
Oracle Linux 7.9	x86_64	5.4.17-2011.6.2.el7uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
Oracle Linux 8.4	x86_64	5.4.17-2102.201.3.el8uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
Oracle Linux 8.6	x86_64	5.4.17-2136.307.3.1.el8uek.x86_64	Primary	✗	✗	✗	✗	✗	✗

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
Oracle Linux 8.7	x86_64	5.15.0-3.60.5.1.el8uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
Oracle Linux 8.8	x86_64	5.15.0-101.103.2.1.el8uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
Oracle Linux 9.0	x86_64	5.15.0-0.30.19.el9uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
Oracle Linux 9.1	x86_64	5.15.0-3.60.5.1.el9uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
Oracle Linux 9.2	x86_64	5.15.0-101.103.2.1.el9uek.x86_64	Primary	✗	✗	✗	✗	✗	✗
OpenSUSE 15.3	AArch64	-	Community	✗	✗	✗	✗	✗	✗
	x86_64	5.3.18-150300.59.43-DEFAULT	Community	✗	✗	✗	✗	✗	✗
OPENEULER20.03 SP1	AArch64	4.19.90-2012.4.0.0053.OE1.AARCH64	Community	✗	✗	✗	✗	✗	✗
	x86_64	4.19.90-2110.8.0.0119.OE1.X86_64	Community	✗	✗	✗	✗	✗	✗
OPENEULER20.03	AArch6	4.19.90-2112.8.0.0131.oe	Primary	✗	✗	✗	✗	✗	✗

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
SP3	4	1.aarch64							
	x86_64	4.19.90-2112.8.0.0131.oe1.x86_64	Primary	✗	✗	✗	✗	✗	✗
OPENEULER22.03	AArch64	5.10.0-60.18.0.50.oe2203.aarch64	Primary	✗	✗	✗	✗	✗	✗
	x86_64	5.10.0-60.18.0.50.oe2203.x86_64	Primary	✗	✗	✗	✗	✗	✗
Photon OS 3.0	x86_64	4.19.225-3.ph3	Community	✗	✗	✗	✗	✗	✗
RHEL/CentOS7.2	x86_64	3.10.0-327.el7.x86_64	Primary	✗	✗	✗	✓	✗	12.2
RHEL/CentOS7.4	x86_64	3.10.0-693.el7.x86_64	Primary	✓	✓	✗	✓	✗	12.2
RHEL/CentOS7.6	x86_64	3.10.0-957.el7.x86_64	Primary	✓	✓	✓	✓	✗	12.2
RHEL/CentOS7.6alternate	aarch64	4.14.0-115.el7a.aarch64	Community	✓	✓	✗	✗	✗	✗
RHEL/CentOS7.7	x86_64	3.10.0-1062.el7.x86_64	Primary	✓	✓	✗	✗	✗	12.2
RHEL/CentOS7.8	x86_64	3.10.0-1127.el7.x86_64	Primary	✓	✓	✓	✓	✗	12.2
RHEL/Ce	x86	3.10.0-	Prima	✓	✓	✓	✓	✗	12.2

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
ntOS7.9	_64	1160.el7.x86_64	ry						
RHEL/CentOS8.0	AArch64	4.18.0-80.el8.aarch64	Primary	✔	✔	✘	✘	✔	12.2
	x86_64	4.18.0-80.el8.x86_64	Primary	✔	✔	✘	✘	✔	12.2
RHEL/CentOS8.1	AArch64	4.18.0-147.el8.aarch64	Primary	✔	✔	✘	✘	✔	12.2
	x86_64	4.18.0-147.el8.x86_64	Primary	✔	✔	✘	✘	✔	12.2
RHEL/CentOS8.2	AArch64	4.18.0-193.el8.aarch64	Primary	✔	✔	✘	✔	✔	12.2
	x86_64	4.18.0-193.el8.x86_64	Primary	✔	✔	✔	✔	✔	12.2
RHEL/CentOS8.3	AArch64	4.18.0-240.el8.aarch64	Primary	✔	✔	✘	✘	✔	12.2
	x86_64	4.18.0-240.el8.x86_64	Primary	✔	✔	✘	✘	✔	12.2
RHEL/CentOS8.4	AArch64	4.18.0-305.el8.aarch64	Primary	✔	✔	✘	✔	✔	12.2
	x86_64	4.18.0-305.el8.x86_64	Primary	✔	✔	✔	✔	✔	12.2

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
RHEL/CentOS/ Rocky8.5	AArch64	4.18.0-348.el8.aarch64	Primary	✔	✔	✘	✔	✔	12.2
	x86_64	4.18.0-348.el8.x86_64	Primary	✔	✔	✔	✔	✔	12.2
RHEL/Rocky8.6	AArch64	AArch644.18.0-372.41.1.el8_6.aarch64	Primary	✔	✔	✘	✔	✔	12.2
	x86_64	4.18.0-372.41.1.el8_6.x86_64	Primary	✔	✔	✔	✔	✔	12.2
RHEL/Rocky8.7	AArch64	4.18.0-425.14.1.el8_7.aarch64	Primary	✔	✘	✘	✔	✔	12.2
	x86_64	4.18.0-425.14.1.el8_7.x86_64	Primary	✔	✘	✔	✔	✔	12.2
RHEL/Rocky8.8	AArch64	4.18.0-477.10.1.el8_8.aarch64	Primary	✔	✘	✘	✔	✔	12.2
	x86_64	4.18.0-477.10.1.el8_8.x86_64	Primary	✔	✘	✔	✔	✔	12.2
RHEL/Rocky9.0	AArch64	5.14.0-70.46.1.el9_0.aarch64	Primary	✔	✘	✘	✔	✔	12.2
	x86_64	5.14.0-70.46.1.el9_0.x86_64	Primary	✔	✘	✔	✔	✔	12.2

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
RHEL/Rocky9.1	AArch64	5.14.0-162.19.1.el9_1.aarch64	Primary	✔	✘	✘	✔	✔	12.2
	x86_64	5.14.0-162.19.1.el9_1.x86_64	Primary	✔	✘	✔	✔	✔	12.2
RHEL/Rocky9.2	AArch64	5.14.0-284.11.1.el9_2.aarch64	Primary	✔	✘	✘	✔	✔	12.2
	x86_64	5.14.0-284.11.1.el9_2.x86_64	Primary	✔	✘	✔	✔	✔	12.2
SLES12.1 SP2	x86_64	4.4.21-69-default	Community	✘	✘	✘	✘	✘	✘
SLES12SP3	x86_64	4.4.73-5-default	Community	✘	✘	✘	✘	✘	✘
SLES12SP4	AArch64	4.12.14-94.41-default	Community	✔	✘	✘	✔	✘	✘
	x86_64	4.12.14-94.41-default	Community	✔	✘	✔	✔	✘	✘
SLES12SP5	AArch64	4.12.14-120-default	Primary	✔	✘	✘	✔	✘	✘
	x86_64	4.12.14-120-default	Primary	✔	✘	✔	✔	✘	✘

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
SLES15SP2	AArch64	5.3.18-22-default	Primary	✔	✔	✘	✔	✘	✘
	x86_64	5.3.18-22-default	Primary	✔	✔	✔	✔	✘	✘
SLES15SP3	AArch64	5.3.18-57-default	Primary	✔	✘	✘	✔	✘	✘
	x86_64	5.3.18-57-default	Primary	✔	✘	✔	✔	✘	✘
SLES15SP4	AArch64	5.14.21-150400.22-default	Primary	✔	✘	✘	✔	✘	✘
	x86_64	5.14.21-150400.22-default	Primary	✔	✘	✔	✔	✘	✘
SLES15SP5	AArch64	5.14.21-150500.53-default	Primary	✔	✘	✘	✔	✘	✘
	x86_64	5.14.21-150500.53-default	Primary	✔	✘	✔	✔	✘	✘
Ubuntu 16.04	x86_64	4.4.0-21-generic	Community	✘	✘	✘	✘	✘	✘
Ubuntu18.04	AArch64	4.15.0-20-generic	Primary	✔	✔	✘	✔	✔	11.6

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
	x86_64	4.15.0-20-generic	Primary	✔	✔	✔	✔	✔	11.6
Ubuntu20.04	AArch64	5.4.0-26-generic	Primary	✔	✔	✘	✔	✔	12.2
	x86_64	5.4.0-26-generic	Primary	✔	✔	✔	✔	✔	12.2
Ubuntu22.04	AArch64	5.15.0-25-generic	Primary	✔	✔	✔	✔	✔	12.2
	x86_64	5.15.0-25-generic	Primary	✔	✔	✔	✔	✔	12.2
Ubuntu23.04	x86_64	6.2.0-20-generic	Primary	✔	✘	✔	✔	✘	✘
Ubuntu 23.10	x86_64	6.5.0-5-generic	Primary	✔	✘	✔	✔	✘	✘
UOS20.1020	AArch64	4.19.90-2109.1.0.0108.up2.uel20.aarch64	Primary	✘	✘	✘	✘	✘	✘
	x86_64	4.19.90-2109.1.0.0108.up2.uel20.x86_64	Primary	✘	✘	✘	✘	✘	✘
UOS20.1040	AArch64	4.19.0-arm64-server	Primary	✘	✘	✘	✘	✘	✘
	x86_64	4.19.0-server-amd64	Primary	✘	✘	✘	✘	✘	✘

Operating System	Architecture	Default Kernel Version (Primary)/ Tested with Kernel Version (Community)	OS Support Model	ASAP^2 OVS-Kernel SR-IOV	ASAP^2 OVS-DPDK SR-IOV	NFS over RDMA	NVMe	GPUDirect Storage (GDS)	UCX-CUDA Version
Citrix XenServer Host7.1	x86_64	4.4.0+2	Primary	✗	✗	✗	✗	✗	✗
Citrix XenServer Host8.2	x86_64	4.19.0+1	Primary	✗	✗	✗	✗	✗	✗
Kernel 6.5	AArch64	6.5	Primary	✓	✗	✗	✓	✗	✗
	x86_64	6.5	Primary	✓	✗	✓	✓	✗	✗

⚠ Warning

- 32 bit platforms are no longer supported in MLNX_OFED
- For RPM-based distributions, to install OFED on a different kernel, create a new ISO image using mlnx_add_kernel_support.sh script (see the MLNX_OFED User Manual for instructions)
- Upgrading MLNX_OFED on a cluster requires upgrading all of its nodes to the newest version as well
- If using MLNX_OFED 4.9 LTS with MLNX_OFED 5.x with upstream verbs, MLNX_OFED 4.9 must be installed with --upstream-libs flag so the verbs libraries match.

- A combination of 4.9 LTS default verbs and MOFED 5.x upstream verbs is not supported.
- All operating systems listed above are fully supported in Paravirtualized and SR-IOV environments with Linux KVM Hypervisor

Upgrade/Downgrade Matrix

This section reflects which versions were tested and verified for upgrade and downgrade.

Target Version	**Versions Verified for Upgrade/Downgrade**	Release Type	Release Date
23.10-0.5.5.0 GA	23.07-0.5.1.2	GA (Update #1)	September 2023
	5.8-3.0.7.0	GA-LTS (Update #3)	June 2023

MLNX_OFED Version Interoperability

This section reflects which versions were tested and verified for multi-version environments.

Target Version	Verified OFED Version Interoperability	Release Type	Release Date
23.10-0.5.5.0 GA	23.07-0.5.1.2	GA (Update #1)	September 2023
	5.8-3.0.7.0	GA-LTS (Update #3)	June 2023

Supported NIC Firmware Versions

⚠ **Warning**

As of version 5.1, ConnectX-3, ConnectX-3 Pro or Connect-IB adapter cards are no longer supported. To work with a version that supports these adapter cards, please refer to version 4.9 long-term support (LTS).

This current version is tested with the following NVIDIA adapter card firmware versions:

Adapter Card	Bundled Firmware Version	Additional Supported Firmware Version
BlueField®-2	24.38.1002	24.37.1300
ConnectX-7	28.39.1002	28.38.1002
ConnectX-6 Lx	26.39.1002	26.38.1002
ConnectX-6 Dx	22.39.1002	22.38.1002
ConnectX-6	20.39.1002	20.38.1002
ConnectX-5/ConnectX-5 Ex	16.35.3006	N/A
BlueField	18.33.1048	N/A
ConnectX-4	12.28.2006	N/A
ConnectX-4 Lx	14.32.1010	N/A

For the official firmware versions, please see https://www.nvidia.com/en-us/networking/ Support Support Firmware Download.

Supported Non-Linux Virtual Machines

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

NIC	Windows Virtual Machine Type	Minimal WinOF Version	Protocol
ConnectX-4	Windows 2012 R2 DC	MLNX_WinOF2 2.50	IB, IPoIB, ETH
ConnectX-4 Lx	Windows 2016 DC	MLNX_WinOF2 2.50	IB, IPoIB, ETH

NIC	Windows Virtual Machine Type	Minimal WinOF Version	Protocol
ConnectX-5 family	All Windows server editions	MLNX_WinOF2 2.50	IPoIB, ETH
ConnectX-6 family		MLNX_WinOF2 2.50	IPoIB, ETH

Support in ASAP2—Accelerated Switch and Packet Processing®

ASAP^2 Requirements	• iproute >= 4.12 (for tc support) • Upstream Open vSwitch >= 2.8 for CentOS 7.2 NVIDIA openvswitch
ASAP^2-Supported Adapter Cards	• ConnectX-5 • ConnectX-6 Dx • ConnectX-6 Lx • ConnectX-7

Lustre Versions Compatible with MLNX_OFED

- Lustre 2.15.2
- Lustre 2.12.9

Hardware and Software Requirements

- 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	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
SLES 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
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

Unsupported Functionalities/Features/NICs

The following are the unsupported functionalities/features/NICs in the current version:

- ConnectX-2 adapter card
- ConnectX-3 adapter card
- ConnectX-3 Pro adapter card
- Connect-IB adapter card
- Soft-RoCE
- RDMA experimental verbs library (mlnx_lib)
- CIFS (Common Internet File System) module installation

- Relational Database Service (RDS)
- mthca InfiniBand driver
- Ethernet IPoIB (eIPoIB)
- InfiniBand Connected transport service

Changes and New Features

Customer Affecting Changes

Customer Affecting Change	Description
Debugfs Directory Path Change	The debugfs directory of each interface can now be found under: /sys/kernel/debug/mlx5//, and not directly under the root of the debugfs filesystem (/sys/debug/kernel).
Deprecation of OFED Public Power PC Installation	Starting from this release, MLNX_OFED releases for Power PC are no longer available for download from the public Download Center web page. Instead, you can find it on the following page: https://network.nvidia.com/support/firmware/ibm-systemp/.
Pre-notification: Deprecation of Older Operating Systems	Starting from next release, MLNX_OFED releases will no longer support operating systems with kernels below v4.18. This includes the following systems: • Rhel7.x • Debian9.13 • Sles12.x • Xenserver7.1

New Features

The following are the new features and changes that were added in this version. The supported adapter cards are specified as follows:

Supported Cards	Description
All HCAs	Supported in the following adapter cards **unless specifically stated otherwise:** ConnectX-4/ConnectX-4 Lx/ConnectX-5/ConnectX-6/ConnectX-6 Dx/ConnectX-6 Lx/ConnectX-7/BlueField-2
ConnectX-6 Dx and above	Supported in the following adapter cards **unless specifically stated otherwise:** ConnectX-6 Dx/ConnectX-6 Lx/ConnectX-7/BlueField-2
ConnectX-6 and above	Supported in the following adapter cards **unless specifically stated otherwise:** ConnectX-6/ConnectX-6 Dx/ConnectX-6 Lx/ConnectX-7/BlueField-2

Supported Cards	Description
ConnectX-5 and above	Supported in the following adapter cards **unless specifically stated otherwise:** ConnectX-5/ConnectX-6/ConnectX-6 Dx/ConnectX-6 Lx/ConnectX-7/BlueField-2
ConnectX-4 and above	Supported in the following adapter cards **unless specifically stated otherwise:** ConnectX-4/ConnectX-4 Lx/ConnectX-5/ConnectX-6/ConnectX-6 Dx/ConnectX-6 Lx/ConnectX-7/BlueField-2

Feature/Change	Description
23.10-0.5.5.0	
ASAP^2 Features	
Rate Limiter Extension for LAG Mode	[ConnectX-5 and above] During rate configuration, this feature considers the aggregated port speed when LAG is configured, so the functions (PF,VF,SF) can utilize the maximum aggregated link speed for transmission.
Core Features	
Live Migration Support with IPsec Full Offload	[ConnectX-7 and above] Added an option to perform live migration on a machine with IPsec enabled. This allows having a VF on VM that has a valid IPsec state and migrate it without having to remove the IPsec. **Note:** This feature is currently not supported when working in MPV mode (Multi port VHCA/Dual port). Such support will be added in future releases.
Exposing the Max SF Configuration via Devlink Resource	[All HCAs] This feature allows user applications to view the maximum number of SFs (Scalable Functions) configured on the system via general Linux API, Devlink resource.
NetDev Features	
NIC Temperature Exposure	[ConnectX-5 and above] Added support for NIC temperature exposure by implementing the hwmon kernel API.
Counters Indicating Packet	[All HCAs] Added two counters through the Devlink health reporter - generated_pkt_steering_fail and handled_pkt_steering_fail. These

Feature/Change	Description
Drops due to Catastrophic Steering Errors	counters indicate packet drops due to catastrophic steering errors.
RDMA Features	
Using the RDMA Tool to Allow Controlled QKEYs Use by Non-privileged Users	[All HCAs] To allow using controlled QKEYs by non-privileged users, the following command can now be used in the iproute2 RDMA tool, instead of the "enforce_qkey_check" module parameter, which will be deprecated: ----- $rdma system set privileged-qkey on
IPsec Configuration Support When All IPsec Operations are Offloaded to the Hardware	[ConnectX-7 and above] Added support for configuring IPsec even when all IPsec operations are completely offloaded to the hardware. This feature not only provides a significant performance improvement, but also enables the use of IPsec over RoCE packets, which are outside the network stack and cannot be used without full hardware offload. As a result, users can now leverage the benefits of IPsec protocol with RoCE V2, even when using SR-IOV VFs.
Security Features	
Configuration from VM/VF for Full Offload Transport Mode	[ConnectX-6 Lx and above] This feature introduces two new boolean attributes of a port function: ipsec_crypto and ipsec_packet (also known as IPsec Full Offload). The goal is to provide a level of granularity for controlling VF IPsec offload capabilities similar to what is currently offered in the software mode. This allows users to decide if they want both types of IPsec offload enabled for a VF, just one of them, or none (which is the default).
General	
General	Bug fixes

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

API Changes in MLNX_OFED

MLNX_OFED Verbs API Migration

As of MLNX_OFED v5.0 release (Q1 of the year 2020), MLNX_OFED Verbs API have migrated from the legacy version of user space verbs libraries (libibervs, libmlx5, etc.) to the Upstream version rdma-core.

For the list of MLNX_OFED verbs APIs that have been migrated, refer to Migration to RDMA-Core document.

Bug Fixes in This Version

Below are the bugs fixed in this version. For a list of fixes previous version, see Bug Fixes History.

Internal Reference Number	Description
3613899	**Description:** Fixed an issue where the srp_daemon service was enabled by default. Starting from this release, the srp_daemon service is disabled by default, and can be enabled after MLNX_OFED installation using the systemctl enable srp_daemon command.
	Keywords: srp_daemon
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3549785	**Description:** Fixed an issue where the NVMe and mlx5_core drivers failed during BFB installation. As a result, Anolis OS could not be installed on the SSD and the mlxfwreset command did not work during Anolis BFB installation.
	Keywords: NVMe, mlx5_core, Anolis OS
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3602955	**Description:** Fixed an issue that occurred when a VF was set to get allmulti traffic. The issue caused the steering rules to send the multicast traffic received by the NIC back to the uplink.

Internal Reference Number	Description
	Keywords: VF, allmulti traffic
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3553766	**Description:** Fixed an issue where the enable_remote_dev_reset Devlink parameter was not supported on kernel versions below v5.10.
	Keywords: Devlink parameter
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3546694	**Description:** Fixed an issue where MAC address configuration for PFs could fail if SR-IOV was enabled at the same time.
	Keywords: PF, MAC address, SR-IOV
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3538018	**Description:** Fixed an issue where firmware sync reset (with the 'mlxfwreset -d <device> -l 3 r --sync 1' command) could fail on a system configured for hotplug on the PCIe slot on which the mlx5 card was mounted.
	Keywords: Firmware sync reset, mlx5 card
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3587834	**Description:** Fixed an issue where the enable_remote_dev_reset Devlink parameter was not supported on kernel versions below v5.10.
	Keywords: Devlink parameter
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3576351	**Description:** Resolved a warning that was triggered when starting the openibd service, which pertained to an unidentified 'ExecRestart'

Internal Reference Number	Description
	value within the 'Service' section.
	Keywords: openibd, warning
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3565980	**Description:** Fixed an issue where deb-based (Ubuntu and Debian) versions of MLNX_OFED did not install the rshim package, and the software needed to access the BlueField guest system from the host on a BlueField-type installation by default. Note that this was only needed (and enabled) on the host side. This issue did not occur on RPM-based systems (RHEL, SLES, Oracle, etc.).
	Keywords: Ubuntu, Debian, rshim package
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3557482	**Description:** Fixed an issue where the 'mlnx_tune -l' list of supported operating systems did not include several operating systems that were actually supported, such as RHEL8.6 and Ubuntu 22.04.
	Keywords: mlnx_tune -l
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3549684	**Description:** Fixed a signature-related issue that occurred when installing DOCA on SLES15SP4 using the repository.
	Keywords: DOCA, SLES15SP4
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0
3380263	**Description:** Fixed an issue where users who attempted to use OFED with Device ID NVD0000000033, had to install the firmware manually.
	Keywords: Device ID NVD0000000033
	Discovered in Release: 23.07-0.5.0.0

Internal Reference Number	Description
	Fixed in Release: 23.10-0.5.5.0
3228788	**Description:** Fixed an issue where running rx-tls-offload over Korg6.0 as its TLS module did not work properly.
	Keywords: NetDev, TLS
	Discovered in Release: 23.07-0.5.0.0
	Fixed in Release: 23.10-0.5.5.0

Known Issues

The following is a list of general limitations and known issues of the current version of the release.

Internal Ref. Number	Issue
3682658	**Description:** While using the RDMA-CM user application and the AF_IB parameter, the kernel uses only the first byte of the private data to set the CMA version. In such scenario, any user data written to this byte will be overwritten.
	Keywords: RDMA-CM user application, AF_IB, private data
	Workaround: Do not use AF_IB for application's private data.
	Discovered in Release: 23.10-0.5.5.0
3640082	**Description:** A potential null pointer dereference might occur due to a missing update in the PCI subsystem code when creating the maximum number of VFs. All kernel versions lacking the following fix are impacted: "PCI: Avoid enabling PCI atomics on VFs."
	Keywords: Maximal VF number
	Workaround: N/A
	Discovered in Release: 23.10-0.5.5.0

Internal Ref. Number	Issue
3653417	**Description:** When offloading IPsec policy rules while in legacy mode there are two options: 1. Software steering - The software stack will handle the task, and no device offload will take place. 2. Changing the steering mode to firmware steering will return unsupported.
	Keywords: IPsec, legacy mode
	Workaround: Perform a devlink reload after changing the steering mode.
	Discovered in Release: 23.10-0.5.5.0
3612274	**Description:** Currently, either IPsec offload or TC offload for a specific interface is allowed. The offloading TC rule to an interface will fail if an IPSec rule is already offloaded on it, and vice-versa.
	Keywords: IPsec offload, TC offload
	Workaround: N/A
	Discovered in Release: 23.10-0.5.5.0
3596126	**Description:** OVS mirroring of both egress and ingress together with modified TTL is not supported by Connectx-5 cards, and may cause packets checksum issues and errors in the dmesg command.
	Keywords: OVS mirroring, Connectx-5
	Workaround: N/A
	Discovered in Release: 23.10-0.5.5.0
3538463	**Description:** A Kernel ABI problem in Sles15SP4 may lead to issues during driver start. This impacts kernels starting from version 5.14.21-150400.24.11.1 up to version 5.14.21-150400.24.63.1 (July 2022 to May 2023), inclusive. For more information, see https://www.suse.com/support/kb/doc/?id=000021137.
	Keywords: Kernel ABI, Sles15SP4, driver start

Internal Ref. Number	Issue
	Workaround: Upgrade to a kernel version newer than 5.14.21-150400.24.63.1 (May 2023).
	Discovered in Release: 23.10-0.5.5.0
3637252	**Description:** When running over REHL7.6 with excessive RDMA/RoCE workload, kernel warnings may be triggered.
	Keywords: REHL7.6, RDMA, RoCE
	Workaround: N/A
	Discovered in Release: 23.10-0.5.5.0

Internal Ref. Number	Issue
3046655	**Description:** A package manager upgrade with zypper (on an SLES system) may prompt a question about vendor change from "Mellanox Technologies" to "OpenFabrics".
	Keywords: Installation, SLES
	Workaround: Either accept the prompted change, or add the /etc/zypp/vendors.d/mlnx_ofed file with the following content: [main] vendors = Mellanox,OpenFabrics
	Discovered in Release: 23.07-0.5.0.0
3392477	**Description:** The ConnectX-7 firmware embedded in this MLNX_OFED version cannot be burnt using the MLNX_OFED installer script.
	Keywords: ConnectX-7, MLNX_OFED installer script
	Workaround: Please download and install the dedicated firmware from the web https://network.nvidia.com/support/firmware/connectx7ib/
	Discovered in Release: 23.07-0.5.0.0
3532756	**Description:** The kernel may crash when restarting the driver while IP sec rules are configured.
	Keywords: IP sec

Internal Ref. Number	Issue
	Workaround: Flush the IP sec configuration before reloading the driver: ip xfrm state flush ip xfrm policy flush
	Discovered in Release: 23.07-0.5.0.0
3472979	**Description:** When a large number of virtual functions are present, the output of the "ip link show" command may be truncated.
	Keywords: virtual functions, ip link show
	Workaround: N/A
	Discovered in Release: 23.07-0.5.0.0
3413938	**Description:** When using the mlnx-sf script, creating and deleting an SF with the same ID number in a stressful manner may cause the setup to hang due to a race between the create and delete commands.
	Keywords: Hang; mlnx-sf
	Workaround: N/A
	Discovered in Release: 23.07-0.5.0.0
3461572	**Description:** Configuring Multiport Eswitch LAG mode can be performed only via devlink from this release onwards. The compat sysfs should not be used to configure mpesw LAG.
	Keywords: Multiport Eswitch, compat sysfs, mpesw LAG
	Workaround: N/A
	Discovered in Release: 23.07-0.5.0.0
3464337	**Description:** Simultaneously adding or removing TC rules while operating on kernel version 6.3 could potentially result in stability issues.
	Keywords: ASAP, rules, TC
	Workaround: Make sure the following fix is part of the kernel: https://lore.kernel.org/netdev/20230504181616.2834983-3-vladbu@nvidia.com/T/

Internal Ref. Number	Issue
	Discovered in Release: 23.07-0.5.0.0
3469484	**Description:** Mirror and connection tracking (CT) offload actions are not supported simultaneously if the kernel version does not support hardware miss to TC actions. Thus, when performing a CT offload test, the actual number of offloaded connections may be lower than expected.
	Keywords: ASAP, CT offload
	Workaround: Make sure to have the following offending commit in the tree: net/sched: act_ct: offload UDP NEW connections Make sure to to have https://www.spinics.net/lists/stable-commits/msg303536.html in the kernel tree to fix this issue.
	Discovered in Release: 23.07-0.5.0.0
3473331	**Description:** When performing a CT offload test, the actual number of offloaded connections may be lower than expected.
	Keywords: ASAP, CT offload
	Workaround: N/A
	Discovered in Release: 23.07-0.5.0.0
3499413	**Description:** Due to the following kernel issue, under heavy load, some connections may not be offloaded, leading to performance issues: "net/sched: act_ct: offload UDP NEW connections."
	Keywords: ASAP, CT offload
	Workaround: N/A
	Discovered in Release: 23.07-0.5.0.0

Internal Ref. Number	Issue
3360710	**Description:** Configuring PFC in parallel to buffer size and prio2buffer commands may lead to misalignment between firmware and software in regards to receiving buffer ownership.

Internal Ref. Number	Issue
	Keywords: NetDev, PFC, Buffer Size, prio2buffer
	Workaround: First, configure PFC on all ports, and then perform other needed QoS (i.e., buffer_size or prio2buffer) configurations accordingly.
	Discovered in Release: 23.04-0.5.3.3
3413879	**Description:** OpenSM may not be started automatically if chkconfig was not installed before OpenSM is installed. Note, however, that chkconfig will fail to install if the directory (rather than symbolic link to directory) /etc/init.d already exists (e.g., from a previous installation of MLNX_OFED).
	Keywords: Installation, OpenSM, chkconfig
	Workaround: Install chkconfig before installing MLNX_OFED. If installing it fails, make sure /etc/init.d does not exist at the time of installing it.
	Discovered in Release: 23.04-0.5.3.3
3424596	**Description:** On SLES 15.4, installing MLNX_OFED using a package repository (with zypper) may trigger an error message about missing dependency for 'librte_eal.so.20.0()(64bit)' . This is because the inbox package libdpdk-20_0 is being uninstalled as it is incompatible with the MLNX_OFED rdma-core packages.
	Keywords: Installation, SLES 15.4
	Workaround: Uninstall the relevant packages: 'zypper uninstall libdpdk-20_0' before installing MLNX_OFED. This will also remove the inbox openvswitch package.
	Discovered in Release: 23.04-0.5.3.3
3433416	**Description:** On systems that were installed with MLNX_OFED 5.9 or older and include a CUDA package (ucx-cuda / hcoll-cuda), an upgrade to MLNX_OFED 23.04 using the package manager ("yum") method will fail. This is because MLNX_OFED up to 5.9 is built with CUDA 11. MLNX_OFED 23.04 is built with CUDA 12 and those CUDA versions are incompatible.
	Keywords: Installation, CUDA, yum

Internal Ref. Number	Issue
	Workaround: Remove CUDA packages included with OFED (ucx-cuda, hcoll-cuda) before upgrading. This will allow to upgrade MLNX_OFED regardless of CUDA version installed. To install them later, CUDA 12 must be installed on the system.
	Discovered in Release: 23.04-0.5.3.3
3420831	**Description:** mlx-steering-dump is not supported on systems in which Python3 is not the default.
	Keywords: mlx-steering-dump, Python3
	Workaround: N/A
	Discovered in Release: 23.04-0.5.3.3
3351989	**Description:** If the underlying persistent device name exceeds 15 characters in length, the operating system will not be able to perform renaming (i.e., the device name will remain "eth ").
	Keywords: Persistant Interface Names
	Workaround: Add the --copy-ifnames-udev flag to the OFED installation command. Note that this flag is only applicable if the persistent name provided by the kernel, without the 'np ' suffix, is 15 characters or fewer.
	Discovered in Release: 23.04-0.5.3.3

Internal Ref. Number	Issue
3324094	**Description:** When working in legacy rq (striding rq off), with large MTU > 3712, a 10-20% degradation in performance might be seen when running UDP stream with 64 bytes message size.
	Keywords: NetDev, MTU, UDP Stream
	Workaround: N/A
	Discovered in Release: 5.9-0.5.6.0
3313137	**Description:** Virtual Functions depend on Physical Functions for device access (e.g, firmware host PAGE management). In addition, VF

Internal Ref. Number	Issue
	may need to access safely the PF 'driver data' to use the command interface as in the VFIO usage to support live migration. While the PF is missing its driver, the VFs are completely unusable. As such, upon PF unload, the SR-IOV is disabled by the PF itself. This is the standard widely seen behavior in Linux drivers today.
	Keywords: Core, SR-IOV, VF, PF
	Workaround: N/A
	Discovered in Release: 5.9-0.5.6.0
3320947	**Description:** When the system is overloaded, there is a possibility that one hour will pass between the creation of DevLink port and it usage/assignment, due to some locking. This will trigger a trace starting with: "Type was not set for devlink port."
	Keywords: Core, DevLink, System Overload
	Workaround: N/A
	Discovered in Release: 5.9-0.5.6.0
3046222	**Description:** Installing OFED with Open vSwitch packages failed over Ubuntu22 OS with inbox Open vSwitch installed on it. Inbox Open vSwitch packages should be removed first.
	Keywords: Installation, Ubuntu22
	Workaround: Use --with-openvswitch flag along with the installation command.
	Discovered in Release: 5.9-0.5.6.0
3262725	**Description:** Devlink reload while deleting namespace may causes a deadlock on kernels older than Linux-6.0.
	Keywords: Devlink, Namespace
	Workaround: N/A
	Discovered in Release: 5.9-0.5.6.0
3253255	**Description:** RHEL 7 does not include built-in support for Python3. There are two potential ways to install it, and both install a package with a different name:

Internal Ref. Number	Issue
	1. EPEL for RHEL7: python36 2. Rhel extra repository Python3 support is needed for using Pyverbs and the Python support of Open vSwitch. MLNX_OFED assumes that on RHEL7.x, if using Python3, that python36 from EPEL is used (otherwise the optional Python3 support cannot be used).
	Keywords: RHEL7, Python3
	Workaround: To use Python3 support on RHEL7, install python36 from the RHEL7 EPEL repository.
	Discovered in Release: 5.9-0.5.6.0

Internal Ref. Number	Issue
3215514	**Description:** On EulerOS 2.0SP11, installation with the yum method may fail with an error that mlnx-iproute2 is missing a dependency on libdb-5.3.so()(64bit).
	Keywords: Installation, EulerOS 2.0SP11, yum
	Workaround: Install in advance the mlnx-iproute2 package with rpm and with the --nodeps option. For example: rpm -Uv --nodeps RPMS/mlnx-iproute2-5.19.0-1.58101.x86_64.rpm
	Discovered in Release: 5.8- 1.0.1.1
3191223	**Description:** In old kernels, /etc/init.d/openibd stop will fail because of an existing TC rule. Because mlx5_ib is already unloaded, mlx5_core and mlx5_ib will be in an inconsistent state.
	Keywords: ASAP^2, eSwitch, TC Rules
	Workaround: Set eSwitch mode to legacy before enabling SR-IOV or reload mlx5_core to change eSwitch mode to legacy.
	Discovered in Release: 5.8- 1.0.1.1
3199628	**Description:** ping -6 -i <interface name> is broken in v5.18.
	Keywords: NetDev, -i flag

Internal Ref. Number	Issue
	Workaround: In all operating systems that are running Kernel 5.18 and below, remove the -i flag.
	Discovered in Release: 5.8- 1.0.1.1
3002932	**Description:** Jumbo MTU must be set on all uplinks (i.e., uplinks of *_sf and *_sf_r) at all times.
	Keywords: NetDev, MTU, Uplink
	Workaround: Configure jumbo MTU (9216) on all uplink-related interfaces.
	Discovered in Release: 5.8- 1.0.1.1
3130859	**Description:** The yum install method might be broken on installer regenerated with --add-kernel-support-build-only.
	Keywords: Installation, yum
	Workaround: Delete the original mlnx-ofed-all-5.* package and recreate the repository with: createrepo RPMS/
	Discovered in Release: 5.8- 1.0.1.1
3149387	**Description:** The package neohost-backend (included in MLNX_OFED) has a strict dependency on Python 2.7 and on the existance of /usr/bin/python. This dependency is because of a pre-installation test (which is a rather non-standard method) for /usr/bin/python will fail the installation if without Python 2.7. As a result, default installation of this on newer systems that do not have a default of Python 2 has been disabled. If there is an explicit request for this installation using the command-line option --with-neohost-backend, this sanity check will be overriden and there will be an attempt to install it regardless. On newer systems, there is likely to not be /usr/bin/python even if Python 2 is installed; as such its installation will fail.
	Keywords: Installation, Python 2
	Workaround: If neohost-backend is needed on a newer system, install Python 2 in advance and create the symbolic link /usr/bin/python -> python2.

Internal Ref. Number	Issue
	Discovered in Release: 5.8- 1.0.1.1
3213777	**Description:** Oracle Enterprise Linux version 9.0 generates kernel module packages that have dependencies that are not provided by their own kernel RPM packages and thus are not installable.
	Keywords: Installation, Oracle Enterprise Linux v9.0
	Workaround: N/A
	Discovered in Release: 5.8- 1.0.1.1
3229904	**Description:** Restart driver failes to load OFED modules after installing OFED on SLES15sp4 with errata kernel 5.14.21-150400.24.21-default.
	Keywords: Installation
	Workaround: Install OFED with --add-kernel-support flag.
	Discovered in Release: 5.8- 1.0.1.1
3189424	**Description:** VLAN naming is limited to 16 characters (like all other interface names). For names longer than 16 charachters, the kernel generates its own interface name VLAN (VID).
	Keywords: Core, VLAN, Interface Name
	Workaround: Select a name which complies to the 16-characters limitation.
	Discovered in Release: 5.8- 1.0.1.1
3220855	**Description:** Creating external SFs on BF ARM when the host (x86) operating system does not support SFs may cause the host to crash.
	Keywords: Core, Scalable Functions
	Workaround: N/A
	Discovered in Release: 5.8- 1.0.1.1
3239291	**Description: I**n some topologies, like logical partitions, mlxfwreset is not supported.
	Keywords: Core, mlxfwreset

Internal Ref. Number	Issue
	Workaround: N/A
	Discovered in Release: 5.8- 1.0.1.1

Internal Ref. Number	Issue
3114823	**Description:** The first attempt to create a new iSER connection fails with the following messages in dmesg: iSCSI Login timeout on Network Portal <iSER_Target_IP_ADDR>:3260 isert: isert_get_login_rx: isert_conn 00000000e9239d52 interrupted before got login req After the error, the iSER Initiator connects to the Target successfully, but the memory allocated for the first connection is not freed correctly. As a result, the failed attempt also causes memory leakage. • kernel.org Kernel 5.18 • RHEL 9.0 • RHEL 8.6 • Ubuntu 22.04 • SLES 15 SP4 The error happens due to a bug in the scsi_transport_iscsi module, which is not a part of MLNX_OFED. As such, the issue cannot be fixed in MLNX_OFED. The bug is already fixed in kernel 5.19 by the commit f6eed15f3ea7 ("scsi: iscsi: Exclude zero from the endpoint ID range").
	Workaround: Update the kernel if the above errors are experienced. If the issue is still reproduced after the kernel update, ask your distro support to apply the bug fix from the upstream kernel.
	Keywords: iSER Initiator
	Discovered in Release: 5.7-1.0.2.0
3096911	**Description:** Installing chkconfig on Rhel9.0 with OFED using yum failed (chkconfig creates /etc/init.d sym link and OFED creates files in this directory, causing a conflict).
	Workaround: Installing chkconfig before OFED.

Internal Ref. Number	Issue
	Keywords: Installation
	Discovered in Release: 5.7-1.0.2.0
3100544	**Description:** On a RHEL9.x system, in some cases where inbox modules do not match for the drivers being build, rebuilding the drivers (--add-kernel-support) works, but fails to install the built package, with many errors such as: kernel(__rdma_block_iter_next) = 0x8e7528da is needed by mlnx-ofa_kernel-modules-5.6-OFED.5.6.2.0.9.1.kver.5.14.0_70.13.1.el9_0.aarch64.aarch64 This was caused by a bug in the scripts that creates the Requires and Provides headers that is confused by dependencies between different modules of the same external package.
	Workaround: dnf install kernel-modules- # in case it is not the newest.
	Keywords: Installation, RHEL9.x
	Discovered in Release: 5.7-1.0.2.0
3132158	**Description:** Building rdma-core package on Rocky 8.6 OS caused failure in OFED build.
	Workaround: N/A
	Keywords: Installation
	Discovered in Release: 5.7-1.0.2.0
3137440	**Description:** Python package is missing, need to install it manually.
	Workaround: Install Python before starting the build.
	Keywords: Installation, Python
	Discovered in Release: 5.7-1.0.2.0
3141506	**Description:** kernel-macros package does not support building with KMP enabled. KMP needs to be disabled.
	Workaround: Build and install MOFED with KMP disabled (without --kmp flag).
	Keywords: Installation
	Discovered in Release: 5.7-1.0.2.0

Internal Ref. Number	Issue
3141506	**Description:** kernel-macros package does not support building with KMP enabled. KMP needs to be disabled.
	Workaround: Build and install MOFED with KMP disabled (without --kmp flag).
	Keywords: Installation
	Discovered in Release: 5.7-1.0.2.0
3129627	**Description:** Kernel module packaging is not supported in CtyunOS.
	Workaround: N/A
	Keywords: Installation
	Discovered in Release: 5.7-1.0.2.0

2971708	**Description:** For OSs in which Devlink supports setting roce-enable/disable, both sysfs roce_enable show and sysfs roce_enable set are disabled, and the RoCE state must be managed exclusively via Devlink.
	The sysfs interface for roce-enable/disable will be removed entirely for these OSs in a future release.
	To determine if Devlink can be used to enable or disable RoCE, execute the following console command after starting OFED:
	`devlink dev param show | grep roce`
	Devlink supports roce enable/disable if the following line is reflected in the output:
	`name enable_roce type generic`
	For OSs which do not allow enabling/disabling RoCE via Devlink, the sysfs interface behaves as in the previous 2 releases:
	1. For OSs which have Devlink reload, but do not allow setting RoCE state via Devlink:
	sysfs roce_enable show works, as does sysfs roce_enable set, but Devlink reload must be performed after setting the RoCE state via sysfs in order to activate the desired roce state.
	2. For OSs which do not have Devlink reload, RoCE state is managed only by the sysfs interface.

Internal Ref. Number	Issue

	'show' displays the RoCE state and 'set' sets the state and activates it.
	To determine if Devlink dev reload is supported, execute the following console command (using the bash shell):
	`devlink dev help 2>&1 | grep reload`
	Reload is supported if the output is:
	`devlink dev reload DEV [ netns { PID | NAME | ID } ]`

	Workaround: N/A
	Keywords: Enabling/Disabling RoCE
	Discovered in Release: 5.7-1.0.2.0

Internal Ref. Number	Issue
2998194	**Description:** On some systems with many (e.g., 64) virtual functions (VFs) attached to a ConnectX interface, 'ip link' may give an error message: "Error: Buffer too small for object." This applies to both IP commands: the inbox iproute package in RHEL8.x and the mlnx-iproute2 package from MLNX_OFED. This is known to work well and not give an error in RHEL7.x kernel regardless of what user-space package is used (including user-space from RHEL8.x).
	Workaround: N/A
	Keywords: NetDev, RHEL, Virtual Functions
	Discovered in Release: 5.6-1.0.3.3
3045436	**Description:** Rebooting the host while the Arm is down may block the shutdown flow till the Arm is up.

Internal Ref. Number	Issue
	Workaround: Restart the driver on the host side before reboot.
	Keywords: Reboot, Arm
	Discovered in Release: 5.6-1.0.3.3
3040350	**Description:** 1. When offload is enabled, removing a physical port from ovs-dpdk bridge requires restarting OVS service. Not doing so will result in wrong configuration of datapath rules. 2. When offload is enabled, the physical port must be attached to a bridge.
	Workaround: 1. When removing a physical port from an ovs-dpdk bridge while offload is enabled, need to restart openvswitch after reattaching it. 2. Attach physical port to a bridge according to the desired topology.
	Keywords: OVS-DPDK, Bridge, Offload
	Discovered in Release: 5.6-1.0.3.3
2973726	**Description:** dec_ttl only work with ConnectX-6. It does not work with ConnectX-5.
	Workaround: N/A
	Keywords: OVS-DPDK, dec_ttl
	Discovered in Release: 5.6-1.0.3.3
2946873	**Description:** Moving to switchdev mode while deleting namespace may cause a deadlock.
	Workaround: Unload mlx5_ib module before moving to Switchdev mode.
	Keywords: ASAP^2, Switchdev, Namespace
	Discovered in Release: 5.6-1.0.3.3

Internal Ref. Number	Issue
2811957	**Description:** If a system is run from a network boot and is connected to the network storage through an NVIDIA ConnectX card, unloading the mlx5_core driver (such as running '/etc/init.d/openibd restart') will render the system unusable and should therefore be avoided.
	Workaround: N/A
	Keywords: Installation, mlx5_core
	Discovered in Release: 5.6-1.0.3.3
2979243	**Description:** The kernel in CentOS 7.6alt (for non-x86 architectures) is different than that of RHEL 7.6alt. Some of the MLNX_OFED kernel modules that were built for the RHEL7.6alt kernel will not load on a system with Centos7.6alt kernel. If you want to install MLNX_OFED on such a system, you should use ./mlnxofedinstall --add-kernelsupport to rebuild the kernel modules for the Centos kernel.
	Workaround: Use add-kernel-support.
	Keywords: Installation,CentOS
	Discovered in Release: 5.6-1.0.3.3
3011440	**Description:** In Debian 11.2, Ubuntu 21.10, and Ubuntu 22.04, attempting to install an "exact" type of metapackage (such as mlnx-ofed-all-exact or mlnx-ofed-basic-exact) may fail with an error regarding the version of mstflint.
	Workaround: Install also mstflint of the exact same version (e.g., apt install mlnx-ofed-all-exact mstflint=4.16.0-1.56xxxx).
	Keywords: Installation,Debian, Ubuntu, MST
	Discovered in Release: 5.6-1.0.3.3
3024520	**Description:** The option --copy-ifnames-udev copy some files under /etc (/etc/udev/rules.d/82-net-setup-link.rules and /etc/infiniband/vf-net-link-name.sh) that are never removed--not in the case this option is not given and not upon uninstallation. Those scripts are merely examples. They are files under /etc to be maintained by the user.
	Workaround: Remove the files, if needed.
	Keywords: Installation

Internal Ref. Number	Issue
	Discovered in Release: 5.6-1.0.3.3
3046601	**Description:** When rebuilding the kernel modules (--add-kernel-support) for some kernel versions (specifically mainline 4.14) do not unset LDFLAGS properly. Rebuilding xpmem in such a case may fail with the error such as "unrecognized option '-Wl,-z,relro'" in the xpmem build log.
	Workaround: Either disable building xpmem by adding --without-xpmem to the command line, or edit the kernel Makefile to make it unset LDFLAGS: `sed -i -e '/^export ARCH/iLDFLAGS :=' /lib/modules/$(uname -r)/Makefile` **Note:** The Makefile may be located elsewhere, such as the top-level directory of the kernel source directory.
	Keywords: Installation, SLES
	Discovered in Release: 5.6-1.0.3.3
3048411	**Description:** After installing OFED with rebuilt kernel modules, error messages indicating that the kernel module mlx5_ib failed to load (e.g. "mlx5_ib: Unknown symbol . . .") appear. These messages could be safely ignored because the module eventually loads.
	Workaround: Run the command 'dracut -f' to update the initramfs.
	Keywords: Installation
	Discovered in Release: 5.6-1.0.3.3
3048444	**Description:** OFED installation failed using yum for --add-kernel-support option (building packages without KMP enabled) if libfabric package is installed.
	Workaround: Remove libfabric package before OFED installation or use installation script.
	Keywords: Installation, RHEL 8.5
	Discovered in Release: 5.6-1.0.3.3
3015210	**Description:** OVS topology where the tunnel device is over a VF and the VF representor is connected to a bond is not supported.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: ASAP^2, Tunnel Over VF, LAG, Connection Tracking
	Discovered in Release: 5.6-1.0.3.3
3028300	**Description:** OVS metering is not support over kernel 5.17.
	Workaround: N/A
	Keywords: ASAP^2,OVS, Meter, Kernel 5.17
	Discovered in Release: 5.6-1.0.3.3
3044255	**Description:** Destroying mlxdevm group while SF is attached to it is not supported.
	Workaround: N/A
	Keywords: ASAP^2, mlxdevm, QoS, Group, Scalable Functions, ConnectX-6 Dx
	Discovered in Release: 5.6-1.0.3.3
2900346	**Description:** On Ubuntu OS, configuring different IP addresses with different subnets to both ports 0 and 1 is currently not supported. When trying to ping from port 0 on one BlueField-2 card to port 0 on the other BlueField-2 card, then both port 0 and port 1 on the receiving side send a reply to the ARP request (a.k.a, ARP flux).
	Workaround: N/A
	Keywords: BlueField-2, Ubuntu, ARP Flux
	Discovered in Release: 5.6-1.0.3.3
3046456	**Description:** Switching between SwitchDev mode and legacy mode quickly on BlueField-2 can prevent the driver from loading successfully and breaks its health recovery.
	Workaround: Pause 60 seconds between state-altering commands to guarantee the driver health recovery is completed successfully.
	Keywords: ASAP^2, Health Recovery
	Discovered in Release: 5.6-1.0.3.3

Internal Ref. Number	Issue
2934149	**Description:** Adding vDPA ports over ConnectX-5 devices in ovs-dpdk is not supported and will cause a crash.
	Workaround: N/A
	Keywords: OVS-DPDK, ConnectX-5
	Discovered in Release: 5.6-1.0.3.3
2934833	**Description:** Running I/O traffic and toggling both physical ports status (UP/DOWN) in a stressful manner on the receiving-end machine may cause traffic loss.
	Workaround: N/A
	Keywords: RDMA, Port Toggle
	Discovered in Release: 5.6-1.0.3.3
2901514	**Description:** Relaxed Ordering is not working properly on Virtual Functions.
	Workaround: N/A
	Keywords: Relaxed Ordering, VF
	Discovered in Release: 5.6-1.0.3.3

Internal Ref. Number	Issue
2870299	**Description:** Managing SFs is possible using the iproute2 with mlxdevm tool only.
	Workaround: N/A
	Keywords: Scalable Functions

Internal Ref. Number	Issue
	Discovered in Release: 5.5-1.0.3.2
2869722	**Description:** OFED packages were built with DKMS disabled since building OFED with DKMS failed due to a problem in the DKMS package on UOS. --dkms flag should not be used.
	Workaround: N/A
	Keywords: Installation, DKMS
	Discovered in Release: 5.5-1.0.3.2
2870367	**Description:** On UOS, IPoIB PKEY may require manual bring up after driver restart.
	Workaround: N/A
	Keywords: Installation, IPoIB, PKEY
	Discovered in Release: 5.5-1.0.3.2
2836032	**Description:** When using Software steering mlx5dv_dr API to create rules containing encapsulation actions in MLNX_OFED v5.5-1.x.x.x, upgrade firmware to the latest version. Otherwise, the maximum number of encapsulation actions that can be created will be limited to only 16K, and degradation for the rule insertion rate is expected compared to MLNX_OFED v5.4-.x.x.x.x.
	Workaround: N/A
	Keywords: Software Steering
	Discovered in Release: 5.5-1.0.3.2
2851639	**Description:** Enabling ARFS in legacy mode and then moving to switchdev mode is not supported and may cause unwanted behavior.
	Workaround: N/A
	Keywords: NetDev, ARFS
	Discovered in Release: 5.5-1.0.3.2
2851639	**Description:** nvme and iser are not enabled on UOS ARM, because of missing UOS kernel support.
	Workaround: N/A

Internal Ref. Number	Issue
	Keywords: nvme, iser, UOS ARM
	Discovered in Release: 5.5-1.0.3.2
2860855	**Description:** Building OFED on RHEL 8.4 with kmp disabled and then installing with yum fails due to some conflicting packages.
	Workaround: Remove libfabric and librpmem packages before OFED installation,or add --allowerasing option to the installation command.
	Keywords: Installation, RHEL 8.4, kmp, yum
	Discovered in Release: 5.5-1.0.3.2
2865983	**Description:** OFED packages were built with kmp disabled. Building with kmp enabled fails due to missing packages.
	Workaround: N/A
	Keywords: Installation, kmp
	Discovered in Release: 5.5-1.0.3.2

Internal Ref. Number	Issue
2658644	**Description:** Only match on lower 32 bit of ct_label is supported.
	Workaround: N/A
	Keywords: ASAP^2, Connection Tracking
	Discovered in Release: 5.4-1.0.3.0
2706345	**Description:** Number of RQ and TIR allocation in the driver depends on total number of MSI-X vectors allocated. Total number of TIRs supported by device is 16K range. Each representor needs number of CPUs worth TIRs, upto maximum of 128.
	Workaround: To use large number of VFs, set PF_NUM_PF_MSIX to a smaller value of around 32.
	Keywords: ASAP^2,VF, PF_NUM_PF_MSIX
	Discovered in Release: 5.4-1.0.3.0

Internal Ref. Number	Issue
2836997	**Description:** An automatic test that checks a flow meter rate fluctuation stays within a fixed threshold (e.g., 10%) may fail because meter precision is dependent on multiple factors (i.e., rate and burst values and shape of the traffic). To pick the best configuration parameters for a flow meter, perform a couple of test measurements using different values of burst size against expected traffic workload and average the results over an extended period of time (tens of minutes).
	Workaround: N/A
	Keywords: ASAP^2,Meter Threshold
	Discovered in Release: 5.4-1.0.3.0
2863456	**Description:** SA limit by packet count (hard and soft) are supported only on traffic originated from the ECPF. Trying to configure them on VF traffic will remove the SA when hard limit is hit, however traffic could still pass as plain text due to the tunnel offload that is used in such configuration.
	Workaround: N/A
	Keywords: ASAP^2, IPsec Full Offload
	Discovered in Release: 5.4-0.5.1.1
2657392	**Description:** OFED installation caused CIFS to break in RHEL 8.4 and above. A dummy module was added so that CIFS will be disabled after OFED installation in RHEL 8.4 and above.
	Workaround: N/A
	Keywords: Installation, RHEL, CIFS
	Discovered in Release: 5.4-0.5.1.1
2800993	**Description:** OpenMPI does not support running across different operating systems and/or CPU architectures.
	Workaround: N/A
	Keywords: OpenMPI

Internal Ref. Number	Issue
2399503	**Description:** O pen vSwitch is not supported on the latest operating systems containing only Python3 support.
	Workaround: N/A
	Keywords: Python, O pen vSwitch
2657392	**Description:** OFED installation caused CIFS to break in RHEL8.4. A dummy module was added so that CIFS will be disabled after OFED installation in RHEL8.4.
	Workaround: N/A
	Keywords: Installation, RHEL8.4, CIFS
	Discovered in Release: 5.4-0.5.1.1
2782406	**Description:** Running yum update will upgrade kylin-release to a higher version. The version of this package is used for kylin10sp2 detection so the script will detect kylin 10 instead of kylin10sp2 and use its repository by mistake.
	Workaround: Because there are no special cases for kylin10sp2, the repository that was detected with adding --add-kernel-support to the installation command can be used.
	Keywords: Upgrade, kylin
	Discovered in Release: 5.4-3.0.3.0
2755632	**Description:** On dual port cards with SR-IOV, when one port link is configured to InfiniBand and the other port link is configured to Ethernet, the Ethernet port will not be able to support VST and QinQ.
	Workaround: N/A
	Keywords: SR-IOV, VST, QinQ
	Discovered in Release: 5.4-3.0.3.0
2780436	**Description:** Non-default MTU (>1500) is not supported with IPsec crypto offload and may cause packet drops.
	Workaround: N/A
	Keywords: IPsec, Crypto Offload, MTU

Internal Ref. Number	Issue
	Discovered in Release: 5.4-3.0.3.0
2726021	**Description:** Building packages on openEuler with kmp enabled requires kernel-rpm-macros package installed. kernel-rpm-macros-30-13.oe1 does not support -p option and kernel-rpm-macros-30-18.oe1 should be installed instead. On kylin OS, the version of kernel-rpm-macros package does not support -p option needed to support kmp, so it will stay disabled.
	Workaround: N/A
	Keywords: Installation, openEuler
	Discovered in Release: 5.4-3.0.3.0

Internal Ref. Number	Issue
2750653	**Description:** Running fragmented traffic in RHEL 8.3 (4.18.0-240.el8.x86_64) may cause call trace in build_skb.
	Workaround: Update to RHEL 8.3 z-stream 4.18.0-240.22.1.el8_3.x86_64.
	Keywords: RHEL 8.3, Kernel Panic, Call Trace, fr
	Discovered in Release: 5.4-1.0.3.0
2629375	**Description:** Matching on CT label is only supported when matching on lower 32 bits. Full match on all 128 bits of CT label is not supported.
	Workaround: N/A
	Keywords: ASAP^2, Connection Tracking, Label
	Discovered in Release: 5.4-1.0.3.0
2707997	**Description:** Installation in the package manager mode under SLES 15.x may require user-intervention if the original libibverbs is installed.

Internal Ref. Number	Issue
	Workaround: zypper install --force-resolution mlnx-ofed-all
	Keywords: Installation, libibverbs
	Discovered in Release: 5.4-1.0.3.0
2708531	**Description:** Installation in the package manager mode under SLES 15.x may require user-intervention if the original libopenvswitch is installed.
	Workaround: zypper install --force-resolution mlnx-ofed-all
	Keywords: Installation
	Discovered in Release: 5.4-1.0.3.0
2703043	**Description:** Congested TCP lock for kTLS TX device offload traffic compromises the performance.
	Workaround: Disable TCP selective acknowledgement: echo 0 > /proc/sys/net/ipv4/tcp_sack
	Keywords: kTLS TX
	Discovered in Release: 5.4-1.0.3.0
2676405	**Description:** If the package interface-rename is active (on XenServer, for example), the interface renaming by the OFED will not be done to eliminate conflicts.
	Workaround: N/A
	Keywords: Interface Renaming
	Discovered in Release: 5.4-1.0.3.0
2687943	**Description:** Offload of rules which redirect from VF on one PF to VF on second PF is not supported on socket-direct devices.
	Workaround: N/A
	Keywords: ASAP^2, Socket-Direct
	Discovered in Release: 5.4-1.0.3.0
2678672	**Description:** When disabling switchdev mode, the qdisc in tunnel device cannot be destroyed and mlx5e_stats_flower() is still called by OVS resulting in NULL pointer panic and memory leak.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: SwitchDev, mlx5, Tunnel Traffic
	Discovered in Release: 5.4-1.0.3.0
2566548	**Description:** On PPC systems when EEH is enabled, running fw sync reset (either by mlxfwreset with flag --sync 1 or by devlink dev reload action fw_activate), the EEHmay catch the PCI reset and take ownership on the flow. When run few times in sequence, the EEH may also decide to disable the device.
	Workaround: Administrator may disable EEH before running firmware sync reset on the device.
	Keywords: PPC, EEH
	Discovered in Release: 5.4-1.0.3.0
2617950	**Description:** TX port timestamp feature is supported for kernel versions 3.15 and greater. On older kernel versions, the feature will not be supported and ptp_tx _* counters will not increment.
	Workaround: N/A
	Keywords: Ethtool
	Discovered in Release: 5.4-1.0.3.0
2390731	**Description:** Ethtool does not display Port Speed advertised/capability above 100Gb/s over and below kernels 5.0, even when supported.
	Workaround: N/A
	Keywords: Ethtool, Port Speed
	Discovered in Release: 5.4-1.0.3.0

Internal Ref. Number	Issue
2687198	**Description:** Activating VF/SF LAG when at least one VF/SF is still bound may lead to an internal error in the firmware.

Internal Ref. Number	Issue
	Workaround: Make sure all VFs/SFs are unbound prior to VF/SF LAG activation/deactivation.
	Keywords: VF, SF, Firmware, Binding
	Discovered in Release: 5.4-1.0.3.0

Internal Ref. Number	Issue
2585575	**Description:** After disabling sync reset by setting enable_remote_dev_reset to false, running firmware sync reset a few times may lead to general protection fault and system may get stuck.
	Workaround: N/A
	Keywords: Firmware Upgrade
	Discovered in Release: 5.3-1.0.0.1
2582565	**Description:** Conducting a firmware reset or unbinding the PF while in switchdev mode may cause a kernel crash.
	Workaround: N/A
	Keywords: SwitchDev, ASAP^2, Unbind, Firmware Reset
	Discovered in Release: 5.3-1.0.0.1
2587802	**Description:** PTP synchronization may be lost while using tx_port_ts private flag.
	Workaround: Toggle private flag: ethtool --set-priv-flags tx_port_ts off ethtool --set-priv-flags tx_port_ts on restart ptp4l application
	Keywords: PTP Synchronization
	Discovered in Release: 5.3-1.0.0.1
2574943	**Description:** When running kernel 5.8 and bellow or RHEL 8.2 and below, sampled packets do not support tunnel information.
	Workaround: N/A

Internal Ref. Number	Issue
	Keywords: ASAP^2, sFLOW
	Discovered in Release: 5.3-1.0.0.1
2568417	**Description:** Upon upgrade to version 5.3, the package manager tool will install the new packages and then remove the old packages, a depmod WARNING on "mlx5_fpga_tools" will appear. This warning can be safely ignored. mlx5_fpga_tools is a module that existed in version 5.2 and was removed in 5.3.
	Workaround: N/A
	Keywords: Upgrade; mlx5_fpga_tools
	Discovered in Release: 5.3-1.0.0.1
2506425	**Description:** When installing kmod packages on EulerOS 2.0SP9 or OpenEuler 20.03, the following error appears: "modprobe: FATAL: could not get modversions of ". This error can be safely ignored. It is caused by incorrectly adding directories to a list of modules processed by /usr/sbin/weak-modules.
	Workaround: N/A
	Keywords: Installation; modules; kmod
	Discovered in Release: 5.3-1.0.0.1
2492509	**Description:** When installing the driver on OpenEuler or on EulerOS 2.0SP9, rebuilding the drivers (--add-kernel-support) with the --kmp option (to create kmod packages) generates packages that are uninstallable because they have a dependency on "/sbin/depmod" that the system does not provide. This dependency is created by a buggy kmod package building tool included with the distribution.
	Workaround: N/A
	Keywords: add-kernel-support
	Discovered in Release: 5.3-1.0.0.1
2479327	**Description:** On SLES 12 SP5, if the kernel was upgraded to 4.12.14-122.46, it is not possible to rebuild kernel modules (--add-kernel-support) without upgrading gcc as well to at least 4.8.5-31.23.2.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: Upgrade; SLES 12; add-kernel-support
	Discovered in Release: 5.3-1.0.0.1
2584441	**Description:** On SLES 12 SP5, if the kernel was upgraded to 4.12.14-122.46, it is not possible to rebuild kernel modules (--add-kernel-support) without upgrading gcc as well to at least 4.8.5-31.23.2.
	Workaround: N/A
	Keywords: Upgrade; SLES 12; add-kernel-support
	Discovered in Release: 5.3-1.0.0.1
2460865	**Description:** When setting MTU to low values, such as 68 bytes, packets may fail on oversize.
	Workaround: N/A
	Keywords: MTU
	Discovered in Release: 5.3-1.0.0.1
2383318	**Description:** On kernels based on RedHat 7.2, the "tx_port_ts" feature, as set by ethtool —set-priv-flags, is disabled.
	Workaround: N/A
	Keywords: RedHat; tx_port_ts
	Discovered in Release: 5.3-1.0.0.1
2575647	**Description:** An OvS-DPDK crash might occur while doing live-migration for VMs that use virtio-interfaces that are accelerated using OvS-DPDK vDPA ports.
	Workaround: N/A
	Keywords: OvS-DPDK vDPA, Live-migration
	Discovered in Release: 5.3-1.0.0.1

Internal Ref. Number	Issue
2430071	**Description:** After reloading devlink in IPoIB setup, the IB link may stay in initialization state and require to run OpenSM to get the IB link to active state.
	Workaround: N/A
	Keywords: IPoIB devlink reload
	Discovered in Release: 5.2-2.2.0.0
2302786	**Description:** On EulerOS 2.0 SP9 systems, the kernel ABI (kABI) between the base vhulk2006 kernel and the errata vhulk2008 kernel has been changed. It is now not possible to install MLNX_OFED compiled with KMP on vhulk2006 kernel on a vhulk2008 system.
	Workaround: Install MLNX_OFED with --add-kernel-support.
	Keywords: EulerOS; kABI; installation; --add-kernel-support
	Discovered in Release: 5.2-1.0.4.0
2398281	**Description:** A crash in the TLS Rx socket cleanup flow may occur due to a kernel issue where a wrong extra call to tls_dev_del is made.
	Workaround: N/A
	Keywords: TLS RX device offload
	Discovered in Release: 5.2-1.0.4.0
2407415	**Description:** OpenEuler 20.03 Aarch64 with errata kernels 4.19.90-2011.6.0.0049.oe1.aarch64 and 4.19.90-2012.5.0.0054.oe1.aarch64 are incompatible with MLNX_OFED kmod-mlnx-ofa_kernel.
	Workaround: Install MLNX_OFED with --add-kernel-support.
	Keywords: OpenEuler; Aarch64; installation; --add-kernel-support
	Discovered in Release: 5.2-1.0.4.0
2348077	**Description:** RDMA device name for VFs may change after resetting all VFs at once.
	Workaround: Either reset interfaces one by one with a delay in between, or use a network interface naming scheme with predictable interface names, such as NAME_PCI or NAME_GUID. Copy

Internal Ref. Number	Issue
	/lib/udev/rules.d/60-rdma-persistent-naming.rules to /etc/udev/rules.d/ and edit the last line accordingly. Note that this will change interface names.
	Keywords: RDMA; VF
	Discovered in Release: 5.2-1.0.4.0
2381713	**Description:** esp4_offload and esp6_offload modules are expected to be loaded according to the list determined by the default kernel. However, these modules cannot be loaded when working over Debian 10 with non-default custom kernel as they are not included in it.
	Workaround: Either install MLNX_OFED using --add-kernel-support, or rebuild the non-default custom kernel to include these modules.
	Keywords: esp4_offload; esp6_offload; kernel, Debian
	Discovered in Release: 5.2-1.0.4.0
2382898	**Description**: On kernel 4.14, there is no traffic for UDP or TCP with payload size larger than 1398 on GENEVE IPv6 over VLAN tag interface.
	Workaround: N/A
	Keywords: GENEVE; stag; VLAN; UDP
	Discovered in Release: 5.2-1.0.4.0
2326155	**Description:** When toggling the link state while running RoCE traffic, the below warning may appear in the dmesg: __ib_cache_gid_add: unable to add gid <gid> error=-28
	Workaround: N/A
	Keywords: RoCE; __ib_cache_gid_add
	Discovered in Release: 5.2-1.0.4.0
2329654	**Description:** Running XDP over an IP tunnel may fail when working with kernels as old as version 4.14.
	Workaround: N/A
	Keywords: XDP, Kernel
	Discovered in Release: 5.2-1.0.4.0

Internal Ref. Number	Issue
2249156	**Description:** MLNX_OFED installation will remove qperf package in case it was done after qperf installation.
	Workaround: Make sure to install qperf package after installing MLNX_OFED, or re-install qperf after installing MLNX_OFED.
	Keywords: Installation; qperf
	Discovered in Release: 5.2-1.0.4.0
2355956	**Description:** OFED installation requires kernel config CONFIG_DEBUG_INFO to be set.
	Workaround: N/A
	Keywords: Installation; CONFIG_DEBUG_INFO
	Discovered in Release: 5.2-1.0.4.0
2362781	**Description:** Openibd may fail to unload the Inbox driver mlx5_ib on Ubuntu 18.04 PPC Boston server due to a bug in the Inbox drivers.
	Workaround: N/A
	Keywords: Openibd; Inbox; Ubuntu; mlx5_ib
	Discovered in Release: 5.2-1.0.4.0
2367659	**Description:** Upgrading the MLNX_OFED version that is configured as a YUM repository may yield warning messages from depmod about unknown symbols, such as: depmod: WARNING: /lib/modules/4.18.0-240.el8.×8664/extra/iser/ib_iser.ko needs unknown symbol ib_fmr_pool_unmap depmod: WARNING: /lib/modules/4.18.0-240.el8.×8664/extra/srp/ib_srp.ko needs unknown symbol ib_create_qp_user These warnings appear since the RPM packages upgrade occurs sequentially, and there is an upgrade dependency between some of the modules, which would create a state of upgrade inconsistency. These warnings are temporary and can be ignored as eventually all modules will be upgraded, and the warnings will no longer appear.
	Workaround: N/A
	Keywords: YUM; RPM; symbol; depmod; ISER; SRP
	Discovered in Release: 5.2-1.0.4.0

Internal Ref. Number	Issue
2385269	**Description:** The number of connections offloaded is limited to 100K when working with Kernel v5.9.
	Workaround: N/A
	Keywords: ASAP2; Connection Tracking; Kernel
	Discovered in Release: 5.2-1.0.4.0
2393169	**Description:** Mirroring is not supported with Connection Tracking when the source port is a VxLAN device.
	Workaround: N/A
	Keywords: ASAP^2; Connection Tracking; Mirroring
	Discovered in Release: 5.2-1.0.4.0
2395082	**Description:** A call trace may take place when moving from SwitchDev mode back to Legacy mode in Kernel v5.9 due to a kernel issue in tcf_block_unbind.
	Workaround: N/A
	Keywords: ASAP^2;SwitchDev; call trace; kernel; tcf_block_unbind
	Discovered in Release: 5.2-1.0.4.0

Internal Ref. Number	Issue
2354899	**Description:** ODP is not supported on RHEL7.x systems when running over an ETH link layer with RoCE disabled.
	Workaround: N/A
	Keywords: ODP, RHEL, RoCE
	Discovered in Release: 5.1-2.5.8.0
2338150	**Description:** Scatter to **CQE** feature should be disabled for the GPUDirect tests to work.
	Workaround: Set the MLX5_SCATTER_TO_CQE environment variable to 0 before the ib_send_bw command. For example: MLX5_SCATTER_TO_CQE=0 ib_send_bw -d <...>

Internal Ref. Number	Issue
	Keywords: CQE, GPUDirect
	Discovered in Release: 5.1-2.5.8.0
2295732	**Description:** Upgrading from legacy (mlnx-libs) to the current rdma-core based build using YUM (package manager) fails.
	Workaround: To perform this upgrade, either use the installer script or uninstall the old packages and install the new packages.
	Keywords: Legacy, mlnx-libs, rdma-core, installation
	Discovered in Release: 5.1-2.5.8.0
2295735	**Description:** Upgrading from legacy (mlnx-libs) to the current rdma-core based build using the apt-get (package manager) fails.
	Workaround: To perform this upgrade, either use the installer script or uninstall the old packages and install the new packages.
	Keywords: Legacy, mlnx-libs, rdma-core, apt, apt-get, installation
	Discovered in Release: 5.1-2.5.8.0
2248996	**Description:** Downgrading the firmware version for ConnectX-6 cards using "mlnx_ofed_install --fw-update-only --force-fw-update" fails.
	Workaround: Manually downgrade the firmware version - please see Firmware Update Instructions.
	Keywords: Firmware, ConnectX-6
	Discovered in Release: 5.1-0.6.6.0
2175930	**Description:** When using OFED 5.1 on PPC architectures with kernels v5.5 or v5.6 and an old ethtool utility, a harmless warning call trace may appear in the dmesg due to mismatch between user space and kernel. The warning call trace mentions ethtool_notify.
	Workaround: Update the ethtool utility to version 5.6 on such systems in order to avoid the call trace.
	Keywords: PPC, ethtool_notify, kernel
	Discovered in Release: 5.1-0.6.6.0

Internal Ref. Number	Issue
2198764	**Description:** If MLNX_OFED is installed on a Debian or Ubuntu system that is run in chroot environment, the openibd service will not be enabled. If the chroot files are being used as a base of a full system, the openibd service is left disabled.
	Workaround: Currently, openibd is a sysv-init script that you can enable manually by running: update-rc.d openibd defaults
	Keywords: chroot, Debian , Ubuntu, openibd
	Discovered in Release: 5.1-0.6.6.0
2237134	**Description:** Running connection tracking (CT) with FW steering may cause CREATE_FLOW_TABLE command to fail with syndrome.
	Workaround: Configure OVS to use a single handler-thread: #ovs-vsctl set Open_vSwitch . other_config:n-handler-threads=1
	Keywords: Connection tracking, ASAP, OVS, FW steering
	Discovered in Release: 5.1-0.6.6.0
2239894	**Description:** Running OpenVSwitch offload with high traffic throughput can cause low insertion rate due to high CPU usage.
	Workaround: Reduce the number of combined channels of the uplink using "ethtool -L".
	Keywords: Insertion rate, ASAP2
	Discovered in Release: 5.1-0.6.6.0
2240671	**Description:** Header rewrite action is not supported over RHEL/CentOS 7.4.
	Workaround: N/A
	Keywords: ASAP, header rewrite, RHEL, RedHat, CentOS, OS
	Discovered in Release: 5.1-0.6.6.0
2242546	**Description:** Tunnel offload (encap/decap) may cause kernel panic if nf_tables module is not probed.
	Workaround: Make sure to probe the nf_tables module before inserting any rule.

Internal Ref. Number	Issue
	Keywords: Kernel v5.7, ASAP, kernel panic
	Discovered in Release: 5.1-0.6.6.0
2143007	**Description:** IPsec packets are dropped during heavy traffic due to a bug in net/xfrm Linux Kernel.
	Workaround: Make sure the Kernel is modified to apply the following patch: "xfrm: Fix double ESP trailer insertion in IPsec crypto offload".
	Keywords: IPsec, xfrm
	Discovered in Release: 5.1-0.6.6.0
2225952	**Description:** VF mirroring with TC policy skip_sw is not supported on RHEL/CentOS 7.4, 7.5 and 7.6 OSs.
	Workaround: N/A
	Keywords: ASAP^2, Mirroring, RHEL, RedHat, OS
	Discovered in Release: 5.1-0.6.6.0
2216521	**Description:** After upgrading MLNX_OFED from v5.0 or earlier, ibdev2netdev utility changes the installation prefix to /usr/sbin. Therefore, it cannot be found while found in the same SHELL environment.
	Workaround: After installing MLNX_OFED, log out and log in again to refresh the SHELL environment.
	Keywords: ibdev2netdev
	Discovered in Release: 5.1-0.6.6.0
2202520	**Description:** Rules with VLAN push/pop, encap/decap and header rewrite actions together are not supported.
	Workaround: N/A
	Keywords: ASAP^2, SwitchDev, VLAN push/pop, encap/decap, header rewrite
	Discovered in Release: 5.1-0.6.6.0
2210752	**Description:** Switching from Legacy mode to SwitchDev mode and vice-

Internal Ref. Number	Issue
	versa while TC rules exist on the NIC will result in failure.
	Workaround: Before attempting to switch mode, make sure to delete all TC rules on the NIC or stop OpenvSwitch.
	Keywords: ASAP^2, Devlink, Legacy SR-IOV
	Discovered in Release: 5.1-0.6.6.0

Internal Ref. Number	Issue
2125036/2125031	**Description:** Upgrading the MLNX_OFED from an UPSTREAM_LIBS based version to an MLNX_LIBS based version fails unless the driver is uninstalled and then re-installed.
	Workaround: Make sure to uninstall and re-install MLNX_OFED to complete the upgrade.
	Keywords: Installation, UPSTREAM_LIBS, MLNX_LIBS
	Discovered in Release: 5.0-2.1.8.0
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: 5.0-2.1.8.0
2110321	**Description:** Multiple driver restarts may cause IPoIB soft lockup.
	Workaround: N/A
	Keywords: Driver restart, IPoIB
	Discovered in Release: 5.0-2.1.8.0
2112251	**Description:** On kernels 4.10-4.14, when Geneve tunnel's remote endpoint is defined using IPv6, packets larger than MTU are not

Internal Ref. Number	Issue
	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: 5.0-2.1.8.0
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: 5.0-2.1.8.0
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: 5.0-2.1.8.0
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: 5.0-1.0.0.0
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.

Internal Ref. Number	Issue
	Keywords: Bonding, MTU, Kernel
	Discovered in Release: 5.0-1.0.0.0
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: 5.0-1.0.0.0
2058535	**Description:** ibdev2netdev command returns duplicate devices with different ports in SwitchDev mode.
	Workaround: Use /opt/mellanox/iproute2/sbin/rdma link show command instead.
	Keywords: ibdev2netdev
	Discovered in Release: 5.0-1.0.0.0
2072568	**Description:** In RHEL/CentOS 7.2 OSs, adding drop rules when act_gact is not loaded may cause a kernel crash.
	Workaround: Preload all needed modules to avoid such a scenario (cls_flower, act_mirred, act_gact, act_tunnel_key and act_vlan).
	Keywords: RHEL/CentOS 7.2, Kernel 4.9, call trace, ASAP
	Discovered in Release: 5.0-1.0.0.0
2093698	**Description:** VF LAG configuration is not supported when the NUM_OF_VFS configured in mlxconfig is higher than 64.
	Workaround: N/A
	Keywords: VF LAG, SwitchDev mode, ASAP
	Discovered in Release: 5.0-1.0.0.0
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: 5.0-1.0.0.0
2000590	**Description:** Sending packets larger than MTU is not supported when working with OVS-DPDK.
	Workaround: N/A
	Keywords: MTU, OVS-DPDK
	Discovered in Release: 5.0-1.0.0.0
2062900	**Description:** Moving VF from SwitchDev mode to Legacy mode while the representor is being used by OVS-DPDK results in a segmentation fault.
	Workaround: To move VF to Legacy mode with no error, make sure to delete the ports from the OVS.
	Keywords: SwitchDev, Legacy, representor, OVS-DPDK
	Discovered in Release: 5.0-1.0.0.0
2075942	**Description:** Huge pages configuration is lost each time the server is configured.
	Workaround: Re-configure the huge pages after each reboot, or configure them as a kernel parameter.
	Keywords: Huge pages, reboot, OVS-DPDK
	Discovered in Release: 5.0-1.0.0.0
2067012	**Description:** MLNX_OFED cannot be installed on Debian 9.11 OS in SwitchDev mode.
	Workaround: Install OFED with the flag --add-kernel-support.
	Keywords: ASAP, SwitchDev, Debian, Kernel
	Discovered in Release: 5.0-1.0.0.0
2036572	**Description:** When using a thread domain and the lockless rdma-core ibv_post_send path, there is an additional CPU penalty due to required barriers around the device MMIO buffer that were omitted in MLNX_OFED.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: rdma-core, write-combining, MMIO buffer
	Discovered in Release: 5.0-1.0.0.0

Internal Ref. Number	Issue
-	**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
1997230	**Description:** Running mlxfwreset or unloading mlx5_core module while contrak flows are offloaded may cause a call trace in the kernel.
	Workaround: Stop OVS service before calling mlxfwreset or unloading mlx5_core module.
	Keywords: Contrak, ASAP, OVS, mlxfwrest, unload
	Discovered in Release: 4.7-3.2.9.0
1955352	**Description**: Moving 2 ports to SwitchDev mode in parallel is not supported.
	Workaround: N/A
	Keywords: ASAP, SwitchDev
	Discovered in Release: 4.7-3.2.9.0
1979958	**Description:** VxLAN IPv6 offload is not supported over CentOS/RHEL v7.2 OSs.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: Tunnel, VXLAN, ASAP, IPv6
	Discovered in Release: 4.7-3.2.9.0
1991710	**Description:** PRIO_TAG_REQUIRED_EN configuration is not supported and may cause call trace.
	Workaround: N/A
	Keywords: ASAP, PRIO_TAG, mstconfig
	Discovered in Release: 4.7-3.2.9.0
1967866	**Description:** Enabling ECMP offload requires the VFs to be unbound and VMs to be shut down.
	Workaround: N/A
	Keywords: ECMP, Multipath, ASAP^2
	Discovered in Release: 4.7-3.2.9.0
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
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
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.

Internal Ref. Number	Issue
	Keywords: Flow steering
	Discovered in Release: 4.7-1.0.0.1

Internal Ref. Number	Issue
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
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
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
-	**Description:** Aging is not functional on bond device in RHEL 7.6.
	Workaround: N/A
	Keywords: VF LAG, ASAP^2
	Discovered in Release: 4.6-1.0.1.1

Internal Ref. Number	Issue
1747774	**Description:** In VF LAG mode, outgoing traffic in load balanced mode is according to the origin ring, thus, half of the rings will be coupled with port 1 and half with port 2. All the traffic on the same ring will be sent from the same port.
	Workaround: N/A
	Keywords: VF LAG, ASAP^2
	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
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
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
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
-	**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: 200Gb/s, advertisement, Ethtool
	Discovered in Release: 4.6-1.0.1.1

Internal Ref. Number	Issue
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
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 • 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.

Internal Ref. Number	Issue
	Workaround: N/A
	Keywords: mlx5 driver, SLES 12 SP1
	Discovered in Release: 4.5-1.0.1.0

User Manual

- Introduction
- Installation
- Features Overview and Configuration
- Programming
- InfiniBand Fabric Utilities
- Troubleshooting
- Common Abbreviations and Related Documents

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.

NVIDIA OFED is a single Virtual Protocol Interconnect (VPI) software stack which operates across all NVIDIA network adapter solutions supporting the following uplinks to servers:

Uplink/Adapter Card	Driver Name	Uplink Speed
BlueField-2	mlx5	• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2]
BlueField		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE

Uplink/Adapter Card	Driver Name	Uplink Speed
ConnectX-7		• InfiniBand: EDR, HDR100, HDR, NDR200, NDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[3]
ConnectX-6 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2]
ConnectX-6 Dx		• Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-6		• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-5/ConnectX-5 Ex		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-4 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE
ConnectX-4		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE[1], 100GbE

1. 56GbE is an NVIDIA proprietary link speed and can be achieved while connecting an NVIDIA adapter card to NVIDIA SX10XX switch series or when connecting an NVIDIA adapter card to another NVIDIA adapter card.

2. Speed that supports both NRZ and PAM4 modes in Force mode and Auto-Negotiation mode.

3. Speed that supports PAM4 mode only.

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:

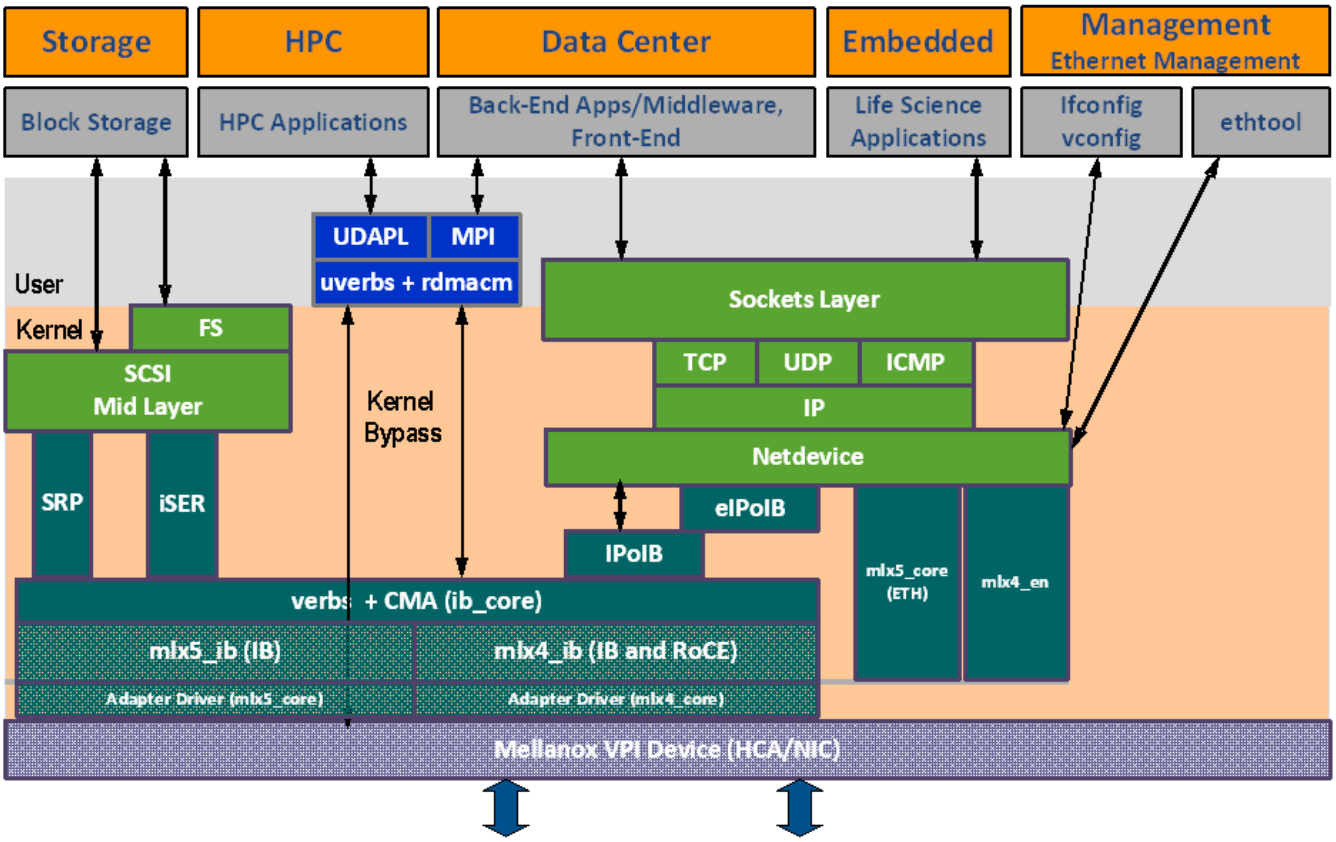
- NVIDIA Messaging Accelerator (VMA™) software: Socket acceleration library that performs OS bypass for standard socket-based applications.
  Please note, VMA support is provided separately from NVIDIA OFED support. For further information, please refer to the VMA documentation (docs.nvidia.com/networking/category/vma).

- 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

> **⚠ Warning**
>
> This driver is no longer supported in MLNX_OFED. To work with ConnectX-3® and ConnectX-3 Pro NICs, please refer to MLNX_OFED LTS version available on the web.

mlx5 Driver

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

> ⚠ **Warning**
>
> Please note that Connect-IB card is no longer supported in MLNX_OFED. To work with this card, please refer to MLNX_OFED LTS version available on the web.

mlx5_core

Acts as a library of common functions (e.g. initializing the device after reset) required by ConnectX-4 and above adapter cards. mlx5_core driver also implements the Ethernet interfaces for ConnectX-4 and above. 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 **Legacy** and **RDMA-Core** 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

The following are libmlx5 **Legacy** only environment variables:

- MLX5_ENABLE_CQE_COMPRESSION
    - Saves PCIe bandwidth by compressing a few CQEs into a smaller amount of bytes on PCIe. Setting this variable enables CQE compression.
    - Default value 0 (disabled)
- MLX5_RELAXED_PACKET_ORDERING_ON
  See "Out-of-Order (OOO) Data Placement" 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 (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. NVIDIA OFED includes the following MPI implementation over InfiniBand:

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

NVIDIA 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 NVIDIA OFED[1].

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

Diagnostic Utilities

NVIDIA 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 (MFT)

The NVIDIA Firmware Tools 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).

Package Contents

ISO Image

NVIDIA 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
        - mlx5_ib
        - mlx5_core (includes 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—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

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)

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 NVIDIA OFED for Linux package on a single host machine with NVIDIA InfiniBand and/or Ethernet adapter hardware installed.

The chapter contains the following sections:

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

Features Overview and Configuration

> ⚠ **Warning**
>
> It is recommended to enable the “above 4G decoding” BIOS setting for features that require large amount of PCIe resources.
>
> Such features are: SR-IOV with numerous VFs, PCIe Emulated Switch, and Large BAR Requests.

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.

Utility	Description
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.
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 ibutils2)	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

Utility	Description
	• Alias GUIDs check • Subnet Manager check • Partition keys check • Nodes information **Note:** This version of ibdiagnet is included in the ibutils2 package, and it is run by default after installing NVIDIA 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 NVIDIA's 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 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)

Utility	Description
	4. Open Wireshark and load the generated file To download Wireshark for a Linux or Windows environment go to www.wireshark.org. **Notes:** • Although ibdump is a Linux application, the generated .pcap file may be analyzed on either operating system. • If one of the HCA's ports 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. Optionally, 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.
ibportstate	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. If the queried port is a switch port, then ibportstate can be used to:

Utility	Description
	• 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.
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.

Utility	Description
	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. **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.

Utility	Description
	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, NVIDIA 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 NVIDIA 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_atom	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

Utility	Description
ic_bw	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.
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

Utility	Description
	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 NVIDIA

representative or NVIDIA Support at networking-support@nvidia.com.

.

The chapter contains the following sections:

- General Issues
- Ethernet Related Issues
- InfiniBand Related Issues
- Installation Related Issues
- Performance Related Issues
- SR-IOV Related Issues
- PXE (FlexBoot) Related Issues
- RDMA Related Issues
- Debugging Related Issues
- OVS Offload Using ASAP2 Direct Related Issues

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)
b	(Small) 'b' is used to indicate size in bits or multiples of bits (e.g., 1Kb = 1024 bits)

Abbreviation/ Acronym	Description
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
Subnet Administrator (SA)	An application (normally part of the Subnet Manager) that implements the interface for querying and manipulating subnet management data

Term	Description
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)	An NVIDIA technology that allows NVIDIA 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 NVIDIA adapter devices	See the Release Notes relevant to your adapter device
MFT User Manual and Release Notes	NVIDIA Firmware Tools (MFT) User Manual and Release Notes documents
WinOF User Manual	Mellanox WinOF User Manual describes the installation, configuration, and operation of NVIDIA Windows driver
VMA User Manual	NVIDIA VMA User Manual describes the installation, configuration, and operation of NVIDIA VMA driver

Documentation History

- Release Notes History
- User Manual Revision History

Release Notes History

- Changes and New Features History
- Bug Fixes History

User Manual Revision History

Release	Date	Description
23.10	November 2023	• Added IPsec Full Offload • Added a note to NVME-oF - NVM Express over Fabrics.
23.07	September 2023	No changes to the User Manual.
	August 2023	• Updated "Interrupt Request (IRQ) Naming" under Ethernet Interface. • Removed the "IPsec Crypto Offload" section from the Ethernet Network chapter. • Removed all "Connected Mode" indications, and the "IPoIB Mode Setting" section from IP over InfiniBand (IPoIB). • Removed Running OpenSM, and updated Running OpenSM as Daemon under NVIDIA SM.

Release	Date	Description
23.04	May 2023	• Removed "Installing NEO-Host Using mlnxofedinstall Script" subsection from the Installing MLNX_OFED section • Added note under "Enrolling NVIDIA's x.509 Public Key On your Systems" subsection in the UEFI Secure Boot section • Updated the NVIDIA SM (OpenSM) section
	Jun 2023	• Added an important note to IPsec Crypto Offload
5.7	August 2022	• Added Out of Order (OOO) under RoCE section • Added section describing Dumping Steering Info
5.6	April 2022	• Added OpenFlow Meters section • Added Installation on Community Operating Systems section
5.5	December 2021	• Added "Forward to Multiple Destinations" section • Added "Open vSwitch Metering" section • Added "Multiport eSwitch Mode" section • Added "Bridge Offload" section • Added "TC Configuration for ConnectX-6 Dx and Above" subsection
5.4-3	October 2021	• Removed LASH Routing Algorithm
5.4-2	August 2021	• Added CT CT NAT section • Added Representor Metering section • Updated VF Metering section • Removed sysfs VXLAN portion of the Enabling VXLAN Hardware Stateless Offloads section
5.4	June 2021	• Updated SR-IOV Live Migration
5.3-1	March 31,2021	• Added PTP Cyc2time Hardware Translation Offload section

Release	Date	Description
		• Added Connection Tracking Performance Tuning section • Added VF Metering section • Added note under OVS-DPDK Hardware Offloads section • Updated Persistent Naming section • Updated command under Connection Tracking Offload section • Added sFlow description under OVS-DPDK Hardware Offloads section
5.2	February 14, 2021	• Added Setting up MLNX_OFED YUM Repository Using --add-kernel-support section • Added Setting up MLNX_OFED apt-get Repository Using --add-kernel-support section
	January 19, 2021	Added OpenSSL with kTLS Offload section
	January 4, 2021	• Added Offloaded Traffic Sniffer section • Added Tx Port Time-Stamping section • Added VLAN Push/Pop section • Added sFLOW section • Added E2E Cache section • Added Geneve Encapsulation/Decapsulation section • Added Parallel Offloads section • Updated SR-IOV VF LAG section • Removed Installing MLNX_OFED on Innova™ IPsec Adapter Cards section • Removed Updating Firmware and FPGA Image on Innova IPsec Cards section
5.1-2	September 17, 2020	Added Packet Pacing for Hairpin Queues section
5.1	July 28, 2020	Updated the content of the entire document following the removal of support for ConnectX-3, ConnectX-3 Pro and Connect-IB adapter cards, as well as the deprecation of RDMA experimental verbs library (mlnx_lib)

Release	Date	Description
		Added SR-IOV Live Migration section
		Added SR-IOV VF LAG section
5.0-2	April 23, 2020	Added Interrupt Request (IRQ) Naming section
	April 6, 2020	Added Kernel Transport Layer Security (kTLS) Offloads section
5.0	March 3, 2020	• Added IPSec Crypto Offload section • Added OVS-DPDK Hardware Offloads section • Updated OVS Hardware Offloads Configuration section
4.7	December 29, 2019	• Added Configuring Uplink Representor Mode section
	December 13, 2019	• Added Performance Tuning Based on Traffic Patterns section • Added "num_of_groups" entry to table mlx5_core Module Parameters • Added .Mediated Devices v5.4-0.5.1.1-Beta section
	September 29, 2019	• Updated Additional Installation Procedures section
4.6	May 13, 2019	• ethtool section updates: Added description of -f flashing option to Ethtool Supported Options table

	April 30, 2019	• ethtool section updates:
		  ◦ Updated the description of ethtool -s eth<x> advertise <N> autoneg on counter under Ethtool
		  ◦ Added the following counters under Ethtool:
		  ▪ ethtool --show-fec eth<x>
		  ▪ ethtool --set-fec eth<x> encoding auto|off|rs|baser
		• Added Devlink Parameters section
		• Added Limit Bandwidth per Group of VFs section
		• Added Disabling RoCE section
		• Added RDMA-CM QP Timeout Control section
		• Added RDMA-CM Application Managed QP section

Release	Date	Description
4.5	December 19, 2018	• Reorganized Chapter 2, “Installation”: Consolidated the separate installation procedures under Installing NVIDIA OFED and Additional Installation Procedures • Added Installing NEO-Host Using mlnxofedinstall Script
	November 29, 2018	Added the following sections: • Local Loopback Disable • Offsweep Balancing

Legal Notices and 3rd Party Licenses

The following are the drivers' software, tools and HCA firmware legal notices and 3rd party licenses.

Product	Version	Legal Notices and 3rd Party Licenses
MLNX_OFED	23.10-0	• License • 3rd Part Notice
Firmware	xx.39.1002	• HCA Firmware EULA • 3rd Party Notice
MFT	4.26.0	• 3rd Party Notice • License
Clusterkit	1.11	• License • 3rd Party Notice
DPCP	1.1.43	• License • 3rd Party Notice
VMA	9.8.40	• 3rd Party Unify Notice • 3rd Party Notice
XLIO	3.20.7	• 3rd Party Unify Notice • 3rd Party Notice
HCOLL	4.8	• License • 3rd Party Notice
SHARP	3.5.0	• License

Product	Version	Legal Notices and 3rd Party Licenses
		• 3rd Party Notice
ibutils2	2.15	• License • 3rd Party Notice
OpenSM	5.17.0	• 3rd Party Unify Notice • 3rd Party Notice
mpitests	3.2.21	• License • 3rd Party Notice

© Copyright 2023, NVIDIA. PDF Generated on 06/05/2024