



## **NVIDIA MLNX\_OFED Documentation Rev 5.7-1.0.2.0**

# Table of contents

<b>Release Notes</b>	11
General Support	15
Changes and New Features	32
Bug Fixes in This Version	36
Known Issues	38
<b>User Manual</b>	79
Introduction	79
Installation	91
Features Overview and Configuration	92
Programming	93
InfiniBand Fabric Utilities	96
Troubleshooting	107
Common Abbreviations and Related Documents	108
<b>Documentation History</b>	112
Release Notes History	112
User Manual Revision History	112

# List of Figures

Figure 0. Image2019 1 10 10 55 56 Version 1 Modificationdate  
1701068080867 Api V2

---

Figure 1. Procedure Heading Icon Version 1 Modificationdate  
1701068083530 Api V2

---

Figure 2. Procedure Heading Icon Version 1 Modificationdate  
1701068083530 Api V2

---

Figure 3. Worddave635fed9c99097774044df72a47e9130 Version 1  
Modificationdate 1701068088487 Api V2

---

Figure 4. Image2019 2 12 10 26 53 Version 1 Modificationdate  
1701068094027 Api V2

---

Figure 5. Procedure Heading Icon Version 1 Modificationdate  
1549377199353 Api V2

---

Figure 6. Procedure Heading Icon Version 1 Modificationdate  
1549377199353 Api V2

---

Figure 7. Procedure Heading Icon Version 1 Modificationdate  
1701068098717 Api V2

---

Figure 8. Procedure Heading Icon Version 1 Modificationdate  
1701068100387 Api V2

---

Figure 9. Procedure Heading Icon Version 1 Modificationdate  
1701068100387 Api V2

---

Figure 10. Procedure Heading Icon Version 1 Modificationdate  
1701068105110 Api V2

---

---

Figure 11. Image2022 7 26 10 29 13 Version 1 Modificationdate  
1701068114427 Api V2

---

Figure 12. Image2022 7 28 14 58 53 Version 1 Modificationdate  
1701068114787 Api V2

---

Figure 13. Image2022 7 28 15 0 34 Version 1 Modificationdate  
1701068115877 Api V2

---

Figure 14. Image2022 7 28 14 59 57 Version 1 Modificationdate  
1701068115523 Api V2

---

Figure 15. Image2022 7 28 15 1 11 Version 1 Modificationdate  
1701068116270 Api V2

---

Figure 16. Image2022 7 28 15 1 42 Version 1 Modificationdate  
1701068116647 Api V2

---

Figure 17. Image2022 7 28 15 2 14 Version 1 Modificationdate  
1701068117027 Api V2

---

Figure 18. Image2022 7 28 15 4 11 Version 1 Modificationdate  
1701068117350 Api V2

---

Figure 19. Image2022 7 28 15 6 40 Version 1 Modificationdate  
1701068117820 Api V2

---

Figure 20. Procedure Heading Icon Version 1 Modificationdate  
1701068152193 Api V2

---

Figure 21. Procedure Heading Icon Version 1 Modificationdate  
1701068152193 Api V2

---

Figure 22. Procedure Heading Icon Version 1 Modificationdate  
1701068152193 Api V2

---

---

Figure 23. Procedure Heading Icon Version 1 Modificationdate  
1701068152193 Api V2

---

Figure 24. Procedure Heading Icon Version 1 Modificationdate  
1701068152193 Api V2

---

Figure 25. Procedure Heading Icon Version 1 Modificationdate  
1701068152193 Api V2

---

Figure 26. Procedure Heading Icon Version 1 Modificationdate  
1701068162393 Api V2

---

Figure 27. Procedure Heading Icon Version 1 Modificationdate  
1701068162393 Api V2

---

Figure 28. Procedure Heading Icon Version 1 Modificationdate  
1701068162393 Api V2

---

Figure 29. Procedure Heading Icon Version 1 Modificationdate  
1701068162393 Api V2

---

Figure 30. Worddavb2ee67a7eb9aae5c536610e39a37dcc5 Version 1  
Modificationdate 1701068167727 Api V2

---

Figure 31. Worddav6931c32564b3b0c166f4a26788219144 Version 1  
Modificationdate 1701068168613 Api V2

---

Figure 32. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 33. Image2019 3 8 12 50 6 Version 1 Modificationdate  
1701068169263 Api V2

---

Figure 34. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

---

Figure 35. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 36. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 37. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 38. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 39. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 40. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 41. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 42. Procedure Heading Icon Version 1 Modificationdate  
1701068168970 Api V2

---

Figure 43. Srios Live Migration Stage 2 Version 1 Modificationdate  
1701068171427 Api V2

---

Figure 44. Image2020 7 26 14 25 6 Version 1 Modificationdate  
1701068173067 Api V2

---

Figure 45. Image2020 7 26 14 26 25 Version 1 Modificationdate  
1701068173417 Api V2

---

Figure 46. Image2020 7 26 14 27 26 Version 1 Modificationdate  
1701068173787 Api V2

---

---

Figure 47. Image2020 7 26 14 28 39 Version 1 Modificationdate  
1701068174167 Api V2

---

Figure 48. Image2020 7 26 14 30 13 Version 1 Modificationdate  
1701068174457 Api V2

---

Figure 49. Image2020 7 26 14 33 10 Version 1 Modificationdate  
1701068174827 Api V2

---

Figure 50. Image2020 7 26 14 34 7 Version 1 Modificationdate  
1701068175230 Api V2

---

Figure 51. Image2020 7 26 14 34 18 Version 1 Modificationdate  
1701068175537 Api V2

---

Figure 52. Image2020 7 26 14 38 28 Version 1 Modificationdate  
1701068175907 Api V2

---

Figure 53. Image2020 7 26 14 39 14 Version 1 Modificationdate  
1701068176597 Api V2

---

Figure 54. Image2020 7 26 14 39 27 Version 1 Modificationdate  
1701068176990 Api V2

---

Figure 55. Procedure Heading Icon Version 1 Modificationdate  
1701068179773 Api V2

---

Figure 56. Worddav336f9b6791fd85e08c8e6897697cd75b Version 1  
Modificationdate 1701068180797 Api V2

---

Figure 57. Procedure Heading Icon Version 1 Modificationdate  
1701068182460 Api V2

---

Figure 58. Procedure Heading Icon Version 1 Modificationdate  
1701068182460 Api V2

---

---

Figure 59. Procedure Heading Icon Version 1 Modificationdate  
1701068182460 Api V2

---

Figure 60. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 61. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 62. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 63. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 64. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 65. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 66. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 67. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 68. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 69. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 70. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

---

Figure 71. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 72. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 73. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 74. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 75. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 76. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 77. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 78. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 79. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 80. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 81. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 82. Procedure Heading Icon Version 1 Modificationdate  
1701068185187 Api V2

---

Figure 83. Procedure Heading Icon Version 1 Modificationdate  
1701068187483 Api V2

---

## **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](https://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

## Release Notes Update History

Revision	Date	Description
5.7-1.0.2.0	July 31, 2022	Initial release of this document version.

### 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	<ul style="list-style-type: none"><li>InfiniBand: SDR, FDR, EDR, HDR</li><li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup></li></ul>
BlueField		<ul style="list-style-type: none"><li>InfiniBand: SDR, QDR, FDR, FDR10, EDR</li><li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE</li></ul>
ConnectX-7		<ul style="list-style-type: none"><li>InfiniBand: EDR, HDR100, HDR, NDR200, NDR</li><li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup>, 200GbE<sup>3</sup>, 400GbE</li></ul>
ConnectX-6 Lx		<ul style="list-style-type: none"><li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup></li></ul>
ConnectX-6 Dx		<ul style="list-style-type: none"><li>Ethernet: 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup>, 200GbE<sup>2</sup></li></ul>
ConnectX-6		<ul style="list-style-type: none"><li>InfiniBand: SDR, FDR, EDR, HDR</li><li>Ethernet: 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup>, 200GbE<sup>2</sup></li></ul>
ConnectX-5/ConnectX-5 Ex		<ul style="list-style-type: none"><li>InfiniBand: SDR, QDR, FDR, FDR10, EDR</li><li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE</li></ul>
ConnectX-4 Lx		<ul style="list-style-type: none"><li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE</li></ul>
ConnectX-4		<ul style="list-style-type: none"><li>InfiniBand: SDR, QDR, FDR, FDR10, EDR</li></ul>

<b>Uplink/Adapter Card</b>	<b>Driver Name</b>	<b>Uplink Speed</b>
		<ul style="list-style-type: none"> <li>• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE<sup>1</sup>, 100GbE</li> </ul>

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	1.7.421-1.57102	BSD
dapl	2.1.10.1.mlnx-OFED.4.9.0.1.4.57102	Dual GPL/BSD/CPL
dpcp	1.1.29-1.57102	Proprietary
dump_pr	1.0- 5.11.0.MLNX20220713.g58a37c9.571 02	GPLv2 or BSD
hcoll	4.8.3217-1.57102	Proprietary
ibdump	6.0.0-1.57102	BSD2+GPL2
ibsim	0.10-1.57102	GPLv2 or BSD
ibutils2	2.1.1- 0.151.MLNX20220720.gcd746c3.571 02	Mellanox Confidential and Proprietary
iser	5.7-OFED.5.7.1.0.2.1	GPLv2
isert	5.7-OFED.5.7.1.0.2.1	GPLv2
kernel-mft	4.21.0-99	Dual BSD/GPL

Package	Revision	Licenses
knem	1.1.4.90mlnx1-OFED.5.6.0.1.6.1	BSD and GPLv2
libvma	9.6.4-1	GPLv2 or BSD
libxlio	1.3.5-1	GPLv2 or BSD
mlnx-en	5.7-1.0.2.0.ge78f334	GPLv2
mlnx-ethtool	5.18-1.57102	GPL
mlnx-iproute2	5.18.0-1.57102	GPL
mlnx-nfsrdma	5.7-OFED.5.7.1.0.2.1	GPLv2
mlnx-nvme	5.7-OFED.5.7.1.0.2.1	GPLv2
mlnx-ofa_kernel	5.7-OFED.5.7.1.0.2.1	GPLv2
mlnx-tools	5.2.0-0.57102	GPLv2
mlx-steering-dump	1.0.0-0.57102	GPLv2
mpi-selector	1.0.3-1.57102	BSD
mpitests	3.2.20-de56b6b.57102	BSD
mstflint	4.16.1-2.57102	GPL/BSD
multiperf	3.0-3.0.57102	BSD 3-Clause, GPL v2 or later
ofed-docs	5.7-OFED.5.7.1.0.2	GPL/BSD
ofed-scripts	5.7-OFED.5.7.1.0.2	GPL/BSD
openmpi	4.1.5a1-1.57102	BSD
opensm	5.12.0.MLNX20220721.3a88a9b-0.1.57102	GPLv2 or BSD
openvswitch	2.17.2-1.57102	ASL 2.0 and LGPLv2+ and SISSL
perftest	4.5-0.17.g6f25f23.57102	BSD 3-Clause, GPL v2 or later
rdma-core	56mlnx40-1.57102	GPLv2 or BSD
rshim	2.0.6-13.gd253708	GPLv2
sharp	3.0.0.MLNX20220713.ea604271-1.57102	Proprietary

Package	Revision	Licenses
sockperf	3.8-0.git31ee322aa82a.57102	BSD
srp	5.7-OFED.5.7.1.0.2.1	GPLv2
ucx	1.14.0-1.57102	BSD
xpmem	2.6.3-1.57102	GPLv2 and GPLv2.1

# General Support

## Supported Operating Systems

Operating System	Architecture	Default Kernel Version
ALIOS7.2	AArch64	4.19.48-006.ali4000.alios7.aarch64
BCLINUX21.10SP2	AArch64	4.19.90-2107.6.0.0098.oe1.bclinux.aarch64
BCLINUX8.2	x86_64	4.19.0-240.23.11.el8_2.bclinux.x86_64
CTYUNOS2.0	AArch64	4.19.90-2102.2.0.0062.ctl2.aarch64
	x86_64	4.19.90-2102.2.0.0062.ctl2.x86_64
Debian10.8	AArch64	4.19.0-14-arm64
	x86_64	4.19.0-14-amd64
Debian10.9	x86_64	4.19.0-16-amd64
Debian11.2	AArch64	5.10.0-10-arm64
	x86_64	5.10.0-10-amd64
EulerOS2.0sp10	AArch64	4.19.90-vhulk2110.1.0.h860.eulerosv2r10.aarch64
	x86_64	4.18.0-147.5.2.4.h694.eulerosv2r10.x86_64
EulerOS2.0sp9	AArch64	4.19.90-vhulk2006.2.0.h171.eulerosv2r9.aarch64
	x86_64	4.18.0-147.5.1.0.h269.eulerosv2r9.x86_64

Operating System	Architecture	Default Kernel Version
Fedora32	x86_64	5.6.6-300.fc32.x86_64
KYLIN10	AArch64	4.19.90-17.ky10.aarch64
	x86_64	4.19.90-17.ky10.x86_64
KYLIN10SP2	AArch64	4.19.90-24.4.v2101.ky10.aarch64
	x86_64	4.19.90-24.4.v2101.ky10.x86_64
Oracle Linux 7.9	x86_64	5.4.17-2011.6.2.el7uek.x86_64
Oracle Linux 8.2	x86_64	5.4.17-2011.1.2.el8uek.x86_64
Oracle Linux 8.3	x86_64	5.4.17-2011.7.4.el8uek.x86_64
Oracle Linux 8.4	x86_64	5.4.17-2102.2013.el8uek.x86_64
Oracle Linux 8.5	x86_64	5.4.17-2136.300.7.el8uek.x86_64
Oracle Linux 8.6	x86_64	5.4.17-2136.307.3.1.el8uek.x86_64
OPENEULER20.03SP1	AArch64	4.19.90-2012.4.0.0053.oe1.aarch64
	x86_64	4.19.90-2012.5.0.0054.oe1.x86_64
OPENEULER20.03SP3	AArch64	4.19.90-2112.8.0.0131.oe1.aarch64
	x86_64	4.19.90-2112.8.0.0131.oe1.x86_64
RHEL/CentOS7.2	x86_64	3.10.0-327.el7.x86_64
RHEL/CentOS7.4	ppc64	3.10.0-693.el7.ppc64
	ppc64le	3.10.0-693.el7.ppc64le
	x86_64	3.10.0-693.el7.x86_64
RHEL/CentOS7.5	ppc64	3.10.0-862.el7.ppc64
	ppc64le	3.10.0-862.el7.ppc64le
	x86_64	3.10.0-862.el7.x86_64
RHEL/CentOS7.5alternate	AArch64	4.14.0-49.el7a.aarch64
RHEL/CentOS7.6	ppc64	3.10.0-957.el7.ppc64
	ppc64le	3.10.0-957.el7.ppc64le

Operating System	Architecture	Default Kernel Version
	x86_64	3.10.0-957.el7.x86_64
RHEL/CentOS7.6alternate	AArch64	4.14.0-115.el7a.aarch64
	ppc64le	4.14.0-115.el7a.ppc64le
RHEL/CentOS7.7	ppc64	3.10.0-1062.el7.ppc64
	ppc64le	3.10.0-1062.el7.ppc64le
	x86_64	3.10.0-1062.el7.x86_64
RHEL/CentOS7.8	ppc64	3.10.0-1127.el7.ppc64
	ppc64le	3.10.0-1127.el7.ppc64le
	x86_64	3.10.0-1127.el7.x86_64
RHEL/CentOS7.9	ppc64	3.10.0-1160.el7.ppc64
	ppc64le	3.10.0-1160.el7.ppc64le
	x86_64	3.10.0-1160.el7.x86_64
RHEL/CentOS8.0	AArch64	4.18.0-80.el8.aarch64
	ppc64le	4.18.0-80.el8.ppc64le
	x86_64	4.18.0-80.el8.x86_64
RHEL/CentOS8.1	AArch64	4.18.0-147.el8.aarch64
	ppc64le	4.18.0-147.el8.ppc64le
	x86_64	4.18.0-147.el8.x86_64
RHEL/CentOS8.2	AArch64	4.18.0-193.el8.aarch64
	ppc64le	4.18.0-193.el8.ppc64le
	x86_64	4.18.0-193.el8.x86_64
RHEL/CentOS8.3	AArch64	4.18.0-240.el8.aarch64
	ppc64le	4.18.0-240.el8.ppc64le
	x86_64	4.18.0-240.el8.x86_64
RHEL/CentOS8.4	AArch64	4.18.0-305.el8.aarch64
	ppc64le	4.18.0-305.el8.ppc64le

Operating System	Architecture	Default Kernel Version
	x86_64	4.18.0-305.el8.x86_64
RHEL/CentOS8.5	AArch64	4.18.0-348.el8.aarch64
	ppc64le	4.18.0-348.el8.ppc64le
	x86_64	4.18.0-348.el8.x86_64
RHEL8.6	AArch64	4.18.0-372.9.1.el8.aarch64
	ppc64le	4.18.0-372.9.1.el8.ppc64le
	x86_64	4.18.0-372.9.1.el8.x86_64
RHEL9.0	AArch64	5.14.0-70.13.1.el9_0.aarch64
	ppc64le	5.14.0-70.13.1.el9_0.ppc64le
	x86_64	5.14.0-70.13.1.el9_0.x86_64
SLES12SP3	ppc64le	4.4.73-5-default
	x86_64	4.4.73-5-default
SLES12SP4	AArch64	4.12.14-94.41-default
	ppc64le	4.12.14-94.41-default
	x86_64	4.12.14-94.41-default
SLES12SP5	AArch64	4.12.14-120-default
	ppc64le	4.12.14-120-default
	x86_64	4.12.14-120-default
SLES15SP2	AArch64	5.3.18-22-default
	ppc64le	5.3.18-22-default
	x86_64	5.3.18-22-default
SLES15SP3	AArch64	5.3.18-57-default
	ppc64le	5.3.18-57-default
	x86_64	5.3.18-57-default
SLES15SP4	AArch64	5.14.21-150400.22-default
	ppc64le	5.14.21-150400.22-default

Operating System	Architecture	Default Kernel Version
	x86_64	5.14.21-150400.22-default
Ubuntu18.04	AArch64	4.15.0-20-generic
	ppc64le	4.15.0-20-generic
	x86_64	4.15.0-20-generic
Ubuntu20.04	AArch64	5.4.0-26-generic
	ppc64le	5.4.0-26-generic
	x86_64	5.4.0-26-generic
Ubuntu22.04	AArch64	5.15.0-25-generic
	ppc64le	5.15.0-25-generic
	x86_64	5.15.0-25-generic
UOS20	AArch64	4.19.0-arm64-server
	x86_64	4.19.0-server-amd64
UOS20.1020	AArch64	4.19.90-2109.1.0.0108.up2.uel20.aarch64
	x86_64	4.19.90-2109.1.0.0108.up2.uel20.x86_64
Citrix XenServer Host7.1	x86_64	4.4.0+2
Citrix XenServer Host8.2	x86_64	4.19.0+1
Kernel 5.17	AArch64	5.17
	ppc64le	5.17
	x86_64	5.17

## Supported Community Operating Systems

Community OS	Architecture	Tested with Kernel Version
BCLINUX7.6	x86_64	3.10.0-957.el7.x86_64

Community OS	Architecture	Tested with Kernel Version
BCLINUX7.7	AArch64	4.19.25-203.el7.bclinux.aarch64
	x86_64	3.10.0-1062.el7.bclinux.x86_64
BCLINUX8.1	x86_64	4.19.0-193.1.3.el8.bclinux.x86_64
Debian9.13	AArch64	4.9.0-13-arm64
	x86_64	4.9.0-13-amd64
EulerOS2.0sp5	x86_64	3.10.0-862.14.1.5.h591.eulerosv2r7.x86_64
EulerOS2.0sp8	AArch64	4.19.36-vhulk1907.1.0.h748.eulerosv2r8.aarch64
RHEL/CentOS7.5	ppc64	3.10.0-862.el7.ppc64
	ppc64le	3.10.0-862.el7.ppc64le
	x86_64	3.10.0-862.el7.x86_64
RHEL/CentOS7.5alternative	AArch64	4.14.0-49.el7a.aarch64
RHEL/CentOS7.6alternative	aarch64	4.14.0-115.el7a.aarch64
	ppc64le	4.14.0-115.el7a.ppc64le
RHEL/CentOS7.7	ppc64	3.10.0-1062.el7.ppc64
	ppc64le	3.10.0-1062.el7.ppc64le
	x86_64	3.10.0-1062.el7.x86_64
RHEL/CentOS7.8	ppc64	3.10.0-1127.el7.ppc64
	ppc64le	3.10.0-1127.el7.ppc64le
	x86_64	3.10.0-1127.el7.x86_64
SLES12SP2	x86_64	4.4.21-69-default
SLES12SP3	ppc64le	4.4.73-5-default
	x86_64	4.4.73-5-default
SLES12SP4	AArch64	4.12.14-94.41-default
	ppc64le	4.12.14-94.41-default
	x86_64	4.12.14-94.41-default

Community OS	Architecture	Tested with Kernel Version
Ubuntu16.04	ppc64le	4.4.0-21-generic
	x86_64	4.4.0-21-generic
Alma 8.5	x86_64	4.18.0-348.12.2.EL8_5.X86_64
Anolis OS 8.4	AArch64	4.18.0-348.2.1.AN8_4.AARCH64
	x86_64	4.18.0-305.AN8.X86_64
CentOS Stream	AArch64	-
	ppc64le	-
	x86_64	4.18.0-365.EL8.X86_64
Fedora 35	x86_64	5.16.8-200.fc35.x86_64
OpenEuler OS 22.03 LTS	AArch64	5.10.0-60.18.0.50.OE2203.AARCH64
	x86_64	5.10.0-60.18.0.50.oe2203.x86_64
OpenSUSE 15.3	AArch64	-
	ppc64le	-
	x86_64	5.3.18-150300.59.43-DEFAULT
Photon OS 3.0	x86_64	4.19.225-3.ph3
Rocky 8.5	AArch64	-
	x86_64	4.18.0-348.12.2.EL8_5.X86_64
Rocky 8.6	AArch64	4.18.0-372.9.1.EL8.AARCH64
	x86_64	4.18.0-372.9.1.el8.x86_64

## ⚠️ 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

## 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
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®

### ASAP2 Supported Operating Systems

## OVS-Kernel SR-IOV Based Supported Operating Systems

Below is a list of all the operating systems that support OVS-Kernel ASAP<sup>2</sup> in the current software package.

- BCLinux 7.7
- BCLinux 8.1
- Debian 10.9
- Debian 11.2
- RHEL/CentOS 7.4
- RHEL/CentOS 7.5
- RHEL/CentOS 7.6
- RHEL/CentOS 7.7
- RHEL/CentOS 7.8
- RHEL/CentOS 8.0
- RHEL/CentOS 8.1
- RHEL/CentOS 8.2
- RHEL/CentOS 8.3
- RHEL/Centos 8.4
- RHEL/Centos 8.5
- RHEL 8.6
- RHEL 9.0
- Oracle Linux 7.8
- Oracle Linux 8.1

- Oracle Linux 8.2
- Oracle Linux 8.3
- Oracle Linux 8.4
- SLES12 SP4
- SLES12 SP5
- SLES15 SP2
- SLES 15 SP3
- SLES 15 SP4
- Ubuntu 18.04
- Ubuntu 20.04
- Ubuntu 22.04

## OVS-DPDK SR-IOV Based Supported OSs

Below is a list of all the operating systems that support OVS-DPDK ASAP<sup>2</sup> in the current software package.

- RHEL/CentOS 7.4
- RHEL/CentOS 7.5
- RHEL/CentOS 7.6
- RHEL/CentOS 7.7
- RHEL/CentOS 7.8
- RHEL/CentOS 8.0
- RHEL/CentOS 8.1

- RHEL/CentOS 8.2
- RHEL/CentOS 8.3
- RHEL/CentOS 8.4
- RHEL/CentOS 8.5
- Ubuntu 18.04
- Ubuntu 20.04
- SLES15 SP2

## ASAP2 Requirements

- iproute >= 4.12 (for tc support)
- Upstream Open vSwitch >= 2.8 for CentOS 7.2 NVIDIA openvswitch

## ASAP2 Supported Adapter Cards

- ConnectX-5
- ConnectX-6 Dx
- ConnectX-6 Lx
- ConnectX-7

## UCX and CUDA Versions Compatible with MLNX\_OFED

UCX Version	Built with CUDA version	Compatible with CUDA versions
1.13	11.6	11.x

## NFS over RDMA (NFSoRDMA) Supported Operating Systems

### NFSoRDMA Supported Operating Systems on Client and Target Sides

- RHEL7.5
- RHEL7.6
- RHEL7.7
- RHEL7.8
- RHEL7.9
- RHEL 8.1
- RHEL8.2
- RHEL8.4
- RHEL8.5
- RHEL8.6
- RHEL9
- SLES12SP4
- SLES12SP5
- SLES15SP2
- SLES15SP3

- SLES15SP4
- Ubuntu-16.04
- Ubuntu-18.04
- Ubuntu-20.04
- Ubuntu-22.04
- Debian 10.9
- Debian 11.2
- Kernel 5.17

## **Lustre Versions Compatible with MLNX\_OFED**

- Lustre 2.15.0
- Lustre 2.12.9

## **NEO-Host Supported Operating Systems**

- RedHat 7.4
- RedHat 7.5
- RedHat 7.6
- RedHat 7.7
- RedHat 7.8
- RedHat 7.9
- RedHat 8.x
- Oracle Linux 7.9

- Oracle Linux 8.1
- Oracle Linux 8.2
- Oracle Linux 8.3
- Fedora 32
- SLES 12 SP3
- SLES 12 SP4
- SLES 12 SP5
- SLES 15 SP2
- SLES 15 SP3
- Ubuntu 18.04
- Ubuntu 20.04

## **GPUDirect Storage (GDS) Supported Operating Systems**

- Rhel 8.x
- Ubuntu 18.04
- Ubuntu 20.04

## **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/Fedora	yum install perl pciutils python gcc-gfortran libxml2-python tcsh libnl.i686 libnl expat glib2 tcl libstdc++ bc tk gtk2 atk cairo numactl pkgconfig ethtool lsof
XenServer	yum install perl pciutils python libxml2-python libnl expat glib2 tcl bc libstdc++ tk pkgconfig ethtool
SLES 12	zypper install pkg-config expat libstdc++6 libglib-2_0_0 lib- gtk-2_0_0 tcl libcairo2 tcsh python bc pciutils libatk-1_0_0 tk python-libxml2 lsof libnl3-200 ethtool lsof
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 libglib-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

## NVMe Supported Operating Systems

- RHEL8.6
- RHEL9.0
- SLES15SP4
- Ubuntu-22.04
- Kernel 5.17
- Fedora35
- RHEL7.2/CentOS7.2

- RHEL7.2/CentOS7.4
- RHEL7.5/CentOS7.5
- RHEL7.6/CentOS7.6
- RHEL7.7/CentOS7.7
- RHEL7.8/CentOS7.8
- RHEL7.9/CentOS7.9
- RHEL8.2
- RHEL8.4
- RHEL8.5
- SLES12SP4
- SLES12SP5
- SLES15SP2
- SLES15SP3
- Ubuntu16.04
- Ubuntu18.04
- Ubuntu20.04
- Debian 10.9
- Debian 11.2

## Supported NIC Firmware Versions



**Warning**

As of version 5.1, ConnectX-3, ConnectX-3 Pro or Connect-IB NICs 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 NIC firmware versions:

### **Warning**

Firmware versions listed are the minimum supported versions.

<b>Adapter Card</b>	<b>Recommended Firmware Version</b>	<b>Additional Firmware Version Supported</b>
BlueField®-2	24.34.1002	24.33.1048
BlueField	18.33.1048	N/A
ConnectX-7	28.34.1002	28.33.2028
ConnectX-6 Lx	26.34.1002	26.33.1048
ConnectX-6 Dx	22.34.1002	22.33.1048
ConnectX-6	20.34.1002	20.33.1048
ConnectX-5/ConnectX-5 Ex	16.34.1002	16.33.1048
ConnectX-4 Lx	14.32.1010	N/A
ConnectX-4	12.28.2006	N/A

For the official firmware versions, please see [https://www.nvidia.com/en-us/networking/Support\\_Support\\_Firmware\\_Download](https://www.nvidia.com/en-us/networking/Support_Support_Firmware_Download).

## **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)

## Changes and New Features

### 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 <b><u>unless specifically stated otherwise:</u></b> 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 <b><u>unless specifically stated otherwise:</u></b> ConnectX-6 Dx / ConnectX-6 Lx / ConnectX-7 / BlueField-2
ConnectX-6 and above	Supported in the following adapter cards <b><u>unless specifically stated otherwise:</u></b>

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

Feature/Change	Description
5.7-1.0.2.0	
Support Representor Metering Over SFs	[ConnectX-6 Dx and above and BlueField-2] Extended the support of representor metering from supporting only VFs representor to also supporting SFs representor.
Exposing Error Counters on a VPort Manager	[ConnectX-4 and above] Added support for exposing error counters on a VPort manager function for all other VPorts. These counters can be used to detect malicious users who are exploiting flows that can slow the device. The counters are exposed through debugfs under: <code>/sys/kernel/debug/mlx5/esw/&lt;func&gt;/vnic_diag/</code>
Ethtool RX Flow Steering for IPoIB	[All HCAs] Enabled steering of IPoIB packets via Ethtool, in the same way it is done today for Ethernet packets.
Memory Consumption Minimization	[ConnectX-4 and above] Added support for providing knobs which enable users to minimize memory consumption of mlx5 functions (PF/VF/SF).
XDP Support for Uplink Representors	[ConnectX-5 and ConnectX-6 Dx] Added XDP support for uplink representors in switchdev mode.
Resiliency to tx_port_ts	[ConnectX-6 Dx and above] Added resiliency to the tx_port_ts feature. private-flag may be enabled via ethtool tx_port_ts which provides a more accurate time-stamp. In very rare cases, the said

Feature/Change	Description
	time-stamp was lost, leading to losing the synchronization altogether. This feature allows for fast recovery and allows to quickly regain synchronization.
Database of Devlink Health Asserts	[ConnectX-4 and above] Health buffer now contains more debug information like the epoch time in sec of the error and the error's severity. The print to dmesg is done with the debug level corresponding to the error's severity. This allows the user to use dmesg attribute: dmesg --level to focus on different severity levels of firmware errors.
Expose FEC Counters via Ethtool	[ConnectX-5 and above] Exposed the following FEC (forward error detection) counters: <b>ETHTOOL_A_FEC_STAT_CORRECTED</b> <ul style="list-style-type: none"> <li>• fc_fec_corrected_blocks_laneX</li> <li>• rs_fec_corrected_blocks</li> </ul> <b>ETHTOOL_A_FEC_STAT_UNCORR</b> <ul style="list-style-type: none"> <li>• fc_fec_uncorrectable_blocks_laneX</li> <li>• rs_fec_uncorrectable_blocks</li> </ul> <b>ETHTOOL_A_FEC_STAT_CORR_BITS</b> <ul style="list-style-type: none"> <li>• phy_corrected_bits</li> </ul> Command: ethtool -I show-fec <ifc>
Application Device Queues	[ConnectX-4 and above] Added driver-level support for Application Device Queues. This feature allows partition defining over the RX/TX queues into groups and isolates traffic of different applications. This mainly improves predictability and tail latency.
Reinjection of Packets Into Kernel	[All HCAs] Added support for a new software steering action, mlx5dv_dr_action_create_dest_root_table(). This action can be used to forward packets back into a level 0 table. As a table with level 0 is the kernel owned table, this will result in injecting packets to the kernel steering pipeline.
DCT LAG	[ConnectX-6 Dx and above] Added firmware support to allow explicit port selection based on steering and not QP affinity. Functionality:

Feature/Change	Description
	1. Use LAG Hash Mode for the HCA with two ports, if supported. 2. Keep port affinity function in LAG Hash Mode if it supports bypass select flow table in non-SwitchDev mode.
AES-XTS in RDMA	Added support for plaintext AES-XTS DEKs.
General	<u>Bug fixes</u>

For additional information on the new features, please refer to [MLNX\\_OFED User Manual](#).

## Customer Affecting Changes

Customer Affecting Change	Description
5.7-1.0.2.0	
Multi-Block Encryption	Multi-block encryption is currently unsupported, due to a hardware limitation.

There are no customer affecting changes in this version.

## 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
3032335	<p><b>Description:</b> Creating multiple steering rules that modify a packet and match on the same packet headers can cause an error to be displayed in dmesg when deleting the steering rules.</p> <p><b>Keywords:</b> Steering Rules</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3011368	<p><b>Description:</b> Some IB spec QP state behaviour on post_send()/recv() is not being fully enforced. The fix makes the QP compliant to IB spec about when it is allowed to post_send()/recv() and when it should return an error.</p> <p><b>Keywords:</b> RDMA, IB spec QP</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3075125	<p><b>Description:</b> When changing trust state from PCP to DSCP, the TC number changes by default to 8, in some cases, disrupting traffic prioritization if trust state is changed back to PCP.</p> <p><b>Keywords:</b> NetDev, QoS</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3054413	<p><b>Description:</b> In the current release, the following OPNs/PSIDs should be manually upgraded:</p> <p>MCX753106AS-HEA-N NVD0000000023 MCX75310AAS-HEA-N NVD0000000024</p> <p><b>Keywords:</b> ConnectX-7, Upgrade</p>

Internal Reference Number	Description
	<p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
	<p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3070653	<p><b>Description:</b> In versions of MLNX_OFED before 5.7, the xpmem kernel module was not signed. When it was installed on systems (mostly RHEL and other compatible systems) the following error message would appear: "xpmem: loading out-of-tree module taints kernel."</p> <p><b>Keywords:</b> Installation, xpmem</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3075357	<p><b>Description:</b> In Debian-based distributions, in /etc/init.d/openibd, the path to enable the firmware tracer is /sys/kernel/debug/tracing/events/mlx5/fw_tracer/enable instead of /sys/kernel/debug/tracing/events/mlx5/mlx5_fw/enable . As a result, firmware tracer will never get enabled even when supported.</p> <p><b>Keywords:</b> Installation, Kernel Trace Debug</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
2688191	<p><b>Description:</b> The minimum Tx rate limit is not supported with link speed of 1Gb/s.</p> <p><b>Keywords:</b> Rate Limit, 1Gb/s</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3044255	<p><b>Description:</b> Destroying mlxdevm group while SF is attached to it is not supported.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, mlxdevm, QoS, Group, Scalable Functions</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>

Internal Reference Number	Description
3047142	<p><b>Description:</b> Using OVS offload with NIC mode (non switchdev mode) causes traffic to drop.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Offload, NIC Mode, OVS</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3123986	<p><b>Description:</b> In some cases VF metering configuration failure caused a deadlock.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, VF Metering</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>
3053842	<p><b>Description:</b> A race condition may cause some connection aging to set to 24 hours instead of 30 seconds.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Connection Tracking, Aging</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p> <p><b>Fixed in Release:</b> 5.7-1.0.2.0</p>

## Known Issues

The following is a list of general limitations and known issues of the current version of the release. For the list of old known issues, please refer to NVIDIA OFED Archived Known

Issues file at

[http://www.mellanox.com/pdf/prod\\_software/MLNX\\_OFED\\_Archived\\_Known\\_Issues.pdf](http://www.mellanox.com/pdf/prod_software/MLNX_OFED_Archived_Known_Issues.pdf)

Internal Ref. Num ber	Issue
3114 823	<p><b>Description:</b> The first attempt to create a new iSER connection fails with the following messages in dmesg:</p> <pre>iSCSI Login timeout on Network Portal &lt;iSER_Target_IP_ADDR&gt;:3260 isert: isert_get_login_rx: isert_conn 00000000e9239d52 interrupted before got login req</pre> <p>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.</p> <ul style="list-style-type: none"><li>• RHEL 9.0</li><li>• RHEL 8.6</li><li>• Ubuntu 22.04</li><li>• SLES 15 SP4</li></ul> <p>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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> iSER Initiator</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3096 911	<p><b>Description:</b> 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).</p> <p><b>Workaround:</b> Installing chkconfig before OFED.</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3100 544	<p><b>Description:</b> 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:</p>

Internal Ref. Num ber	Issue
	<p>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</p> <p>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.</p> <p><b>Workaround:</b> dnf install kernel-modules- # in case it is not the newest.</p> <p><b>Keywords:</b> Installation, RHEL9.x</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3132 158	<p><b>Description:</b> Building rdma-core package on Rocky 8.6 OS caused failure in OFED build.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3137 440	<p><b>Description:</b> Python package is missing, need to install it manually.</p> <p><b>Workaround:</b> Install Python before starting the build.</p> <p><b>Keywords:</b> Installation, Python</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3141 506	<p><b>Description:</b> kernel-macros package does not support building with KMP enabled. KMP needs to be disabled.</p> <p><b>Workaround:</b> Build and install MOFED with KMP disabled (without --kmp flag).</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3141 506	<p><b>Description:</b> kernel-macros package does not support building with KMP enabled. KMP needs to be disabled.</p> <p><b>Workaround:</b> Build and install MOFED with KMP disabled (without --kmp flag).</p> <p><b>Keywords:</b> Installation</p>

Internal Ref. Num ber	Issue
	<p><b>Discovered in Release:</b> 5.7-1.0.2.0</p> <p><b>Description:</b> Starting firmware version XX.34.0350, a new NVCONFIG has been added to the ARM side only: MANAGEMENT_PF_MODE. If this config is on, the user will see a PCI Function (PF) which failed to probe:</p> <pre>[ 6.837102] mlx5_core 0000:03:00.2: mlx5_cmd_check:756:(pid 206): ENABLE_HCA(0x104) op_mod(0x0) failed, status bad parameter(0x3), syndrome (0x6ca1f5) [ 6.864227] mlx5_core 0000:03:00.2: mlx5_peer_pf_init:40:(pid 206): Failed to enable peer PF HCA err(-22) [ 6.883453] mlx5_core 0000:03:00.2: mlx5_load:1129:(pid 206): Failed to init embedded CPU [ 8.261268] mlx5_core 0000:03:00.2: init_one:1365:(pid 206): mlx5_load_one failed with error code -22 [ 8.280056] mlx5_core: probe of 0000:03:00.2 failed with error -22</pre>
3142 212	<p><b>Workaround:</b> To clear the error, disable MANAGEMENT_PF_MODE NVCONFIG (set it to 2).</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
3129 627	<p><b>Description:</b> Kernel module packaging is not supported in CtyunOS.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>
2971 708	<p><b>Description:</b> 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.</p>

Internal Ref. Num ber	Issue
	<p>To determine if Devlink can be used to enable or disable RoCE, execute the following console command after starting OFED:</p> <pre data-bbox="323 508 866 544">devlink dev param show   grep roce</pre> <p>Devlink supports roce enable/disable if the following line is reflected in the output:</p> <pre data-bbox="323 699 796 734">name enable_roce type generic</pre> <p>For OSs which do not allow enabling/disabling RoCE via Devlink, the sysfs interface behaves as in the previous 2 releases:</p> <ol style="list-style-type: none"> <li>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.</li> <li>2. For OSs which do not have Devlink reload, RoCE state is managed only by the sysfs interface.        '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):</li> </ol> <pre data-bbox="401 1353 936 1389">devlink dev help 2&gt;&amp;1   grep reload</pre> <p>Reload is supported if the output is:</p> <pre data-bbox="401 1491 1168 1526">devlink dev reload DEV [ netns { PID   NAME   ID } ]</pre> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Enabling/Disabling RoCE</p> <p><b>Discovered in Release:</b> 5.7-1.0.2.0</p>

Internal Ref. Num ber	Issue
2998 194	<p><b>Description:</b> 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).</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> NetDev, RHEL, Virtual Functions</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
3045 436	<p><b>Description:</b> Rebooting the host while the Arm is down may block the shutdown flow till the Arm is up.</p> <p><b>Workaround:</b> Restart the driver on the host side before reboot.</p> <p><b>Keywords:</b> Reboot, Arm</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
3040 350	<p><b>Description:</b></p> <ol style="list-style-type: none"> <li>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.</li> <li>2. When offload is enabled, the physical port must be attached to a bridge.</li> </ol> <p><b>Workaround:</b></p> <ol style="list-style-type: none"> <li>1. When removing a physical port from an ovs-dpdk bridge while offload is enabled, need to restart openvswitch after reattaching it.</li> <li>2. Attach physical port to a bridge according to the desired topology.</li> </ol> <p><b>Keywords:</b> OVS-DPDK, Bridge, Offload</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>

Internal Ref. Num ber	Issue
2973 726	<p><b>Description:</b> dec_ttl only work with ConnectX-6. It does not work with ConnectX-5.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> OVS-DPDK, dec_ttl</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
	<p><b>Description:</b> Moving to switchdev mode while deleting namespace may cause a deadlock.</p> <p><b>Workaround:</b> Unload mlx5_ib module before moving to Switchdev mode.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Switchdev, Namespace</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, mlx5_core</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Use add-kernel-support.</p> <p><b>Keywords:</b> Installation,CentOS</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>

Internal Ref. Num ber	Issue
3011 440	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Install also mstflint of the exact same version (e.g., apt install mlnx-ofed-all-exact mstflint=4.16.0-1.56xxxx).</p> <p><b>Keywords:</b> Installation, Debian, Ubuntu, MST</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Remove the files, if needed.</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Either disable building xpmem by adding --without-xpmem to the command line, or edit the kernel Makefile to make it unset LDFLAGS:</p> <pre>sed -i -e '/^export ARCH/iLDFLAGS :=' /lib/modules/\$(uname -r)/Makefile</pre> <p><b>Note:</b> The Makefile may be located elsewhere, such as the top-level directory of the kernel source directory.</p> <p><b>Keywords:</b> Installation, SLES</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
3046 601	<p><b>Description:</b> A package manager upgrade with zypper (on a SLES system) may prompt a question about vendor change from "Mellanox Technologies" to</p>

Internal Ref. Num ber	Issue
	<p>"OpenFabrics".</p> <p><b>Workaround:</b> Either accept this when prompted or add the file /etc/zypp/vendors.d/mlnx_ofed with the following content:</p> <pre>[main] vendors = Mellanox,OpenFabrics</pre>
	<p><b>Keywords:</b> Installation, SLES</p>
	<p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
3048	<p><b>Description:</b> 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.</p>
411	<p><b>Workaround:</b> Run the command 'dracut -f' to update the initramfs.</p>
	<p><b>Keywords:</b> Installation</p>
	<p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
3048	<p><b>Description:</b> OFED installation failed using yum for --add-kernel-support option (building packages without KMP enabled) if libfabric package is installed.</p>
444	<p><b>Workaround:</b> Remove libfabric package before OFED installation or use installation script.</p>
	<p><b>Keywords:</b> Installation, RHEL 8.5</p>
	<p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
3015	<p><b>Description:</b> OVS topology where the tunnel device is over a VF and the VF representor is connected to a bond is not supported.</p>
210	<p><b>Workaround:</b> N/A</p>
	<p><b>Keywords:</b> ASAP<sup>2</sup>, ConnectX-6 Dx, Tunnel Over VF, LAG, Connection Tracking</p>
	<p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>

Internal Ref. Num ber	Issue
3028 300	<b>Description:</b> OVS metering is not support over kernel 5.17.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP <sup>2</sup> , OVS, Meter, Kernel 5.17
	<b>Discovered in Release:</b> 5.6-1.0.3.3
3044 255	<b>Description:</b> Destroying mlxdevm group while SF is attached to it is not supported.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP <sup>2</sup> , mlxdevm, QoS, Group, Scalable Functions, ConnectX-6 Dx
	<b>Discovered in Release:</b> 5.6-1.0.3.3
2900 346	<b>Description:</b> 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).
	<b>Workaround:</b> N/A
	<b>Keywords:</b> BlueField-2, Ubuntu, ARP Flux
	<b>Discovered in Release:</b> 5.6-1.0.3.3
3046 456	<b>Description:</b> Switching between SwitchDev mode and legacy mode quickly on BlueField-2 can prevent the driver from loading successfully and breaks its health recovery.
	<b>Workaround:</b> Pause 60 seconds between state-altering commands to guarantee the driver health recovery is completed successfully.
	<b>Keywords:</b> ASAP <sup>2</sup> , BlueField-2, Health Recovery
	<b>Discovered in Release:</b> 5.6-1.0.3.3
2934 149	<b>Description:</b> Adding vDPA ports over ConnectX-5 devices in ovs-dpdk is not supported and will cause a crash.

Internal Ref. Num ber	Issue
	<p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> OVS-DPDK, ConnectX-5</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
2934 833	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> RDMA, Port Toggle</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>
2901 514	<p><b>Description:</b> Relaxed Ordering is not working properly on Virtual Functions.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Relaxed Ordering, VF</p> <p><b>Discovered in Release:</b> 5.6-1.0.3.3</p>

Internal Ref. Num ber	Issue
2870 299	<p><b>Description:</b> Managing SFs is possible using the iproute2 with mlxdevm tool only.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Scalable Functions</p>

Internal Ref. Num ber	Issue
	<b>Discovered in Release:</b> 5.5-1.0.3.2
2869 722	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, DKMS</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>
2870 367	<p><b>Description:</b> On UOS, IPoIB PKEY may require manual bring up after driver restart.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, IPoIB, PKEY</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>
2836 032	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Software Steering</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>
2851 639	<p><b>Description:</b> Enabling ARFS in legacy mode and then moving to switchdev mode is not supported and may cause unwanted behavior.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> NetDev, ARFS</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>

Internal Ref. Num ber	Issue
2851 639	<p><b>Description:</b> nvme and iser are not enabled on UOS ARM, because of missing UOS kernel support.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> nvme, iser, UOS ARM</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>
2860 855	<p><b>Description:</b> Building OFED on RHEL 8.4 with kmp disabled and then installing with yum fails due to some conflicting packages.</p> <p><b>Workaround:</b> Remove libfabric and librpmem packages before OFED installation, or add --allowerasing option to the installation command.</p> <p><b>Keywords:</b> Installation, RHEL 8.4, kmp, yum</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>
2865 983	<p><b>Description:</b> OFED packages were built with kmp disabled. Building with kmp enabled fails due to missing packages.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, kmp</p> <p><b>Discovered in Release:</b> 5.5-1.0.3.2</p>

Internal Ref. Num ber	Issue
26586 44	<p><b>Description:</b> Only match on lower 32 bit of ct_label is supported.</p> <p><b>Workaround:</b> N/A</p>

Internal Ref. Number	Issue
	<p><b>Keywords:</b> ASAP<sup>2</sup>, Connection Tracking</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
27063 45	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> To use large number of VFs, set PF_NUM_PF_MSIX to a smaller value of around 32.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, VF, PF_NUM_PF_MSIX</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
28369 97	<p><b>Description:</b> 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).</p> <p>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).</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Meter Threshold</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
28634 56	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, IPsec Full Offload</p> <p><b>Discovered in Release:</b> 5.4-0.5.1.1</p>

Internal Ref. Number	Issue
26573 92	<p><b>Description:</b> OFED installation caused CIFS to break in RHEL8.4 and RHEL8.5. A dummy module was added so that CIFS will be disabled after OFED installation in RHEL8.4 and RHEL8.5.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, RHEL8.4, RHEL8.5, CIFS</p> <p><b>Discovered in Release:</b> 5.4-0.5.1.1</p>
28009 93	<p><b>Description:</b> OpenMPI does not support running across different operating systems and/or CPU architectures.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> OpenMPI</p>
23995 03	<p><b>Description:</b> Open vSwitch is not supported on the latest operating systems containing only Python3 support.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Python, Open vSwitch</p>
26573 92	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, RHEL8.4, CIFS</p> <p><b>Discovered in Release:</b> 5.4-0.5.1.1</p>
27824 06	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> 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.</p> <p><b>Keywords:</b> Upgrade, kylin</p>

Internal Ref. Number	Issue
	<b>Discovered in Release:</b> 5.4-3.0.3.0
27556 32	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> SR-IOV, VST, QinQ</p> <p><b>Discovered in Release:</b> 5.4-3.0.3.0</p>
27804 36	<p><b>Description:</b> Non-default MTU (&gt;1500) is not supported with IPsec crypto offload and may cause packet drops.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> IPsec, Crypto Offload, MTU</p> <p><b>Discovered in Release:</b> 5.4-3.0.3.0</p>
27260 21	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation, openEuler</p> <p><b>Discovered in Release:</b> 5.4-3.0.3.0</p>

Internal Ref. Number	Issue
27506 53	<b>Description:</b> Running fragmented traffic in RHEL 8.3 (4.18.0-240.el8.x86_64) may cause call trace in build_skb.

Internal Ref. Number	Issue
	<p><b>Workaround:</b> Update to RHEL 8.3 z-stream 4.18.0-240.22.1.el8_3.x86_64.</p> <p><b>Keywords:</b> RHEL 8.3, Kernel Panic, Call Trace, fr</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
26293 75	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Connection Tracking, Label</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
27079 97	<p><b>Description:</b> Installation in the package manager mode under SLES 15.x may require user-intervention if the original libibverbs is installed.</p> <p><b>Workaround:</b> zypper install --force-resolution mlnx-ofed-all</p> <p><b>Keywords:</b> Installation, libibverbs</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
27085 31	<p><b>Description:</b> Installation in the package manager mode under SLES 15.x may require user-intervention if the original libopenvswitch is installed.</p> <p><b>Workaround:</b> zypper install --force-resolution mlnx-ofed-all</p> <p><b>Keywords:</b> Installation</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
27030 43	<p><b>Description:</b> Congested TCP lock for kTLS TX device offload traffic compromises the performance.</p> <p><b>Workaround:</b> Disable TCP selective acknowledgement: echo 0 &gt; /proc/sys/net/ipv4/tcp_sack</p> <p><b>Keywords:</b> kTLS TX</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
26764 05	<p><b>Description:</b> If the package interface-rename is active (on XenServer, for example), the interface renaming by the OFED will not be done to eliminate</p>

Internal Ref. Number	Issue
	<p>conflicts.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Interface Renaming</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
26879 43	<p><b>Description:</b> Offload of rules which redirect from VF on one PF to VF on second PF is not supported on socket-direct devices.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Socket-Direct</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
26786 72	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> SwitchDev, mlx5, Tunnel Traffic</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
25665 48	<p><b>Description:</b> 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 EEH may 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.</p> <p><b>Workaround:</b> Administrator may disable EEH before running firmware sync reset on the device.</p> <p><b>Keywords:</b> PPC, EEH</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
26179 50	<p><b>Description:</b> 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.</p>

Internal Ref. Number	Issue
	<p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Ethtool</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>
23907 31	<p><b>Description:</b> Ethtool does not display Port Speed advertised/capability above 100Gb/s over and below kernels 5.0, even when supported.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Ethtool, Port Speed</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>

Internal Ref. Number	Issue
	<p><b>Description:</b> Activating VF/SF LAG when at least one VF/SF is still bound may lead to an internal error in the firmware.</p> <p><b>Workaround:</b> Make sure all VFs/SFs are unbound prior to VF/SF LAG activation/deactivation.</p> <p><b>Keywords:</b> VF, SF, Firmware, Binding</p> <p><b>Discovered in Release:</b> 5.4-1.0.3.0</p>

Internal Ref. Number	Issue
2585 575	<p><b>Description:</b> After disabling sync reset by setting enable_remote_dev_reset to false, running firmware sync reset a few times may lead to general protection</p>

Internal Ref. Num ber	Issue
	<p>fault and system may get stuck.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Firmware Upgrade</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2582 565	<p><b>Description:</b> Conducting a firmware reset or unbinding the PF while in switchdev mode may cause a kernel crash.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> SwitchDev, ASAP<sup>2</sup>, Unbind, Firmware Reset</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2587 802	<p><b>Description:</b> PTP synchronization may be lost while using tx_port_ts private flag.</p> <p><b>Workaround:</b> Toggle private flag:  ethtool --set-priv-flags tx_port_ts off  ethtool --set-priv-flags tx_port_ts on  restart ptp4l application</p> <p><b>Keywords:</b> PTP Synchronization</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2574 943	<p><b>Description:</b> When running kernel 5.8 and bellow or RHEL 8.2 and below, sampled packets do not support tunnel information.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, sFLOW</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2568 417	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p>

Internal Ref. Num ber	Issue
	<p><b>Keywords:</b> Upgrade; mlx5_fpga_tools</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2506 425	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation; modules; kmod</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2492 509	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> add-kernel-support</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2479 327	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Upgrade; SLES 12; add-kernel-support</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2584 441	<p><b>Description:</b> 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.</p>

Internal Ref. Num ber	Issue
	<p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Upgrade; SLES 12; add-kernel-support</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2460 865	<p><b>Description:</b> When setting MTU to low values, such as 68 bytes, packets may fail on oversize.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> MTU</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2383 318	<p><b>Description:</b> On kernels based on RedHat 7.2, the "tx_port_ts" feature, as set by ethtool —set-priv-flags, is disabled.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> RedHat; tx_port_ts</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>
2575 647	<p><b>Description:</b> An OvS-DPDK crash might occur while doing live-migration for VMs that use virtio-interfaces that are accelerated using OvS-DPDK vDPA ports.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> OvS-DPDK vDPA, Live-migration</p> <p><b>Discovered in Release:</b> 5.3-1.0.0.1</p>

Internal Ref. Num ber	Issue
24300 71	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p>

Internal Ref. Number	Issue
	<p><b>Keywords:</b> IPoIB devlink reload</p> <p><b>Discovered in Release:</b> 5.2-2.2.0.0</p>
23027 86	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Install MLNX_OFED with --add-kernel-support.</p> <p><b>Keywords:</b> EulerOS; kABI; installation; --add-kernel-support</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>
23982 81	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> TLS RX device offload</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>
24074 15	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Install MLNX_OFED with --add-kernel-support.</p> <p><b>Keywords:</b> OpenEuler; Aarch64; installation; --add-kernel-support</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>
23480 77	<p><b>Description:</b> RDMA device name for VFs may change after resetting all VFs at once.</p> <p><b>Workaround:</b> 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 /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.</p> <p><b>Keywords:</b> RDMA; VF</p>

Internal Ref. Number	Issue
	<b>Discovered in Release:</b> 5.2-1.0.4.0
23817 13	<b>Description:</b> 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.
	<b>Workaround:</b> Either install MLNX_OFED using --add-kernel-support, or rebuild the non-default custom kernel to include these modules.
	<b>Keywords:</b> esp4_offload; esp6_offload; kernel, Debian
	<b>Discovered in Release:</b> 5.2-1.0.4.0
23828 98	<b>Description:</b> 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.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> GENEVE; stag; VLAN; UDP
	<b>Discovered in Release:</b> 5.2-1.0.4.0
23261 55	<b>Description:</b> When toggling the link state while running RoCE traffic, the below warning may appear in the dmesg: <code>_ib_cache_gid_add: unable to add gid &lt;gid&gt; error=-28</code>
	<b>Workaround:</b> N/A
	<b>Keywords:</b> RoCE; <code>_ib_cache_gid_add</code>
	<b>Discovered in Release:</b> 5.2-1.0.4.0
23296 54	<b>Description:</b> Running XDP over an IP tunnel may fail when working with kernels as old as version 4.14.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> XDP, Kernel
	<b>Discovered in Release:</b> 5.2-1.0.4.0
22491 56	<b>Description:</b> MLNX_OFED installation will remove qperf package in case it was done after qperf installation.

Internal Ref. Number	Issue
	<p><b>Workaround:</b> Make sure to install qperf package after installing MLNX_OFED, or re-install qperf after installing MLNX_OFED.</p> <p><b>Keywords:</b> Installation; qperf</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>
23559 56	<p><b>Description:</b> OFED installation requires kernel config CONFIG_DEBUG_INFO to be set.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Installation; CONFIG_DEBUG_INFO</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>
23627 81	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Openibd; Inbox; Ubuntu; mlx5_ib</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>
23676 59	<p><b>Description:</b> Upgrading the MLNX_OFED version that is configured as a YUM repository may yield warning messages from depmod about unknown symbols, such as:</p> <pre data-bbox="279 1353 1405 1495">depmod: WARNING: /lib/modules/4.18.0-240.el8.x86_64/extra/iser/ib_iser.ko needs unknown symbol ib_fmr_pool_unmap depmod: WARNING: /lib/modules/4.18.0-240.el8.x86_64/extra/srp/ib_srp.ko needs unknown symbol ib_create_qp_user</pre> <p>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.</p> <p>These warnings are temporary and can be ignored as eventually all modules will be upgraded, and the warnings will no longer appear.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> YUM; RPM; symbol; depmod; ISER; SRP</p> <p><b>Discovered in Release:</b> 5.2-1.0.4.0</p>

Internal Ref. Number	Issue
23852 69	<b>Description:</b> The number of connections offloaded is limited to 100K when working with Kernel v5.9.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP2; Connection Tracking; Kernel
	<b>Discovered in Release:</b> 5.2-1.0.4.0
23931 69	<b>Description:</b> Mirroring is not supported with Connection Tracking when the source port is a VxLAN device.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP <sup>2</sup> ; Connection Tracking; Mirroring
	<b>Discovered in Release:</b> 5.2-1.0.4.0
23950 82	<b>Description:</b> 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.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP <sup>2</sup> ; SwitchDev; call trace; kernel; tcf_block_unbind
	<b>Discovered in Release:</b> 5.2-1.0.4.0

Internal Ref. Number	Issue
235489 9	<b>Description:</b> ODP is not supported on RHEL7.x systems when running over an ETH link layer with RoCE disabled.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ODP, RHEL, RoCE

Internal Ref. Number	Issue
	<b>Discovered in Release:</b> 5.1-2.5.8.0
2338150	<b>Description:</b> Scatter to <b>CQE</b> feature should be disabled for the GPUDirect tests to work.
	<b>Workaround:</b> 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 <...>
	<b>Keywords:</b> CQE, GPUDirect
	<b>Discovered in Release:</b> 5.1-2.5.8.0
2295732	<b>Description:</b> Upgrading from legacy (mlnx-libs) to the current rdma-core based build using YUM (package manager) fails.
	<b>Workaround:</b> To perform this upgrade, either use the installer script or uninstall the old packages and install the new packages.
	<b>Keywords:</b> Legacy, mlnx-libs, rdma-core, installation
	<b>Discovered in Release:</b> 5.1-2.5.8.0
2295735	<b>Description:</b> Upgrading from legacy (mlnx-libs) to the current rdma-core based build using the apt-get (package manager) fails.
	<b>Workaround:</b> To perform this upgrade, either use the installer script or uninstall the old packages and install the new packages.
	<b>Keywords:</b> Legacy, mlnx-libs, rdma-core, apt, apt-get, installation
	<b>Discovered in Release:</b> 5.1-2.5.8.0
2248996	<b>Description:</b> Downgrading the firmware version for ConnectX-6 cards using "mlnx_ofed_install --fw-update-only --force-fw-update" fails.
	<b>Workaround:</b> Manually downgrade the firmware version - please see <a href="#">Firmware Update Instructions</a> .
	<b>Keywords:</b> Firmware, ConnectX-6
	<b>Discovered in Release:</b> 5.1-0.6.6.0

Internal Ref. Number	Issue
2175930	<p><b>Description:</b> 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.</p>
	<p><b>Workaround:</b> Update the ethtool utility to version 5.6 on such systems in order to avoid the call trace.</p>
	<p><b>Keywords:</b> PPC, ethtool_notify, kernel</p>
	<p><b>Discovered in Release:</b> 5.1-0.6.6.0</p>
2198764	<p><b>Description:</b> 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.</p>
	<p><b>Workaround:</b> Currently, openibd is a sysv-init script that you can enable manually by running: update-rc.d openibd defaults</p>
	<p><b>Keywords:</b> chroot, Debian , Ubuntu, openibd</p>
	<p><b>Discovered in Release:</b> 5.1-0.6.6.0</p>
2237134	<p><b>Description:</b> Running connection tracking (CT) with FW steering may cause CREATE_FLOW_TABLE command to fail with syndrome.</p>
	<p><b>Workaround:</b> Configure OVS to use a single handler-thread: #ovs-vsctl set Open_vSwitch . other_config:n-handler-threads=1</p>
	<p><b>Keywords:</b> Connection tracking, ASAP, OVS, FW steering</p>
	<p><b>Discovered in Release:</b> 5.1-0.6.6.0</p>
2239894	<p><b>Description:</b> Running OpenVSwitch offload with high traffic throughput can cause low insertion rate due to high CPU usage.</p>
	<p><b>Workaround:</b> Reduce the number of combined channels of the uplink using "ethtool -L".</p>
	<p><b>Keywords:</b> Insertion rate, ASAP2</p>
	<p><b>Discovered in Release:</b> 5.1-0.6.6.0</p>

Internal Ref. Number	Issue
224067 1	<b>Description:</b> Header rewrite action is not supported over RHEL/CentOS 7.4.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP, header rewrite, RHEL, RedHat, CentOS, OS
	<b>Discovered in Release:</b> 5.1-0.6.6.0
224254 6	<b>Description:</b> Tunnel offload (encap/decap) may cause kernel panic if nf_tables module is not probed.
	<b>Workaround:</b> Make sure to probe the nf_tables module before inserting any rule.
	<b>Keywords:</b> Kernel v5.7, ASAP, kernel panic
	<b>Discovered in Release:</b> 5.1-0.6.6.0
214300 7	<b>Description:</b> IPsec packets are dropped during heavy traffic due to a bug in net/xfrm Linux Kernel.
	<b>Workaround:</b> Make sure the Kernel is modified to apply the following patch: "xfrm: Fix double ESP trailer insertion in IPsec crypto offload".
	<b>Keywords:</b> IPsec, xfrm
	<b>Discovered in Release:</b> 5.1-0.6.6.0
222595 2	<b>Description:</b> VF mirroring with TC policy skip_sw is not supported on RHEL/CentOS 7.4, 7.5 and 7.6 OSs.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP <sup>2</sup> , Mirroring, RHEL, RedHat, OS
	<b>Discovered in Release:</b> 5.1-0.6.6.0
221652 1	<b>Description:</b> 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.
	<b>Workaround:</b> After installing MLNX_OFED, log out and log in again to refresh the SHELL environment.
	<b>Keywords:</b> ibdev2netdev

Internal Ref. Number	Issue
	<b>Discovered in Release:</b> 5.1-0.6.6.0
2202520	<p><b>Description:</b> Rules with VLAN push/pop, encap/decap and header rewrite actions together are not supported.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, SwitchDev, VLAN push/pop, encap/decap, header rewrite</p> <p><b>Discovered in Release:</b> 5.1-0.6.6.0</p>
2210752	<p><b>Description:</b> Switching from Legacy mode to SwitchDev mode and vice-versa while TC rules exist on the NIC will result in failure.</p> <p><b>Workaround:</b> Before attempting to switch mode, make sure to delete all TC rules on the NIC or stop OpenvSwitch.</p> <p><b>Keywords:</b> ASAP<sup>2</sup>, Devlink, Legacy SR-IOV</p> <p><b>Discovered in Release:</b> 5.1-0.6.6.0</p>

Internal Ref. Number	Issue
212503 6/21250 31	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Make sure to uninstall and re-install MLNX_OFED to complete the upgrade.</p> <p><b>Keywords:</b> Installation, UPSTREAM_LIBS, MLNX_LIBS</p> <p><b>Discovered in Release:</b> 5.0-2.1.8.0</p>
210544 7	<b>Description:</b> hns_roce warning messages will appear in the dmesg after reboot on Euler2 SP3 OSS.

Internal Ref. Number	Issue
	<p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> hns_roce, dmesg, Euler</p> <p><b>Discovered in Release:</b> 5.0-2.1.8.0</p>
211032 1	<p><b>Description:</b> Multiple driver restarts may cause IPoIB soft lockup.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Driver restart, IPoIB</p> <p><b>Discovered in Release:</b> 5.0-2.1.8.0</p>
211225 1	<p><b>Description:</b> On kernels 4.10-4.14, when Geneve tunnel's remote endpoint is defined using IPv6, packets larger than MTU are not fragmented, resulting in no traffic sent.</p> <p><b>Workaround:</b> Define geneve tunnel's remote endpoint using IPv4.</p> <p><b>Keywords:</b> Kernel, Geneve, IPv4, IPv6, MTU, fragmentation</p> <p><b>Discovered in Release:</b> 5.0-2.1.8.0</p>
210290 2	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> TLS offload, mlx5e</p> <p><b>Discovered in Release:</b> 5.0-2.1.8.0</p>
211153 4	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> TLS offload, mlx5e</p> <p><b>Discovered in Release:</b> 5.0-2.1.8.0</p>
203595 0	<p><b>Description:</b> 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.</p> <ol style="list-style-type: none"> <li>1. Removing PF from the bond (using ifdown, ip link or any other function)</li> </ol>

Internal Ref. Number	Issue
	<p>2. Attempting to disable SR-IOV</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> VF LAG, binding, firmware, FW, PF, SR-IOV</p> <p><b>Discovered in Release:</b> 5.0-1.0.0.0</p>
204454 4	<p><b>Description:</b> When working with OSs with Kernel v4.10, bonding module does not allow setting MTUs larger than 1500 on a bonding interface.</p> <p><b>Workaround:</b> Upgrade your Kernel version to v4.11 or above.</p> <p><b>Keywords:</b> Bonding, MTU, Kernel</p> <p><b>Discovered in Release:</b> 5.0-1.0.0.0</p>
188293 2	<p><b>Description:</b> Libibverbs dependencies are removed during OFED installation, requiring manual installation of libraries that OFED does not reinstall.</p> <p><b>Workaround:</b> Manually install missing packages.</p> <p><b>Keywords:</b> libibverbs, installation</p> <p><b>Discovered in Release:</b> 5.0-1.0.0.0</p>
205853 5	<p><b>Description:</b> ibdev2netdev command returns duplicate devices with different ports in SwitchDev mode.</p> <p><b>Workaround:</b> Use /opt/mellanox/iproute2/sbin/rdma link show command instead.</p> <p><b>Keywords:</b> ibdev2netdev</p> <p><b>Discovered in Release:</b> 5.0-1.0.0.0</p>
207256 8	<p><b>Description:</b> In RHEL/CentOS 7.2 OSs, adding drop rules when act_gact is not loaded may cause a kernel crash.</p> <p><b>Workaround:</b> Preload all needed modules to avoid such a scenario (cls_flower, act_mirred, act_gact, act_tunnel_key and act_vlan).</p> <p><b>Keywords:</b> RHEL/CentOS 7.2, Kernel 4.9, call trace, ASAP</p> <p><b>Discovered in Release:</b> 5.0-1.0.0.0</p>

Internal Ref. Number	Issue
209369 8	<b>Description:</b> VF LAG configuration is not supported when the NUM_OF_VFS configured in mlxconfig is higher than 64.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> VF LAG, SwitchDev mode, ASAP
	<b>Discovered in Release:</b> 5.0-1.0.0.0
209374 6	<b>Description:</b> Devlink health dumps are not supported on kernels lower than v5.3.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> Devlink, health report, dump
	<b>Discovered in Release:</b> 5.0-1.0.0.0
200059 0	<b>Description:</b> Sending packets larger than MTU is not supported when working with OVS-DPDK.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> MTU, OVS-DPDK
	<b>Discovered in Release:</b> 5.0-1.0.0.0
206290 0	<b>Description:</b> Moving VF from SwitchDev mode to Legacy mode while the representor is being used by OVS-DPDK results in a segmentation fault.
	<b>Workaround:</b> To move VF to Legacy mode with no error, make sure to delete the ports from the OVS.
	<b>Keywords:</b> SwitchDev, Legacy, representor, OVS-DPDK
	<b>Discovered in Release:</b> 5.0-1.0.0.0
207594 2	<b>Description:</b> Huge pages configuration is lost each time the server is configured.
	<b>Workaround:</b> Re-configure the huge pages after each reboot, or configure them as a kernel parameter.
	<b>Keywords:</b> Huge pages, reboot, OVS-DPDK
	<b>Discovered in Release:</b> 5.0-1.0.0.0

Internal Ref. Number	Issue
206701 2	<b>Description:</b> MLNX_OFED cannot be installed on Debian 9.11 OS in SwitchDev mode.
	<b>Workaround:</b> Install OFED with the flag --add-kernel-support.
	<b>Keywords:</b> ASAP, SwitchDev, Debian, Kernel
	<b>Discovered in Release:</b> 5.0-1.0.0.0
203657 2	<b>Description:</b> 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.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> rdma-core, write-combining, MMIO buffer
	<b>Discovered in Release:</b> 5.0-1.0.0.0

Internal Ref. Number	Issue
-	<b>Description:</b> 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.
	<b>Workaround:</b> Install the argparse module manually.
	<b>Keywords:</b> Python, MFT, argparse, installation
	<b>Discovered in Release:</b> 4.7-3.2.9.0
1997230	<b>Description:</b> Running mlxfwreset or unloading mlx5_core module while contrak flows are offloaded may cause a call trace in the kernel.
	<b>Workaround:</b> Stop OVS service before calling mlxfwreset or unloading mlx5_core module.
	<b>Keywords:</b> Contrak, ASAP, OVS, mlxfwrest, unload

Internal Ref. Number	Issue
	<b>Discovered in Release:</b> 4.7-3.2.9.0
1955352	<b>Description:</b> Moving 2 ports to SwitchDev mode in parallel is not supported.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP, SwitchDev
	<b>Discovered in Release:</b> 4.7-3.2.9.0
1979958	<b>Description:</b> VxLAN IPv6 offload is not supported over CentOS/RHEL v7.2 OSs.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> Tunnel, VXLAN, ASAP, IPv6
	<b>Discovered in Release:</b> 4.7-3.2.9.0
1991710	<b>Description:</b> PRIO_TAG_REQUIRED_EN configuration is not supported and may cause call trace.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ASAP, PRIO_TAG, mstconfig
	<b>Discovered in Release:</b> 4.7-3.2.9.0
1967866	<b>Description:</b> Enabling ECMP offload requires the VFs to be unbound and VMs to be shut down.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> ECMP, Multipath, ASAP <sup>2</sup>
	<b>Discovered in Release:</b> 4.7-3.2.9.0
1921981	<b>Description:</b> On Ubuntu, Debian and RedHat 8 and above OSS, parsing the mfa2 file using the mstarchive might result in a segmentation fault.
	<b>Workaround:</b> Use mlxarchive to parse the mfa2 file instead.
	<b>Keywords:</b> MFT, mfa2, mstarchive, mlxarchive, Ubuntu, Debian, RedHat, operating system
	<b>Discovered in Release:</b> 4.7-1.0.0.1

Internal Ref. Number	Issue
1840288	<b>Description:</b> MLNX_OFED does not support XDP features on RedHat 7 OS, despite the declared support by RedHat.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> XDP, RedHat
	<b>Discovered in Release:</b> 4.7-1.0.0.1
1821235	<b>Description:</b> 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.
	<b>Workaround:</b> Use the both actions at the end of the flow process.
	<b>Keywords:</b> Flow steering
	<b>Discovered in Release:</b> 4.7-1.0.0.1
1892663/180 0633	<b>Description:</b> mlnx_tune script does not support python3 interpreter.
	<b>Workaround:</b> Run mlnx_tune with python2 interpreter only.
	<b>Keywords:</b> mlnx_tune, python3, python2
	<b>Discovered in Release:</b> 4.7-1.0.0.1

Internal Ref. Number	Issue
15047 85	<b>Description:</b> A lost interrupt issue in pass-through virtual machines may prevent the driver from loading, followed by printing managed pages errors to the dmesg.
	<b>Workaround:</b> Restart the driver.
	<b>Keywords:</b> VM, virtual machine
	<b>Discovered in Release:</b> 4.6-1.0.1.1

Internal Ref. Number	Issue
17644 15	<p><b>Description:</b> Unbinding PFs on LAG devices results in a "Failed to modify QP to RESET" error message.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> RoCE LAG, unbind, PF, RDMA</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
	<p><b>Description:</b> 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.</p>
	<p><b>Workaround:</b> Use RoCEv2 default GID.</p>
	<p><b>Keywords:</b> RoCE</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
18065 65	<p><b>Description:</b> Aging is not functional on bond device in RHEL 7.6.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> VF LAG, ASAP<sup>2</sup></p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
	<p><b>Description:</b> 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.</p>
	<p><b>Workaround:</b> N/A</p>
	<p><b>Keywords:</b> VF LAG, ASAP<sup>2</sup></p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
17477 74	<p><b>Description:</b> 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:</p> <p>bond: link status down for interface eth1, disabling it in 100 ms</p>

Internal Ref. Number	Issue
	<p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> Bonding, slave</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
17120 68	<p><b>Description:</b> Uninstalling MLNX_OFED automatically results in the uninstallation of several libraries that are included in the MLNX_OFED package, such as InfiniBand-related libraries.</p> <p><b>Workaround:</b> If these libraries are required, reinstall them using the local package manager (yum/dnf).</p> <p><b>Keywords:</b> MLNX_OFED libraries</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
-	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> MFT compatible</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
17308 40	<p><b>Description:</b> On ConnectX-4 HCAs, GID index for RoCE v2 is inconsistent when toggling between enabled and disabled interface modes.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> RoCE v2, GID</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>
17174 28	<p><b>Description:</b> On kernels 4.10-4.14, MTUs larger than 1500 cannot be set for a GRE interface with any driver (IPv4 or IPv6).</p> <p><b>Workaround:</b> Upgrade your kernel to any version higher than v4.14.</p> <p><b>Keywords:</b> Fedora 27, gretap, ip_gre, ip_tunnel, ip6_gre, ip6_tunnel</p> <p><b>Discovered in Release:</b> 4.6-1.0.1.1</p>

Internal Ref. Number	Issue
17483 43	<b>Description:</b> Driver reload takes several minutes when a large number of VFs exists.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> VF, SR-IOV
	<b>Discovered in Release:</b> 4.6-1.0.1.1
17339 74	<b>Description:</b> Running heavy traffic (such as 'ping flood') while bringing up and down other mlx5 interfaces may result in "INFO: rcu_preempt detected stalls on CPUS/tasks:" call traces.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> mlx5
	<b>Discovered in Release:</b> 4.6-1.0.1.1
-	<b>Description:</b> On ConnectX-6 HCAs and above, an attempt to configure advertisement (any bitmap) will result in advertising the whole capabilities.
	<b>Workaround:</b> N/A
	<b>Keywords:</b> 200Gb/s, advertisement, Ethtool
	<b>Discovered in Release:</b> 4.6-1.0.1.1

Internal Ref. Number	Issue
1699289	<b>Description:</b> 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.
	<b>Workaround:</b> Make sure to enable HW LRO in the driver: <code>ethtool -k &lt;intf&gt; lro</code>

Internal Ref. Number	Issue
	<p>ethtool --set-priv-flag &lt;intf&gt; hw_lro on</p> <p><b>Keywords:</b> HW LRO, ConnectX-5 and above</p> <p><b>Discovered in Release:</b> 4.5-1.0.1.0</p>
1403313	<p><b>Description:</b> 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.</p> <p><b>Workaround:</b> Make sure to update the Kernel version to v4.15 or above.</p> <p><b>Keywords:</b> VF, PF, IOMMU, Kernel, OS</p> <p><b>Discovered in Release:</b> 4.5-1.0.1.0</p>
-	<p><b>Description:</b> NEO-Host is not supported on the following OSs:</p> <ul style="list-style-type: none"> <li>• SLES12 SP3</li> <li>• SLES12 SP4</li> <li>• SLES15</li> <li>• Fedora 28</li> <li>• RHEL7.1</li> <li>• RHEL7.4 ALT (Pegas1.0)</li> <li>• REL 7.5</li> <li>• RHEL7.6</li> <li>• XenServer 4.9</li> </ul> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> NEO-Host, operating systems</p> <p><b>Discovered in Release:</b> 4.5-1.0.1.0</p>
1521877	<p><b>Description:</b> On SLES 12 SP1 OSs, a kernel tracepoint issue may cause undefined behavior when inserting a kernel module with a wrong parameter.</p> <p><b>Workaround:</b> N/A</p> <p><b>Keywords:</b> mlx5 driver, SLES 12 SP1</p> <p><b>Discovered in Release:</b> 4.5-1.0.1.0</p>



---

# 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	<ul style="list-style-type: none"><li>• InfiniBand: SDR, FDR, EDR, HDR</li><li>• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup></li></ul>
BlueField		<ul style="list-style-type: none"><li>• InfiniBand: SDR, QDR, FDR, FDR10, EDR</li><li>• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE</li></ul>

<b>Uplink/Adapter Card</b>	<b>Driver Name</b>	<b>Uplink Speed</b>
ConnectX-7		<ul style="list-style-type: none"> <li>InfiniBand: EDR, HDR100, HDR, NDR200, NDR</li> <li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup>, 200GbE<sup>3</sup></li> </ul>
ConnectX-6 Lx		<ul style="list-style-type: none"> <li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup></li> </ul>
ConnectX-6 Dx		<ul style="list-style-type: none"> <li>Ethernet: 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup>, 200GbE<sup>2</sup></li> </ul>
ConnectX-6		<ul style="list-style-type: none"> <li>InfiniBand: SDR, FDR, EDR, HDR</li> <li>Ethernet: 10GbE, 25GbE, 40GbE, 50GbE<sup>2</sup>, 100GbE<sup>2</sup>, 200GbE<sup>2</sup></li> </ul>
ConnectX-5/ConnectX-5 Ex		<ul style="list-style-type: none"> <li>InfiniBand: SDR, QDR, FDR, FDR10, EDR</li> <li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE</li> </ul>
ConnectX-4 Lx		<ul style="list-style-type: none"> <li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE</li> </ul>
ConnectX-4		<ul style="list-style-type: none"> <li>InfiniBand: SDR, QDR, FDR, FDR10, EDR</li> <li>Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE<sup>1</sup>, 100GbE</li> </ul>

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.

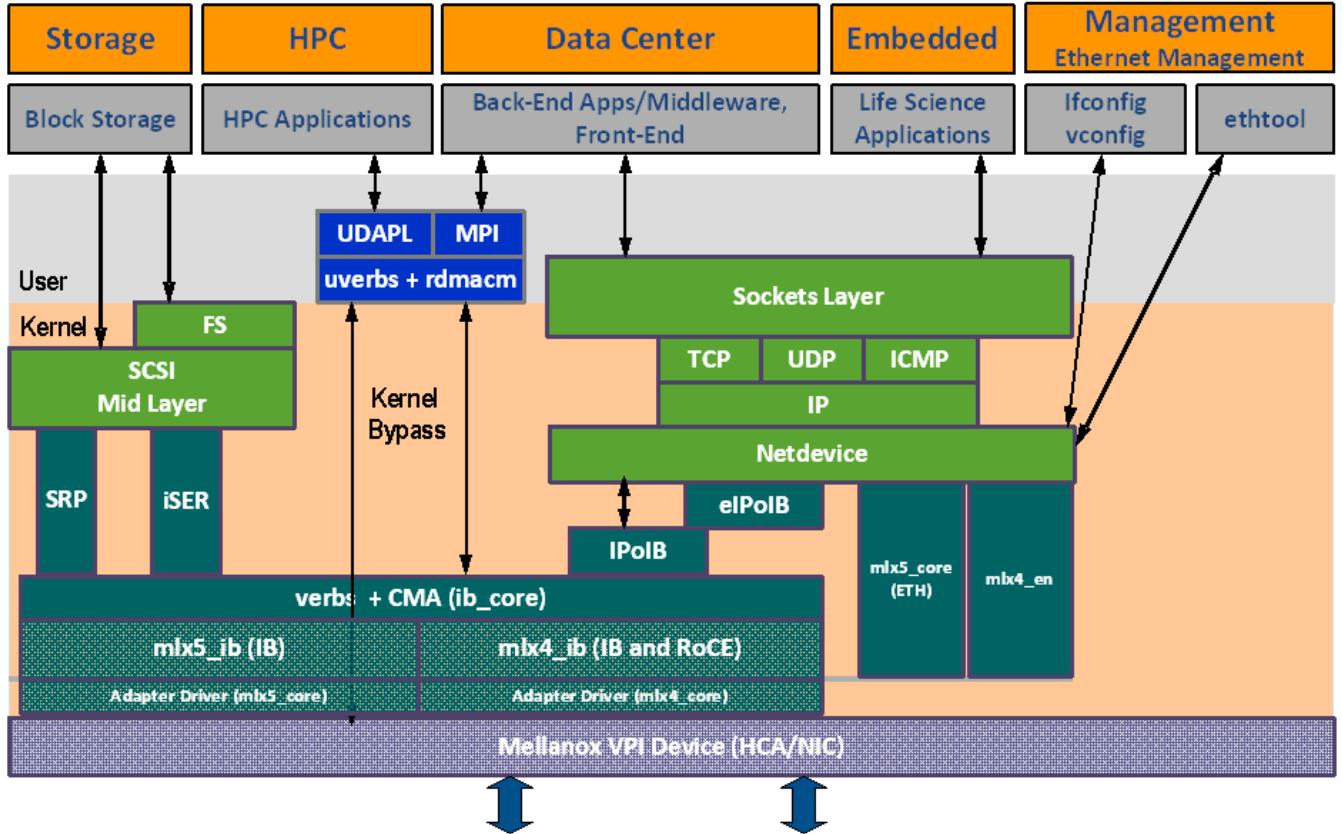
All NVIDIA network adapter cards are compatible with OpenFabrics-based RDMA protocols and software and are supported by major operating system distributions.

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.mellanox.com/category/vma](https://docs.mellanox.com/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 BeBenchmark and 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<sup>1</sup>.

1. OpenSM is disabled by default. See “[OpenSM](#)” 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](https://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.

<i>prof_sel</i>	<p>The parameter name for selecting the profile. The supported values for profiles are:</p> <ul style="list-style-type: none"> <li>• 0—for medium resources, medium performance</li> <li>• 1—for low resources</li> </ul>
-----------------	---

	<ul style="list-style-type: none"> <li>• 2—for high performance (int) (default)</li> </ul>
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_connected_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 ASAP<sup>2</sup> 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
<b>du</b> <b>mp</b> <b>_fts</b>	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
<b>iba ddr</b>	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.
<b>ibc ach eed it</b>	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.
<b>ibcc config</b>	Supports the configuration of congestion control settings on switches and HCAs. For further information, please refer to the tool's man page.
<b>ibcc query</b>	Supports the querying of settings and other information related to congestion control. For further information, please refer to the tool's man page.
<b>ibc ongest</b>	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.
<b>ibd ev2 net dev</b>	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.
<b>ibdi agn et (of ibu tils 2)</b>	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: <ul style="list-style-type: none"> <li>• Fabric discovery</li> <li>• Duplicated GUIDs detection</li> <li>• Links in INIT state and unresponsive links detection</li> <li>• Counters fetch</li> <li>• Error counters check</li> <li>• Routing checks</li> <li>• Link width and speed checks</li> </ul>

Utility	Description
	<ul style="list-style-type: none"> <li>• Alias GUIDs check</li> <li>• Subnet Manager check</li> <li>• Partition keys check</li> <li>• Nodes information</li> </ul> <p><b>Note:</b> 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:</p> <ol style="list-style-type: none"> <li>1. Run ibdiagnet -H</li> <li>Or</li> <li>2. Refer to <a href="https://docs.mellanox.com/display/ibdiagnetUserManualv10">https://docs.mellanox.com/display/ibdiagnetUserManualv10</a></li> </ol>
<b>ibdiagpath</b>	<p>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.</p> <p>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.</p> <p>Example: ibdiagpath --src_lid 1 --dest_lid 28</p> <p>For further information, please refer to the tool's -help flag.</p>
<b>ibdump</b>	<p>Dump InfiniBand traffic that flows to and from NVIDIA's ConnectX® family adapters InfiniBand ports.</p> <p>Note the following:</p> <ul style="list-style-type: none"> <li>• ibdump is not supported for Virtual functions (SR-IOV)</li> <li>• Infiniband traffic sniffing is supported on all HCAs</li> </ul> <p>The dump file can be loaded by the Wireshark tool for graphical traffic analysis. The following describes a workflow for local HCA (adapter) sniffing:</p> <ol style="list-style-type: none"> <li>1. Run ibdump with the desired options</li> <li>2. Run the application that you wish its traffic to be analyzed</li> <li>3. Stop ibdump (CTRL-C) or wait for the data buffer to fill (in --mem-mode)</li> </ol>

Utility	Description
	<p>4. Open Wireshark and load the generated file</p> <p>To download Wireshark for a Linux or Windows environment go to <a href="http://www.wireshark.org">www.wireshark.org</a>.</p> <p><b>Notes:</b></p> <ul style="list-style-type: none"> <li>• Although ibdump is a Linux application, the generated .pcap file may be analyzed on either operating system.</li> <li>• If one of the HCA's ports is configured as InfiniBand, ibdump requires IPoIB DMFS to be enabled. For further information, please refer to <a href="#">Flow Steering Configuration</a> section.</li> </ul> <p>For further information, please refer to the tool's man page.</p>
<b>iblinkinfo</b>	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.
<b>ibnetdiscover</b>	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).
<b>iscnover</b>	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.
<b>ibnethsplit</b>	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.
<b>ibnode</b>	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.
<b>ibping</b>	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.
<b>ibportstate</b>	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
	<ul style="list-style-type: none"> <li>• disable, enable or reset the port</li> <li>• validate the port's link width and speed against the peer port</li> </ul> <p>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:</p> <ul style="list-style-type: none"> <li>• The first ACTIVE port that is found.</li> <li>• If not found, the first port that is UP (physical link state is LinkUp).</li> </ul> <p>For further information, please refer to the tool's man page.</p>
<b>ib quer</b> <b>rors</b>	<p>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.</p> <p>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.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibr out</b> <b>e</b>	<p>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.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibstat</b>	<p>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.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibstatu</b> <b>s</b>	<p>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.</p>
<b>ibs wit</b> <b>ches</b>	<p>Traces the InfiniBand subnet topology or uses an already saved topology file to extract the InfiniBand switches.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibsys</b> <b>st</b> <b>at</b>	<p>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.</p>

Utility	Description
	For further information, please refer to the tool's man page.
<b>ibtopodiff</b>	<p>Compares a topology file and a discovered listing of subnet.lst/ibdiagnet.lst and reports mismatches.</p> <p>Two different algorithms provided:</p> <ul style="list-style-type: none"> <li>• Using the -e option is more suitable for MANY mismatches it applies less heuristics and provide details about the match</li> <li>• Providing the -s, -p and -g starts a detailed heuristics that should be used when only small number of changes are expected</li> </ul> <p>For further information, please refer to the tool's man page.</p>
<b>ibtrace rt</b>	<p>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.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibv_async_watch</b>	<p>Display asynchronous events forwarded to userspace for an InfiniBand device.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibv_devices</b>	<p>Lists InfiniBand devices available for use from userspace, including node GUIDs.</p> <p>For further information, please refer to the tool's man page.</p>
<b>ibv_dev_info</b>	<p>Queries InfiniBand devices and prints about them information that is available for use from userspace.</p> <p>For further information, please refer to the tool's man page.</p>
<b>mstflint</b>	<p>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.</p> <p>To run mstflint, you must know the device location on the PCI bus.</p> <p><b>Note:</b> If you purchased a standard NVIDIA network adapter card, please download the firmware image from <a href="http://nvidia.com/en-us/networking/">nvidia.com/en-us/networking/</a> Support Support <a href="#">Firmware Download</a>. If you purchased a non-standard card from a vendor other than NVIDIA, please contact your vendor.</p>

Utility	Description
	For further information, please refer to the tool's man page.
<b>perfquery</b>	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.
<b>saquer</b> <b>y</b>	Issues the selected SA query. Node records are queried by default. For further information, please refer to the tool's man page.
<b>smi</b> <b>nfo</b>	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. <b>Note:</b> 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.
<b>sm</b> <b>par</b> <b>que</b> <b>ry</b>	Sends SMP query for adaptive routing and private LFT features. For further information, please refer to the tool's man page.
<b>sm</b> <b>pdu</b> <b>mp</b>	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.
<b>sm</b> <b>pqu</b> <b>ery</b>	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
<b>ib_atom</b>	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
<b>ic_bw</b>	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.
<b>ib_atom_ic_lat</b>	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.
<b>ib_read_bw</b>	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.
<b>ib_read_lat</b>	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.
<b>ib_send_bw</b>	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.
<b>ib_send_lat</b>	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.
<b>ib_writ_e_b_w</b>	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.
<b>ib_writ_e_la_t</b>	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.
<b>raw_earn_et_bw</b>	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.
<b>raw_earn_et_lat</b>	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](mailto: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)

<b>Abbreviation/ Acronym</b>	<b>Description</b>
FW	Firmware
HCA	Host Channel Adapter
HW	Hardware
IB	InfiniBand
iSER	iSCSI RDMA Protocol
LSB	Least significant <i>byte</i>
lsb	Least significant <i>bit</i>
MSB	Most significant <i>byte</i>
msb	Most significant <i>bit</i>
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*.

<b>Term</b>	<b>Description</b>
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

<b>Term</b>	<b>Description</b>
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

<b>Document Name</b>	<b>Description</b>
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 PDF file relevant to your adapter device on <a href="http://mellanox.com">mellanox.com</a>
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

- [Release Notes Change Log History](#)
- [Bug Fixes History](#)

## User Manual Revision History

Release	Date	Description
5.7	August 2022	<ul style="list-style-type: none"><li>• Added <a href="#">Out of Order (OOO)</a> under RoCE section</li><li>• Added section describing <a href="#">Dumping Steering Info</a></li></ul>
5.6	April 2022	<ul style="list-style-type: none"><li>• Added <a href="#">OpenFlow Meters</a> section</li><li>• Added <a href="#">Installation on Community Operating Systems</a> section</li></ul>
5.5	December 2021	<ul style="list-style-type: none"><li>• Added "<a href="#">Forward to Multiple Destinations</a>" section</li><li>• Added "<a href="#">Open vSwitch Metering</a>" section</li><li>• Added "<a href="#">Multiport eSwitch Mode</a>" section</li><li>• Added "<a href="#">Bridge Offload</a>" section</li><li>• Added "<a href="#">TC Configuration for ConnectX-6 Dx and Above</a>" subsection</li></ul>
5.4-3	October 2021	<ul style="list-style-type: none"><li>• Removed <a href="#">LASH Routing Algorithm</a></li></ul>

Release	Date	Description
5.4.2	August 2021	<ul style="list-style-type: none"> <li>Added <a href="#">CT CT NAT</a> section</li> <li>Added <a href="#">Representor Metering</a> section</li> <li>Updated <a href="#">VF Metering</a> section</li> <li>Removed sysfs VXLAN portion of the <a href="#">Enabling VXLAN Hardware Stateless Offloads</a> section</li> </ul>
5.4	June 2021	<ul style="list-style-type: none"> <li>Updated <a href="#">SR-IOV Live Migration</a></li> </ul>
5.3.1	March 31,2021	<ul style="list-style-type: none"> <li>Added <a href="#">PTP Cyc2time Hardware Translation Offload</a> section</li> <li>Added <a href="#">Connection Tracking Performance Tuning</a> section</li> <li>Added <a href="#">VF Metering</a> section</li> <li>Added note under <a href="#">OVS-DPDK Hardware Offloads</a> section</li> <li>Updated <a href="#">Persistent Naming</a> section</li> <li>Updated command under <a href="#">Connection Tracking Offload</a> section</li> <li>Added <a href="#">sFlow</a> description under OVS-DPDK Hardware Offloads section</li> </ul>
5.2	February 14, 2021	<ul style="list-style-type: none"> <li>Added <a href="#">Setting up MLNX_OFED YUM Repository Using --add-kernel-support</a> section</li> <li>Added <a href="#">Setting up MLNX_OFED apt-get Repository Using --add-kernel-support</a> section</li> </ul>
	January 19, 2021	Added <a href="#">OpenSSL with kTLS Offload</a> section
	January 4, 2021	<ul style="list-style-type: none"> <li>Added <a href="#">Offloaded Traffic Sniffer</a> section</li> <li>Added <a href="#">Tx Port Time-Stamping</a> section</li> <li>Added <a href="#">VLAN Push/Pop</a> section</li> <li>Added <a href="#">sFLOW</a> section</li> <li>Added <a href="#">E2E Cache</a> section</li> <li>Added <a href="#">Geneve Encapsulation/Decapsulation</a> section</li> <li>Added <a href="#">Parallel Offloads</a> section</li> <li>Updated <a href="#">SR-IOV VF LAG</a> section</li> <li>Removed Installing MLNX_OFED on Innova™ IPsec Adapter Cards section</li> </ul>

Release	Date	Description
		<ul style="list-style-type: none"> <li>Removed Updating Firmware and FPGA Image on Innova IPsec Cards section</li> </ul>
5.1-2	September 17, 2020	Added <a href="#">Packet Pacing for Hairpin Queues</a> 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)
		Added <a href="#">SR-IOV Live Migration</a> section
		Added <a href="#">SR-IOV VF LAG</a> section
5.0-2	April 23, 2020	Added <a href="#">Interrupt Request (IRQ) Naming</a> section
	April 6, 2020	Added <a href="#">Kernel Transport Layer Security (kTLS) Offloads</a> section
5.0	March 3, 2020	<ul style="list-style-type: none"> <li>Added <a href="#">IPSec Crypto Offload</a> section</li> <li>Added <a href="#">OVS-DPDK Hardware Offloads</a> section</li> <li>Updated <a href="#">OVS Hardware Offloads Configuration</a> section</li> </ul>
4.7	December 29, 2019	<ul style="list-style-type: none"> <li>Added <a href="#">Configuring Uplink Represetor Mode</a> section</li> </ul>
	December 13, 2019	<ul style="list-style-type: none"> <li>Added <a href="#">Performance Tuning Based on Traffic Patterns</a> section</li> <li>Added "num_of_groups" entry to table <a href="#">mlx5_core Module Parameters</a></li> <li>Added <a href="#">Mediated Devices v5.4-0.5.1.1-Beta</a> section</li> </ul>
	September 29, 2019	<ul style="list-style-type: none"> <li>Updated <a href="#">Additional Installation Procedures</a> section</li> </ul>
4.6	May 13, 2019	<ul style="list-style-type: none"> <li>ethtool section updates: Added description of -f flashing option to <a href="#">Ethtool Supported Options</a> table</li> </ul>
	April 30, 2019	<ul style="list-style-type: none"> <li>ethtool section updates: <ul style="list-style-type: none"> <li>Updated the description of <a href="#">ethtool -s eth&lt;x&gt; advertise &lt;N&gt; autoneg on</a> counter under Ethtool</li> </ul> </li> </ul>

Release	Date	Description
		<ul style="list-style-type: none"> <li>○ Added the following counters under Ethtool:           <ul style="list-style-type: none"> <li>■ <a href="#">ethtool --show-fec eth&lt;x&gt;</a></li> <li>■ <a href="#">ethtool --set-fec eth&lt;x&gt; encoding auto off rs baser</a></li> </ul> </li> <li>● Added <a href="#">Devlink Parameters</a> section</li> <li>● Added <a href="#">Limit Bandwidth per Group of VFs</a> section</li> <li>● Added <a href="#">Disabling RoCE</a> section</li> <li>● Added <a href="#">RDMA-CM QP Timeout Control</a> section</li> <li>● Added <a href="#">RDMA-CM Application Managed QP</a> section</li> </ul>
4.5	December 19, 2018	<ul style="list-style-type: none"> <li>● Reorganized Chapter 2, “Installation”: Consolidated the separate installation procedures under <a href="#">Installing NVIDIA OFED</a> and <a href="#">Additional Installation Procedures</a></li> <li>● Added <a href="#">Installing NEO-Host Using mlnxofedinstall Script</a></li> </ul>
	November 29, 2018	<p>Added the following sections:</p> <ul style="list-style-type: none"> <li>● <a href="#">Local Loopback Disable</a></li> <li>● <a href="#">Offsweep Balancing</a></li> </ul>

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