Mellanox InfiniBand OFED Driver for VMware vSphere 5.1 and 5.5 User Manual

Rev 1.8.2.4

www.mellanox.com

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT (“PRODUCT(S)”) AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCTO(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

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

Table of Contents

List of Tables

Document Revision History

Table 1 - Document Revision History

Revision	Date	Change Description
1.8.2.4	March 2014	Added the following section: • Section 2.3, "Loading/Unloading Driver Kernel Modules", on page 12 Updated the following sections: • Section 1.3, "Mellanox OFED Package", on page 10 • Section 2.1, "Installing Mellanox InfiniBand OFED Driver for VMware vSphere", on page 11 • Section 2.2, "Removing Mellanox InfiniBand OFED Driver", on page 11 • Section 3.2.2, "IPoIB Configuration", on page 15 • Section 4.1, "Configuring an Uplink", on page 16 • Section 4.2, "Configuring VMware ESXi Server Settings", on page 16 Removed section "Supported Hardware Compatibility"
1.8.2	September 2013	No changes
1.8.1	February 2013	Added Section 3.1, "SCSI RDMA Protocol", on page 13 and its subsections.
1.8.0	June 2012	Document restructuring and minor content updates
1.4.1-2.0.000	May 2011	Initial release

About this Manual

This document provides instructions for installing and using drivers for Mellanox Technologies ConnectX®-2, ConnectX®-3 and ConnectX®-3 Pro based network adapter cards in a VMware ESXi Server environment.

Intended Audience

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

Documentation Conventions

Table 2 - Documentation Conventions

Description	Convention	Example
File names	file.extension	
Directory names	directory	
Commands and their parameters	command param1	mts3610-1 > show hosts
Required item	< >	
Optional item	[ ]	
Mutually exclusive parameters	{ p1, p2, p3 } or {p1 | p2 | p3}	
Optional mutually exclusive parameters	[p1 | p2 | p3]	
Variables for which users supply specific values	Italic font	*enable*
Emphasized words	Italic font	*These are emphasized words*
Note	 <text>	 This is a note.
Warning	 <text>	 May result in system instability.

Common Abbreviations and Acronyms

Table 3 - Abbreviations and Acronyms

Abbreviation / Acronym	Whole Word / Description
B	(Capital) 'B' is used to indicate size in bytes or multiples of bytes (e.g., 1KB = 1024 bytes, and 1MB = 1048576 bytes)
b	(Small) 'b' is used to indicate size in bits or multiples of bits (e.g., 1Kb = 1024 bits)
FW	Firmware
HCA	Host Channel Adapter
HW	Hardware
IB	InfiniBand
LSB	Least significant *byte*
lsb	Least significant *bit*
MSB	Most significant *byte*
msb	Most significant bit
NIC	Network Interface Card
SW	Software
IPoIB	IP over InfiniBand
PFC	Priority Flow Control
PR	Path Record
SL	Service Level
SRP	SCSI RDMA Protocol
ULP	Upper Level Protocol
VL	Virtual Lanes

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*.

Table 4 - Glossary

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	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.
LID	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. See Subnet Manager.
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 agency of the master SM. See Subnet Manager.
Subnet Administrator (SA)	An application (normally part of the Subnet Manager) that implements the interface for querying and manipulating subnet management data.
Subnet Manager (SM)	One of several entities involved in the configuration and control of the subnet.
Unicast Linear Forwarding Tables (LFT)	A table that exists in every switch providing the port through which packets should be sent to each LID.

Related Documentation

Table 5 - Reference Documents

Document Name	Description
MFT User Manual	Mellanox Firmware Tools User Manual. See www.mellanox.com > Products > Adapter IB/VPI SW > Firmware Tools
MFT Release Notes	Release Notes for the Mellanox Firmware Tools. See www.mellanox.com > Products > Adapter IB/VPI SW > Firmware Tools

Support and Updates Webpage

Please visit http://www.mellanox.com > Products > Software > InfiniBand/VPI Drivers/VMware Drivers for downloads, FAQ, troubleshooting, future updates to this manual, etc.

1 Mellanox InfiniBand OFED Driver for VMware® vSphere Overview

1.1 Introduction to Mellanox InfiniBand OFED for VMware

Mellanox OFED is a single Virtual Protocol Interconnect (VPI) software stack based on the OpenFabrics (OFED) Linux stack adapted for VMware, and operates across all Mellanox network adapter solutions supporting up to 56Gb/s InfiniBand (IB) and 2.5 or 5.0 GT/s PCI Express 2.0 and 3.0 uplinks to servers.

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

1.2 Introduction to Mellanox InfiniBand Adapters

Mellanox InfiniBand (IB) adapters, which are based on Mellanox ConnectX® family adapter devices, provide leading server and storage I/O performance with flexibility to support the myriad of communication protocols and network fabrics over a single device, without sacrificing functionality when consolidating I/O. For example, IB-enabled adapters can support:

- Connectivity to 10, 20, 40 and 56Gb/s InfiniBand switches
- A unified application programming interface with access to communication protocols including: Networking (TCP, IP, UDP, sockets), Storage (NFS, CIFS, iSCSI, SRP and Clustered Storage), Clustering (MPI, DAPL, RDS, sockets), and Management (SNMP, SMI-S)
- Communication protocol acceleration engines including: networking, storage, clustering, virtualization and RDMA with enhanced quality of service

1.3 Mellanox OFED Package

1.3.1 Software Components

Please refer to the MLNX_OFED_VMware Release Notes for further information.

1.3.2 mlx4 InfiniBand Driver

lease refer to the MLNX_OFED_VMware Release Notes for further information.

1.3.3 ULPs

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 interface supports unicast, multicast and broadcast. For details, see Chapter 3.2, “IP over InfiniBand”.

On VMware ESXi Server, IPoIB supports Unreliable Datagram (UD) mode only, note that Reliable Connected (RC) mode is not supported.

2 Driver Usage

VMware uses a file package called a VIB (VMware Installation Bundle) as the mechanism for installing or upgrading software packages on an ESXi server.

Mellanox InfiniBand OFED driver consists of several dependent kernel modules, each with its own .vib file. In order to install the driver, the VIBs need to be installed together.

For this, Mellanox InfiniBand OFED driver provides a bundle file, a zip file that contain each module VIB file and meta data file that describes the dependencies between them.

The following steps describe how to install, load/unload, remove the driver.

2.1 Installing Mellanox InfiniBand OFED Driver for VMware vSphere

Please uninstall any previous Mellanox driver packages prior to installing the new version.

➢ ***To install the driver:***

1. Log into the ESXi server with root permissions.
2. Install the driver.

```
#> esxcli software vib install -d <path>/<bundle_file>
```

Example:

```
#> esxcli software vib install -d <path>/MLNX-OFED-ESX-1.8.2.4-10EM-
500.0.0.472560.zip
```

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

```
#> esxcli software vib list | grep Mellanox
net-ib-cm         1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-ib-core       1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-ib-ipoib      1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-ib-mad        1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-ib-sa         1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-ib-umad       1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-mlx4-core     1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
net-mlx4-ib       1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
scsi-ib-srp       1.8.2.4-1OEM.500.0.0.472560  Mellanox  PartnerSupported 2014-03-06
```

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

2.2 Removing Mellanox InfiniBand OFED Driver

➢ ***To remove all the drivers:***

1. Log into the ESXi server with root permissions.

2. List the existing InfiniBand OFED driver modules. (see Step 4 in Section 2.1, on page 11)
3. Remove each module.

```
#> esxcli software vib remove -n scsi-ib-srp
#> esxcli software vib remove -n net-ib-ipoib
#> esxcli software vib remove -n net-mlx4-ib
#> esxcli software vib remove -n net-ib-cm
#> esxcli software vib remove -n net-ib-umad
#> esxcli software vib remove -n net-ib-sa
#> esxcli software vib remove -n net-ib-mad
#> esxcli software vib remove -n net-ib-core
#> esxcli software vib remove -n net-mlx4-core
```

To remove the modules, the command must be run in the same order as shown in the example above.

4. Reboot the server.

2.3 Loading/Unloading Driver Kernel Modules

- ***To unload the driver:***

```
#> /opt/mellanox/bin/openibd.sh stop
```

- ***To load the driver:***

```
#> /opt/mellanox/bin/openibd.sh start
```

- ***To restart the driver:***

```
#> /opt/mellanox/bin/openibd.sh restart
```

3 Driver Features

3.1 SCSI RDMA Protocol

3.1.1 SRP Overview

The InfiniBand package includes a storage module called SRP, which causes each InfiniBand port on the VMware ESXi Server machine to be exposed as one or more physical storage adapters, also referred to as vmhbas. To verify that all supported InfiniBand ports on the host are recognized and up, perform the following steps:

1. Connect to the VMware ESX Server machine using the interface of VMware vSphere Client.
2. Select the "Configuration" tab.
3. Click the "Storage Adapters" entry which appears in the "Hardware" list. A "Storage Adapters" list is displayed, describing per device the "Device" it resides on and its type. InfiniBand storage adapters will appear as SCSI adapters.

   InfiniBand storage adapters are listed under HCA name as follow:

   - MT25418[ConnectX VPI - 10GigE/IB DDR, PCIe 2.0 2.5GT/s]
     - vmhba_mlx4_0.1.1 SCSI
     - vmhba_mlx4_0.2.1 SCSI
4. Click on the storage device to display a window with additional details (e.g. Model, number of targets, Logical Units Numbers and their details).

All the features supported by a storage adapter are also supported by an InfiniBand SCSI storage adapter. Setting the features is performed transparently.

3.1.1.1 Module Configuration

The SRP module is configured upon installation to default values. You can use the **esxcfg-module** utility (available in the service console) to manually configure SRP.

1. To disable the SRP module run:

   ```
   cos# esxcfg-module ib_srp -d
   ```
2. Additionally, you can modify the default parameters of the SRP module, such as the maximum number of targets per SCSI host. To retrieve the list and description of the parameters supported by the SRP module, run:

   ```
   cos# vmkload_mod ib_srp -s
   ```
3. To check which parameters are currently used by SRP module, run:

   ```
   cos# esxcfg-module ib_srp -g
   ```
4. To set a new parameter, run:

   ```
   cos# esxcfg-module ib_srp -s <param=value>
   ```
5. To apply your changes, reboot the machine:

   ```
   cos# reboot
   ```

For example, to set the maximum number of SRP targets per SCSI host to four, run:

```
cos# esxcfg-module ib_srp -s 'max_srp_targets=4'
```

6. To find out all SRP's parameters, run:

```
cos# vmkload_mod -s ib_srp
```

Default values are usually optimum per performance however, if you need to manually set the system to achieve better performance, tune the following parameters:

- **srp_sg_tablesize** - Maximum number of scatter lists supported per I/O.
- **srp_cmd_per_num** - Maximum number of commands can queue per lun.
- **srp_can_queue** - Maximum number of commands can queue per vmhba.

7. To check performance:
   a. Windows VM - Assign luns to VM as raw or RDM devices Run iometer or xdd
   b. Linux VM - Assign luns to VM
   c. In case of multiple VMs, generate I/Os for performance testing

3.1.1.2 Multiple Storage Adapter

SRP has the ability to expose multiple storage adapters (also known as vmhbas) over a single InfiniBand port. By default, one storage adapter is configured for each physical InfiniBand port. This setting is determined by a module parameter (called max_vmhbas), which is passed to the SRP module.

1. To change it, log into the service console and run:

```
cos# esxcfg-module ib_srp -s 'max_vmhbas=n'
cos# reboot
```

As a result, <n> storage adapters (vmhbas) will be registered by the SRP module on your machine.

2. To list all LUNs available on your machine, run:

```
cos# esxcfg-mpath -l
```

3.2 IP over InfiniBand

3.2.1 IPoIB Overview

The IP over IB (IPoIB) driver is a network interface implementation over InfiniBand. IPoIB encapsulates Datagram transport service. The IPoIB driver, ib_ipoib, exploits the following ConnectX®-2/ConnectX®-3/ConnectX®-3 Pro capabilities:

- Uses any CX IB ports (one or two)
- Inserts IP/UDP/TCP checksum on outgoing packets
- Calculates checksum on received packets
- Support net device TSO through CX LSO capability to defragment large datagrams to MTU quantas.
- Unreliable Datagram

IPoIB also supports the following software based enhancements:

- Large Receive Offload
- Ethtool support

3.2.2 IPoIB Configuration

1. Install the Mellanox OFED driver for VMware.
2. Verify the driver VIBs are installed correctly. (see Step 4 in Section 2.1, on page 11)
3. Verify the uplinks state is "up".

```
# esxcli network nic list
```

See your VMware distribution documentation for additional information about configuring IP addresses.

4 Configuring the Mellanox InfiniBand OFED Driver for VMware vSphere

4.1 Configuring an Uplink

To configure an Uplink:

1. Add the device as an uplink to an Existing vSwitch using the CLI.
   a. Log into the ESXi server with root permissions.
   b. Add an uplink to a vSwitch.

```
#> esxcli network vSwitch standard uplink add <uplink_name> -v <vswitch_name>
```

2. Verify the uplink is added successfully.

```
#> esxcli network vswitch standard list -v <vswitch name>
```

➢ ***To remove the device locally:***

1. Log into the ESXi server with root permissions.
2. Remove an uplink from a vSwitch.

```
#> esxcli network vswitch standard uplink remove <uplink_name> -v <vswitch_name>
```

For further information, please refer to:

http://pubs.vmware.com/vsphere-50/topic/com.vmware.vcli.migration.doc_50/cos_upgrade_technote.1.9.html#1024629

4.2 Configuring VMware ESXi Server Settings

VMware ESXi Server settings can be configured using the vSphere Client. Once the InfiniBand OFED driver is installed and configured, the administrator can make use of InfiniBand software available on the VMware ESXi Server machine. The InfiniBand package provides networking and storage over InfiniBand. The following sub-sections describe their configuration.

This section includes instructions for configuring various module parameters.

From ESXi use the following command to view all the available module parameters and default settings.

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

When using ESXi, use vMA or remote CLI vicfg-module.pl to configure the module parameters in a similar way to what is done in the Service Console (COS) for ESXi.

4.2.1 Subnet Manager

The driver package requires InfiniBand Subnet Manager (SM) to run on the subnet. The driver package does not include an SM.

If your fabric includes a managed switch/gateway, please refer to the vendor's user's guide to activate the built-in SM.

If your fabric does not include a managed switch/gateway, an SM application should be installed on at least one non-ESXi Server machine in the subnet. You can download an InfiniBand SM such as OpenSM from www.openfabrics.org under the Downloads section.

4.2.2 Networking

The InfiniBand package includes a networking module called IPoIB, which causes each InfiniBand port on the VMware ESXi Server machine to be exposed as one or more physical network adapters, also referred to as uplinks or vmnics. To verify that all supported InfiniBand ports on the host are recognized and up, perform the following steps:

1. Connect to the VMware ESXi Server machine using the interface of VMware vSphere Client.
2. Select the "Configuration" tab.
3. Click the "Network Adapters" entry which appears in the "Hardware" list.

   A "Network Adapters" list is displayed, describing per uplink the "Device" it resides on, the port "Speed", the port "Configured" state, and the "vSwitch" name it connects to.

To create and configure virtual network adapters connected to InfiniBand uplinks, follow the instructions in the ESXi Server Configuration Guide document.

All features supported by Ethernet adapter uplinks are also supported by InfiniBand port uplinks (e.g., VMware® VMotionTM, NIC teaming, and High Availability), and their setting is performed transparently.

4.2.3 Virtual Local Area Network (VLAN) Support

To support VLAN for VMware ESXi Server users, one of the elements on the virtual or physical network must tag the Ethernet frames with an 802.1Q tag. There are three different configuration modes to tag and untag the frames as virtual machine frames:

1. Virtual Machine Guest Tagging (VGT Mode).
2. ESXi Server Virtual Switch Tagging (VST Mode).
3. External Switch Tagging (EST Mode).

EST is supported for Ethernet switches and can be used beyond IB/Eth Gateways transparently to VMware ESXi Servers within the InfiniBand subnet.

To configure VLAN for InfiniBand networking, the following entities may need to be configured according to the mode you intend to use:

- Subnet Manager Configuration

  Ethernet VLANs are implemented on InfiniBand using Partition Keys (See RFC 4392 for information). Thus, the InfiniBand network must be configured first. This can be done by configuring the Subnet Manager (SM) on your subnet. Note that this configuration is needed for both VLAN configuration modes, VGT and VST.

  For further information on the InfiniBand Partition Keys configuration for IPoIB, see the Subnet Manager manual installed in your subnet.

  The maximum number of Partition Keys available on Mellanox HCAs is:

  - 128 for ConnectX® IB family
  - Check with IB switch documentation for the number of supported partition keys
- Guest Operating System Configuration

For VGT mode, VLANs need to be configured in the installed guest operating system. This procedure may vary for different operating systems. See your guest operating system manual on VLAN configuration.

In addition, for each new interface created within the virtual machine, at least one packet should be transmitted. For example:

Create a new interface (e.g., <eth1>) with IP address <ip1>.

➢ ***To guarantee that a packet is transmitted from the new interface:***

```
arping -I <eth1> <ip1> -c 1
```

- Virtual Switch Configuration

  For VST mode, the virtual switch using an InfiniBand uplink needs to be configured. For further information, see the *ESXi Server 3 Configuration Guide* and *ESXi Server 3 802.1Q VLAN Solutions* documents.

4.2.4 Maximum Transmit Unit (MTU) Configuration

On VMware ESXi Server machines, the MTU is set to 1500 bytes by default. IPoIB supports larger values and allows Jumbo Frames (JF) traffic up to 4052 bytes on VI3 and 4092 bytes on vSphere 5. The maximum value of JF supported by the InfiniBand device is:

- 2044 bytes for the InfiniHost III family
- 4052 / 4092 bytes for ConnectX® IB family (vSphere 5)

Running a datagram IPoIB MTU of 4092 requires that the InfiniBand MTU is set to 4k.

It is the administrator's responsibility to make sure that all the machines in the network support and work with the same MTU value. For operating systems other than VMware ESXi Server, the default value is set to 2044 bytes.

The procedure for changing the MTU may vary, depending on the OS. For example, to change it to 1500 bytes:

- On Linux - if the IPoIB interface is named ib0:

```
ifconfig ib0 mtu 1500
```

- On Microsoft® Windows - execute the following steps:
  a. Open "Network Connections"
  b. Select the IPoIB adapter and right click on it
  c. Select "Properties"
  d. Press "Configure" and then go to the "Advanced" tab
  e. Select the payload MTU size and change it to 1500
  f. Make sure that the firmware of the HCAs and the switches supports the MTU you wish to set.
  g. Configure your Subnet Manager (SM) to set the MTU value in the configuration file. The SM configuration for MTU value is per Partition Key (PKey).

     For example, to enable 4K MTUs on a default PKey using the OpenSM SM6, log into the Linux machine (running OpenSM) and perform the following commands:

h. Edit the file:

```
/usr/local/ofed/etc/opensm/partitions.conf
```

and include the line:

```
key0=0x7fff,ipoib,mtu=5 : ALL=full;
```

i. Restart OpenSM:

```
/etc/init.d/opensmd restart
```

To enable 4k mtu support: run **"esxcli system module parameters set -m=mlx4_core -p=mtu_4k=1"**.
Changes will take effect after the reboot.

4.2.5 High Availability

High Availability is supported for both InfiniBand network and storage adapters. A failover port can be located on the same HCA card or on a different HCA card on the same system (for hardware redundancy).

To define a failover policy for InfiniBand networking and/or storage, follow the instructions in the *ESXi Server Configuration Guide* document.