Mellanox OFED for Linux Release Notes

Rev 2.1-1.0.0

Last Modified: 18 February, 2014

www.mellanox.com

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

Mellanox Technologies, Ltd.
Beit Mellanox
PO Box 586 Yokneam 20692
Israel
www.mellanox.com
Tel: +972 (0)74 723 7200
Fax: +972 (0)4 959 3245

Table of Contents

List Of Tables

1 Overview

These are the release notes of Mellanox OFED for Linux Driver, Rev 2.1-1.0.0. Mellanox OFED is a single Virtual Protocol Interconnect (VPI) software stack and operates across all Mellanox network adapter solutions supporting the following uplinks to servers:

- 10, 20, 40 and 56 Gb/s InfiniBand (IB)
- 10, 40 and 56[1] Gb/s Ethernet
- 2.5 or 5.0 GT/s PCI Express 2.0
- 8 GT/s PCI Express 3.0

2 Main Features in This Release

MLNX_OFED Rev 2.1-1.0.0 provides the following new features:

- Signature Verbs (T10-PI) (at beta level)
- RoCE Time Stamping
- PeerDirect
- Inline-Receive
- Ethernet Performance Counters
- Memory Window
- VMA bundled with MLNX_OFED
- DCT support (at beta level)
- eIPoIB multicast support

1. 56 GbE is a Mellanox propriety link speed and can be achieved while connected to Mellanox SX10XX switch series

3 Content of Mellanox OFED for Linux

Mellanox OFED for Linux software contains the following components:

Table 1 - Mellanox OFED for Linux Software Components

Components	Description
OpenFabrics core and ULPs	• IB HCA drivers (mlx4, mlx5) • core • Upper Layer Protocols: IPoIB, SRP and iSER Initiator
OpenFabrics utilities	• OpenSM: IB Subnet Manager with Mellanox proprietary Adaptive Routing • Diagnostic tools • Performance tests
MPI	• OSU MPI (mvapich2-1.9-1) stack supporting the InfiniBand interface • Open MPI stack 1.6.5 and later supporting the InfiniBand interface • MPI benchmark tests (OSU benchmarks, Intel MPI benchmarks, Presta)
PGAS	• ScalableSHMEM v2.2 supporting InfiniBand, MXM and FCA • ScalableUPC v2.2 supporting InfiniBand, MXM and FCA
HPC Acceleration packages	• Mellanox MXM v2.1 (p2p transport library acceleration over Infiniband) • Mellanox FCA v2.5 (MPI/PGAS collective operations acceleration library over InfiniBand) • KNEM, Linux kernel module enabling high-performance intra-node MPI/PGAS communication for large messages
Extra packages	• ibutils2 • ibdump • MFT
Sources of all software modules (under conditions mentioned in the modules' LICENSE files) except for MFT, OpenSM plugins, ibutils2, and ibdump	
Documentation	

4 Supported Platforms and Operating Systems

The following are the supported OSs in MLNX_OFED Rev 2.1-1.0.0:

Table 2 - Supported Platforms and Operating Systems

Operating System	Platform
RHEL/CentOS 6.3	x86_64 / PPC64
RHEL/CentOS 6.4	x86_64 / PPC64
RHEL/CentOS 6.5	x86_64
SLES11 SP1	x86_64
SLES11 SP2	x86_64 / PPC64
SLES11 SP3	x86_64 / PPC64
OEL 6.1	x86_64
OEL 6.2	x86_64
OEL 6.3	x86_64
OEL 6.4	x86_64
Citrix XenServer Host 6.x	i686
Fedora 18	x86_64
Fedora 19	x86_64
Ubuntu 12.04	x86_64
Ubuntu 13.04	x86_64
Ubuntu 13.10	x86_64
Debian 6.0.7	x86_64
Debian 6.0.8	x86_64
Debian 7.1	x86_64
Debian 7.2	x86_64
kernel 3.10	
kernel 3.11	
kernel 3.12	

If you wish to install OFED on a different kernel, you need to create a new ISO image, using mlnx_add_kernel_support.sh script.
See the MLNX_OFED User Guide for instructions.

Upgrading MLNX_OFED on your cluster requires upgrading all of its nodes to the newest version as well.

5 Hardware and Software Requirements

The following are the hardware and software requirements of MLNX_OFED Rev 2.1-1.0.0.

- 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:

Table 3 - Additional Software Packages

Operating System	Required Packages Installation Command
RHEL/OEL/Fedora	yum install pciutils python gcc-gfortran libxml2-python tcsh libnl.i686 libnl libnl-devel expat glib2 tcl libstdc++ bc tk
XenServer	yum install pciutils python libxml2-python libnl expat glib2 tcl bc libstdc++ tk
OpenSUSE	zypper install glib2-tools pciutils python libxml2-python tcsh libnl-1_1-32bit libstdc++46 expat libnl-1_1-devel libnl-1_1 tcl bc tk
SLES 11 SP1	zypper install pciutils python libxml2-python tcsh libnl libstdc++43 libnl-devel expat glib2 tcl bc libnl.i586 tk
SLES 11 SP2	zypper install pciutils python libnl-32bit libxml2-python tcsh libnl libnl-devel libstdc++46 expat glib2 tcl bc tk
SLES 11 SP3	zypper install pciutils python libnl-32bit libxml2-python tcsh libstdc++43 libnl libnl-devel expat glib2 tcl bc tk
Ubuntu/Debian	apt-get install dpkg autotools-dev autoconf libtool automake1.10 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

6 Supported HCAs

MLNX_OFED Rev 2.1-1.0.0 supports the following Mellanox network adapter cards:

- Connect-IB™ (Rev 10.10.2000 and above)
- ConnectX®-3 Pro (Rev 2.30.8000 and above)
- ConnectX®-3 (Rev 2.30.8000 and above)
- ConnectX®-2 (Rev 2.9.1000 and above)[1]

1. ConnectX®-2 does not support all the new functionality of MLNX_OFED 2.0.3-XXX. For the complete list of the supported features per HCA, please refer to the MLNX_OFED User Manual.

For official firmware versions please see:

http://www.mellanox.com/content/pages.php?pg=firmware_download

7 Compatibility

MLNX_OFED Rev 2.1-1.0.0 is compatible with the following:

Table 4 - MLNX_OFED Rev 2.1-1.0.0 Compatibility Matrix

Mellanox Product	Description/Version
SwitchX®	• InfiniBand - MSX6036, MSX6035, MSX6536 w/w MLNX-OS® version 3.3.3000 • Ethernet - MSX1036, MSX1016, MSX1024 w/w MLNX-OS® version 3.3.3000
FabricIT™ EFM	Tested IPoIB, Verbs and OpenSM priority handover • SLES 11 x64 w/w ConnectX VPI PCIe 2.0 5GT/s - IB QSFP QDR / 10GigE, ConnectX VPI - 10GigE / IB QDR • IS5030 w/w FabricIT EFM version 1.1.2700
FabricIT™ BXM	MBX5020 w/w FabricIT BXM version 2.1.2000
Unified Fabric Manager (UFM®)	v4.6
MXM	v2.1
ScalableUPC	v2.2
ScalableSHMEM	v2.2
FCA	v2.5
OMPI	v1.6.4
MVAPICH	v1.9a
CD	v1.0

8 Change Log History

8.1 Changes in Rev 2.1-1.0.0 From Rev 2.0-3.0.0

Table 5 - New Features, Changes and Fixes in v2.1-1.0.0

Category	Description
EoIB	EoIB is supported only in SLES11SP2 and RHEL6.4
Connect-IB™	Added the ability to resize CQs
IPoIB	Reusing DMA mapped SKB buffers: Performance improvements when IOMMU is enabled
mlnx_en	Added reporting autonegotiation support
	Added Transmit Packet Steering (XPS) support
	Added reporting 56Gbit/s link speed support
	Added Receive Flow Steering (RFS) support in UDP
	Added Low Latency Socket (LLS) support
	Added check for dma_mapping errors
eIPoIB	Added non-virtual environment support
Hypervisor support	KVM and XenServer

8.2 Changes in Rev 2.0-3.0.0 From Rev 2.0-2.0.5

Table 6 - New Features, Changes and Fixes in v2.0-3.0.0 (Sheet 1 of 2)

Category	Description
Operating Systems	Additional OS support: • SLES11SP3 • Fedora16, Fedora17
Drivers	Added Connect-IB™ support
Installation	Added ability to install MLNX_OFED with SR-IOV support.
	Added Yum installation support
EoIB	EoIB (at beta level) is supported only in SLES11SP2 and RHEL6.4
mlx4_core	Modified module parameters to associate configuration values with specific PCI devices identified by their bus/device/function value format
mlx4_en	Reusing DMA mapped buffers: major performance improvements when IOMMU is enabled
	Added Port level QoS support

Table 6 - New Features, Changes and Fixes in v2.0-3.0.0 (Sheet 2 of 2)

Category	Description
IPoIB	Reduced memory consumption
	Limited the number TX and RX queues to 16
	Default IPoIB mode is set to work in Datagram, except for Connect-IB™ adapter card which uses IPoIB with Connected mode as default.
Storage	iSER (at GA level)

8.3 New Features in Rev 2.0-2.0.5[1]

- SR-IOV for both Ethernet and InfiniBand (at Beta level)
- RoCE over SR-IOV (at Beta level)
- eIPoIB to enable IPoIB in a Para-Virtualized environment (at Alpha level)
- Contiguous pages:
  - Internal memory allocation improvements
  - Register shared memory
  - Control objects (QPs, CQs)
- Ethernet Performance Enhancements (NUMA related and others) for 10G and 40G
- OFED_VMA integration to a single branch
- Ethernet Time Stamping (at Beta level)
- Flow Steering for Ethernet and InfiniBand. (at Beta level)
- Raw Eth QPs:
  - Checksum TX/RX
  - Flow Steering
- Errata Kernel upgrade support
- YUM update support
- Storage – iSER (at Beta level) and SRP
- 64bit wide counters (port xmit/recv data/packets unicast/mcast)
- VERSION query API: library and headers

1. SR-IOV, Ethernet Time Stamping and Flow Steering are ConnectX®-3 HCA capability.

9 Known Issues

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

Table 7 - Known Issues

Index	Issue	Description	Workaround
1.	IPoIB	When user increases receive/send a buffer, it might consume all the memory when few child's interfaces are created.	-
2.		The hardware address suffix of IPoIB interfaces in MLNX_OFED v2.0-3.0.0 is `'a'` instead of `'8'` to indicate the TSS support.	-
3.		The size of send queue in Connect-IB™ cards cannot exceed 1K.	-
4.		In 32 bit devices, the maximum number of child interfaces that can be created is 16. Creating more that, might cause out-of-memory issues.	-
5.		The default IPoIB operating mode in ConnectX® family adapter cards is UD and CM in Connect-IB™.	-
6.		Changing the IPoIB mode (CM vs UD) requires the interface to be in `'down'` state.	-
7.		IPoIB interface does not function properly if a third party application changes the PKey table. We recommend modifying PKey tables via OpenSM.	-
8.		When creating a new child interface in an over-loaded kernel, a `dmesg` print is displayed advising the user to try again in a few seconds.	-
9.		Out-of memory issue might occur due to overload of interfaces created.	To calculate the allowed memory per each IPoIB interface check the following: • Num-rings = min(num-cores-on-that-device, 16) • Ring-size = 512 (by default, it is module parameter) • UD memory: 2 * num-rings * ring-size * 8K • CM memory: ring-size * 64k • Total memory = UD mem + CM mem

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
10.		The physical port MTU (indicates the port capability) default value was changed to 4k, whereas the IPoIB port MTU ("logical" MTU) default value is 2k as it is set by the OpenSM.	In order to change the IPoIB MTU to 4k, edit the OpenSM partition file in the section of IPoIB setting as follow: `Default=0xffff, ipoib, mtu=5 : ALL=full;` *Where `"mtu=5"` indicates that all IPoIB ports in the fabric are using 4k MTU, (`"mtu=4"` indicates 2k MTU)
11.		Occasionally, when using IPoIB in Connected mode, the connection might get closed and recovered only after several minutes.	Use the Datagram mode
12.		Fallback to the primary slave of an IPoIB bond does not work with ARP monitoring. (https://bugs.openfabrics.org/show_bug.cgi?id=1990)	-
13.		Whenever the IOMMU parameter is enabled in the kernel it can decrease the number of child interfaces on the device according to resource limitation. The driver will stuck after unknown amount of child interfaces creation.	To avoid such issue: • Decrease the amount of the RX receive buffers (module parameter, the default is 512) • Decrease the number of RX rings (sys/fs or ethtool in new kernels) • Avoid using IOMMU if not required

14.		System might crash in `skb_checksum_help()` while performing TCP retransmit involving packets with 64k packet size.
		A similar out to the below will be printed:
		`kernel BUG at net/core/dev.c:1707!`
		`invalid opcode: 0000 [#1] SMP`
		`RIP: 0010:[<ffffffff81448988>]`
		`skb_checksum_help+0x148/0x160`
		`Call Trace:`
		`<IRQ>`
		`[<ffffffff81448d83>]`
		`dev_hard_start_xmit+0x3e3/0x530`
		`[<ffffffff8144c805>]`
		`dev_queue_xmit+0x205/0x550`
		`[<ffffffff8145247d>]`
		`neigh_connected_output+0xbd/0x1`
		…
		…
			Use UD mode in ipoib

15.		Changing the GUID of a specific SR-IOV guest after the driver has been started, causes the ping to fail. Hence, no traffic can go over that InfiniBand interface.	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
16.		`send_queue_size` over Connect-IB™ adapter cards cannot be larger than 1024	-
17.	Ethernet	Ethernet PV VLAN Guest transparent Tagging (VGT) is only supported in openvswitch and not in standard Linux vBridges and libvirt For more information please see : http://libvirt.org/formatnetwork.html (Setting VLAN tag section)	-

18.		Changing the ring size on 32-bit system may result in failure due to lack of memory. Therefore, mlx4_en will not be able to vmap enough memory and the below message will be printed in dmesg:
		`vmap allocation for size 528384 failed: use vmalloc=<size> to increase size`
		`In this case user can enlarge the vmalloc memory by adding vmalloc=<size> to grub.conf`
		`Default vmalloc setting is 128M. It is recommended to add each time 64M of memory until desired ring size can be allocated.`
		Please note, that in case vmalloc size is too big, the OS will fail to boot, so please use caution when adding additional memory.
		For more info refer to: http://www.mythtv.org/wiki/Common_Problem:_vmalloc_too_small
			-

19.		On OEL6.1 with uek1 (2.6.32-x.x.x.el6uek kernel), when the number of RX ring is smaller than TX rings (kernel issue), the following call trace will be shown in the kernel log: `WARNING: at net/core/dev.c:2077 get_rps_cpu+0x70/0x2b9()`	-
20.		Kernel panic might occur during traffic over IPv6 on kernels between 3.12-rc7 and 3.13-rc1 (kernel issue)	-
21.		Kernel panic might occur during fio splice in kernels before 2.6.34-rc4.	Use kernel v2.6.34-rc4 which provides the following solution: `baff42a net: Fix oops from tcp_collapse() when using splice()`
22.		On Debian-6.0.7, kernel panic may occur when changing the number of TX channels above the default value (8).	-
23.		On kernels that do not support multiqueues, the number of TX channels represents the number of TX rings. The maximal number of TX channels is 16.	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
24.		Transmit timeout might occur on RH6.3 as a result of lost interrupt (OS issue). In this case, the following message will be shown in dmesg: `do_IRQ: 0.203 No irq handler for vector (irq -1)`	-
25.	eIPoIB	On rare occasions, upon driver restart the following message is shown in the dmesg: `'cannot create duplicate filename '/class/net/eth_ipoib_interfaces'`	-
26.		No indication is received when eIPoIB is non functional.	Run `'ps -ef | grep ipoibd'` to verify its functionality.
27.		eIPoIB requires libvirtd, python	-
28.		eIPoIB supports only active-backup mode for bonding.	-
29.		eIPoIB supports only VLAN Switch Tagging (VST) mode on guests.	-
30.		IPv6 is currently not supported in eIPoIB	-
31.	XRC	Legacy API is deprecated, thus when recompiling applications over MLNX_OFED v2.0-3.x.x, warnings such as the below are displayed. `rdma.c:1699: warning: 'ibv_open_xrc_domain' is deprecated (declared at /usr/include/infiniband/ofa_verbs.h:72)` `rdma.c:1706: warning: 'ibv_create_xrc_srq' is deprecated (declared at /usr/include/infiniband/ofa_verbs.h:89)` These warnings can be safely ignored.	-
32.		XRC is not functional in heterogeneous clusters containing non Mellanox HCAs.	-
33.		XRC options do not work when using qperf tool.	Use perftest instead
34.		XRC over ROCE in SR-IOV mode is not functional	-
35.		Out-of memory issue might occur due to overload of XRC receive QP with non zero receive queue size created. XRC QPs do not have receive queues.	-
36.	mlx4_ib module	The `dev_assign_str` module parameter is not backward compatible. In the current version, this parameter is using decimal number to describe the InfiniBand device and not hexadecimal number as it was in previous versions in order to uniform the mapping of device function numbers to InfiniBand device numbers as defined for other module parameters (e.g. `num_vfs` and `probe_vf`).	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
37.	ABI Compatibility	MLNX_OFED Rev 2.1-1.0.0 is not ABI compatible with previous MLNX_OFED/OFED versions.	Recompile the application over the new MLNX_OFED version
38.	System Time	Loading the driver using the openibd script when no InfiniBand vendor module is selected (for example mlx4_ib), may cause the execution of the /sbin/start_udev' script. In RedHat 6.x and OEL6.x this may change the local system time.	-
39.	Verbs	Verbs for the following features are subject to change: • Core-Direct • Shared memory region • Contiguous pages • Flow steering Verbs subject to changes are: • ibv_post_task • ibv_query_values_ex • ibv_query_device_ex • ibv_poll_cq_ex • ibv_reg_shared_mr_ex • ibv_reg_shared_mr • ibv_modify_cq • ibv_create_cq_ex • ibv_modify_qp_ex • ibv_reg_mr • ibv_post_send • ibv_dealloc_mw, • ibv_alloc_mw, • ibv_bind_mw • ibv_query_device • ibv_poll_cq • ibv_create_qp_ex • ibv_modify_qp	-
40.		Using libnl1_1_3~26 or earlier, requires ibv_create_ah protection by a lock for multi-threaded applications.	-
41.	Driver Start	When reloading the driver using the `"/etc/init.d/openibd restart"` command on XenServer6.1, loading of mlx4_en driver might fail with "Unresolved Symbols" errors. This message can safely be ignored.	-
42.		"Out of memory" issues may rise during drivers load depending on the values of the driver module parameters set (e.g. log_num_cq).	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
43.		When reloading/starting the driver using the `/etc/init.d/openibd` the following messages are displayed if there is a third party RPM or driver installed: "Module mlx4_core does not belong to MLNX_OFED" or "Module mlx4_core belong to <rpm name> which is not a part of MLNX_OFED"	Remove the third party RPM/non MLNX_OFED drivers directory, run: `"depmod"` and then rerun `"/etc/init.d/openibd restart"`
44.		Occasionally, when trying to repetitively reload the nes hardware driver on SLES11 SP2, a soft lockups occurs that required reboot.	-
45.		In ConnectX-2, if the driver load succeeds, the informative message below is presented conveying the below limitations: • If port type is IB the number of maximum supported VLs is 4 • If port type is ETH then the maximum priority for VLAN tagged is 3 `"mlx4_core 0000:0d:00.0: command SET_PORT (0xc) failed: in_param=0x120064000, in_mod=0x2, op_mod=0x0, fw status = 0x40"`	
46.	Operating Systems	RHEL 5.X and SLES 10 SPX are currently not supported.	-
47.	SR-IOV	When using legacy VMs with OFED 2.0-2.0.5 hypervisor, the `'enable_64b_cqe_eqe'` parameter must be set to zero on the hypervisor. It should be set in the same way that other module parameters are set for mlx4_core at module load time. For example, add `"options mlx4_core enable_64b_cqe_eqe=0"` as a line in the file `/etc/modprobe.d/mlx4_core.conf`.	-
48.		Enabling SR-IOV requires appending the `"intel_iommu=on"` option to the relevant OS in file `/boot/grub/grub.conf/`. Without that SR-IOV cannot be loaded.	-
49.		rdma_cm does not support UD QPs	-
50.		SR-IOV can be enabled only when using the firmware version embedded in the MLNX_OFED v2.0-3.0.0 driver.	-
51.		When SR-IOV is disabled in the system BIOS, a PCI issue is noticed in Ubuntu v12.04.3 with Linux kernel v3.8 which affects NICs of several manufacturers including Mellanox's, preventing them from operating.	Enable Sr-IOV in the BIOS

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
52.	Port Type Management	OpenSM must be stopped prior to changing the port protocol from InfiniBand to Ethernet.	-
53.		After changing port type using `connectx_port_config` interface ports' names can be changed. For example. ib1 -> ib0 if port1 changed to be Ethernet port and port2 left IB.	Use `udev` rules for persistent naming configuration. For further information, please refer to the User Manual
54.	Flow Steering	Flow Steering is disabled by default.	To enable it, set the parameter below as follow: `log_num_mgm_entry_size should set to -1`
55.		IPv4 rule with source IP cannot be created in SLES 11	-
56.		RFS is not supported in SLES11	-
57.	Quality of Service	QoS is not supported in XenServer, Debian 6.0 and in OEL6.1 and 6.2 with uek kernel	-
58.	Driver Uninstall	A Kernel panic occurs if you uninstall the driver without deleting the SR-IOV module params (`mlx4_core's num_vfs`) in the file `/etc/modprobe.d/mlx4_core.conf`. On the next boot, you will get the panic, and machine will boot up.	Remove the midule after uninstalling and prior to restarting the driver.
59.	Installation	When upgrading from an earlier Mellanox OFED version, the installation script does not stop the earlier version prior to uninstalling it.	Stop the old OFED stack (`/etc/init.d/openibd stop`) before upgrading to this new version.
60.		Upgrading from the previous OFED installation to this release, does not unload the kernel module ipoib_helper.	Reboot after installing the driver.
61.		Installation using Yum does not update HCA firmware.	See "Updating Firmware After Installation" in OFED User Manual
62.		On SLES11.1 the package 'libnl.i586' is required to install MLNX_OFED.	Perform one of the following: • Install the 'libnl.i586' RPM from the SLES11.1 32bit installation disk • Install MLNX_OFED with the following flag "--without-32bit"
63.		When using bonding on Ubuntu OS, the "ifenslave" package must be installed.	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
64.		`"--total-vfs <0-63>"` installation parameter is no longer supported	Use `'--enable-sriov'` installation parameter to burn firmware with SR-IOV support. The number of virtual functions (VFs) will be set to 16. For further information, please refer to the User Manual.

65.	Driver Unload	`"openibd stop"` can sometime fail with the error:
		`Unloading ib_cm [FAILED]`
		`ERROR: Module ib_cm is in use by ib_ipoib`
			Re-run `"openibd stop"`

66.	Fork Support	Fork support from kernel 2.6.12 and above is available provided that applications do not use threads. `fork()` is supported as long as the parent process does not run before the child exits or calls `exec()`. The former can be achieved by calling `wait(childpid)`, and the latter can be achieved by application specific means. The Posix system() call is supported.	-
67.	ISCSI over IPoIB	When working with ISCSI over IPoIB, LRO must be disabled (even if IPoIB is set to connected mode) due to a a bug in older kernels which causes a kernel panic.	-
68.	MLNX_OFED sources	MLNX_OFED includes the OFED source RPM packages used as a build platform for kernel code but does not include the sources of Mellanox proprietary packages.	-
69.	InfiniBand Utilities	When running the ibdiagnet check `nodes_info` on the fabric, a warning specifying that the card does not support general info capabilities for all the HCAs in the fabric will be displayed.	Run `ibdiagnet --skip nodes_info`
70.	mlx5 Driver	Atomic Operations over Connect-IB™ are not supported.	-
71.	General	On ConnectX-2/ConnectX-3 Ethernet adapter cards, there is a mismatch between the GUID value returned by firmware management tools and that returned by fabric/driver utilities that read the GUID via device firmware (e.g., using ibstat). Mlxburn/flint return 0xffff as GUID while the utilities return a value derived from the MAC address. For all driver/firmware/software purposes, the latter value should be used.	N/A. Please use the GUID value returned by the fabric/driver utilities (not 0xffffff).
72.	Uplinks	On rare occasions, ConnectX®-3 Pro adapter card may fail to link up when performing parallel detect to 40GbE.	Restart the driver

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
73.	Resources Limitation	The device capabilities reported may not be reached as it depends on the system on which the device is installed and whether the resource is allocated in the kernel or the userspace.	-

74.		Occasionally, a user process might experience some memory shortage and not function properly due to Linux kernel occupation of the system's free memory for its internal cache.	To free memory to allow it to be allocated in a user process, run the `drop_caches` procedure below.
			Performing the following steps will cause the kernel to flush and free pages, dentries and inodes caches from memory, causing that memory to become free.
			
			Note: As this is a non-destructive operation and dirty objects are not freeable, run `` `sync' `` first.
			
			• To free the pagecache:
			`echo 1 > /proc/sys/vm/drop_caches`
			• To free dentries and inodes:
			`echo 2 > /proc/sys/vm/drop_caches`
			• To free pagecache, dentries and inodes:
			`echo 3 > /proc/sys/vm/drop_caches`

75.		Setting more IP addresses than the available GID entries in the table results in failure and the "update_gid_table error message is displayed: `GID table of port 1 is full. Can't add <address>" message.`	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
76.	Ethernet Performance Counters	In a system with more than 61 VFs, the 62nd VF and onwards is assigned with the SINKQP counter, and as a result will have no statistics, and loopback prevention functionality for SINK counter.	-
77.		Since each VF tries to allocate 2 more QP counter for its RoCE traffic statistics, in a system with less than 61 VFs, if there is free resources it receives new counter otherwise receives the default counter which is shared with Ethernet. In this case RoCE statistics is not available.	-
78.		In ConnectX®-3, when we enable function-based loopback prevention for Ethernet port by default (i.e., based on the QP counter index), the dropped self-loopback packets increase the IfRxErrorFrames/Octets counters.	-
79.	RoCE	Not configuring the Ethernet devices or independent VMs with a unique IP address in the physical port, may result in RoCE GID table corruption.	Restart the driver
80.		If RDMA_CM is not used for connection management, then the source and destination GIDs used to modify a QP or create AH should be of the same type - IPv4 or IPv6.	-
81.		Since the number of GIDs per port is limited to 128, there cannot be more than the allowed IP addresses configured to Ethernet devices that are associated with the port. Allowed number is: • "127" for a single function machine • "15" for a hypervisor in a multifunction machine • "n" for a guest in a multifunction machine (where n is the number of virtual functions)	-
82.		A working IP connectivity between the RoCE devices is required when creating an address handle or modifying a QP with an address vector.	-
83.		MLNX_OFED v2.1-1.0.0 is not interoperable with older versions of MLNX_OFED.	-
84.		Unloading mlx4_en while a rdma_cm session is established can cause a kernel panic.	-
85.	Storage	SLES11-SP1: When running multipath rescan while new devices are added to mpath tables, multipath may not find all the device-mappers.	-
86.		Older versions of rescan_scsi_bus.sh may not recognize some newly created LUNs.	If encountering such issues, it is recommended to use the '-c' flag.
87.	SRP	Reconnecting to a target during host reset stage may result in devices going Offline.	Run `rescan-scsi-bus.sh -r`
88.	SRP Interop	The driver is tested with Storage target vendors recommendations for multipath.conf extensions (ZFS, DDN, TMS, Nimbus, NetApp).	-

Table 7 - Known Issues (Continued)

Index	Issue	Description	Workaround
89.	DDN Storage Fusion 10000 target	DDN does not accept non-default P_Key connection establishment.	-
90.	Oracle Sun ZFS storage 7420	Occasionally the first command to a LUN may not be serviced, aborted, and cause a successful re-connection to the target	-
91.		ZFS does not accept non-default P_Key connection establishment.	-
92.		Ungraceful power cycle of an initiator connected with Targets DDN, Nimbus, NetApp may result in temporary "stale connection" messages when initiator reconnects.	-
93.	iSER	On SLES11, the `ib_iser` module does not load on boot	Add a dummy interface using iscsiadm: • `# iscsiadm -m iface -I ib_iser -o new` • `# iscsiadm -m iface -I ib_iser -o update -n iface.transport_name -v ib_iser`
		In SLES10 SP3 and Ubuntu12.04 need to update user space open-iscsi package to version 2.0.873.	-
		Trying to disconnect a session while the session is undergoing a reconnect flow may result in disconnection hang.	Restart iscsid. **Note:** Please be aware that doing so might cause `rmmod` process to hang as the ib_iser module will not be unloaded.
		Unloading ib_iser during session disconnect event may result in kernel panic.	-
	iSER interop - Oracle Sun ZFS storage 7420	Connection establishment occurs twice which may cause iSER to log a stack trace	-

10 API Changes

10.1 API Changes in MLNX_OFED Rev 2.1-1.0.0

The following are the API changes in MLNX_OFED Rev 2.1-1.0.0:

Table 8 - API Changes in MLNX_OFED Rev 2.0-3.0.0

Name	Description
Dynamically Connected (DC)	The following verbs were added: • `struct ibv_dct *ibv_exp_create_dct(struct ibv_context *context, struct ibv_exp_dct_init_attr *attr)` • `int ibv_exp_destroy_dct(struct ibv_dct *dct)` • `int ibv_exp_query_dct(struct ibv_dct *dct, struct ibv_exp_dct_attr *attr)`

10.1.1 Verbs Extension and Verbs Experimental APIs

- Verbs Extension API

  Verbs extension API defines OFA APIs extension scheme to detect ABI compatibility and enable backward and forward compatibility support.

- Verbs Experimental API

  Verbs experimental API defines MLNX-OFED APIs extension scheme which is similar to the “Verbs extension API”. This extension provides a way to introduce new features before they are integrated into the formal OFA API and to the upstream kernel and libs.

The following are the Verbs Extension and Verbs Experimental APIs in MLNX_OFED Rev 2.1-1.0.0:

Table 9 - Verbs Extension and Verbs Experimental APIs

API Type	APIs
Verbs Extension API	• ibv_post_task • ibv_query_values_ex • ibv_query_device_ex • ibv_create_flow • ibv_destroy_flow • ibv_poll_cq_ex • ibv_reg_shared_mr_ex • ibv_open_xrcd • ibv_close_xrcd • ibv_modify_cq • ibv_create_srq_ex • ibv_get_srq_num • ibv_create_qp_ex • ibv_create_cq_ex • ibv_open_qp • ibv_modify_qp_ex

Table 9 - Verbs Extension and Verbs Experimental APIs

API Type	APIs
Verbs Experimental API	• ibv_exp_create_qp • ibv_exp_query_device • ibv_exp_create_dct • ibv_exp_destroy_dct • ibv_exp_query_dct

10.2 API Changes in MLNX_OFED Rev 2.0-3.0.0

The following are the API changes in MLNX_OFED Rev 2.0-3.0.0:

Table 10 - API Changes in MLNX_OFED Rev 2.0-3.0.0

Name	Description
XRC	The following verbs have become deprecated: • `struct ibv_xrc_domain *ibv_open_xrc_domain` • `struct ibv_srq *ibv_create_xrc_srq` • `int ibv_close_xrc_domain` • `int ibv_create_xrc_rcv_qp` • `int ibv_modify_xrc_rcv_qp` • `int ibv_query_xrc_rcv_qp` • `int ibv_reg_xrc_rcv_qp` • `int ibv_unreg_xrc_rcv_qp`

10.3 API Changes in MLNX_OFED Rev 2.0-2.0.5

The following are the API changes in MLNX_OFED v2.0-2.0.5:

Table 11 - API Changes in MLNX_OFED Rev 2.0-2.0.5

Name	Description
Libibverbs	
Extended speeds	• Missing the `ext_active_speed attribute` from the struct `ibv_port_attr` • Removed function `ibv_ext_rate_to_int` • Added functions `ibv_rate_to_mbps` and `mbps_to_ibv_rate`
Raw QPs	QP types `IBV_QPT_RAW_PACKET` and `IBV_QPT_RAW_ETH` are not supported
Contiguous pages	• Added Contiguous pages support • Added function `ibv_reg_shared_mr`
Libmverbs	

Table 11 - API Changes in MLNX_OFED Rev 2.0-2.0.5

Name	Description

	• The enumeration `IBV_M_WR_CALC` was renamed to `IBV_M_WR_CALC_SEND`
	• The enumeration `IBV_M_WR_WRITE_WITH_IMM` was added
	• In the structure `ibv_m_send_wr`, the union `wr.send` was renamed to `wr.calc_send` and `wr.rdma` was added
	• The following enumerations were renamed:
	  • From `IBV_M_WQE_SQ_ENABLE_CAP` to `IBV_M_WQE_CAP_SQ_ENABLE`
	  • From `IBV_M_WQE_RQ_ENABLE_CAP` to `IBV_M_WQE_CAP_RQ_ENABLE`
	  • From `IBV_M_WQE_CQE_WAIT_CAP` to `IBV_M_WQE_CAP_CQE_WAIT`
	  • From `IBV_M_WQE_CALC_CAP` to `IBV_M_WQE_CAP_CALC_SEND`
	• The enumerations `IBV_M_WQE_CAP_CALC_RDMA_WRITE_WITH_IMM` was added

11 Bug Fixes History

Table 12 lists the bugs fixed in this release.

Table 12 - Fixed Bugs List

#	Issue	Description	Discovered in Release	Fixed in Release
1.	mlx4_core	Restored port types as they were when recovering from an internal error.	2.0-2.0.5	2.1-1.0.0
2.		Added an N/A port type to support port_type_array module param in an HCA with a single port	2.0-2.0.5	2.1-1.0.0
3.	SR-IOV	Fixed memory leak in SR-IOV flow.	2.0-2.0.5	2.0-3.0.0
4.		Fixed communication channel being stuck	2.0-2.0.5	2.0-3.0.0
5.	mlx4_en	Fixed ALB bonding mode failure when enslaving Mellanox interfaces	2.0-3.0.0	2.1-1.0.0
6.		Fixed leak of mapped memory	2.0-3.0.0	2.1-1.0.0
7.		Fixed TX timeout in Ethernet driver.	2.0-2.0.5	2.0-3.0.0
8.		Fixed ethtool stats report for Virtual Functions.	2.0-2.0.5	2.0-3.0.0
9.		Fixed an issue of VLAN traffic over Virtual Machine in paravirtualized mode.	2.0-2.0.5	2.0-3.0.0
10.		Fixed ethtool operation crash while interface down.	2.0-2.0.5	2.0-3.0.0
11.	IPoIB	Fixed memory leak in Connected mode.	2.0-2.0.5	2.0-3.0.0
12.		Fixed an issue causing IPoIB to avoid pkey value 0 for child interfaces.	2.0-2.0.5	2.0-3.0.0