Mellanox Technologies

InfiniBand OFED Driver for

VMware® Infrastructure 3

Installation Guide

Rev 1.21
# Table of Contents

Revision History .................................................................................................................................................. 4

1  Introduction ................................................................................................................................................ 4

2  Hardware Support ....................................................................................................................................... 9

3  Software Support ....................................................................................................................................... 9
    3.1  User Layer Protocols (ULPs) Support ............................................................................................... 9
    3.2  Tools ................................................................................................................................................... 10

4  Setup and Configuration ............................................................................................................................ 10
    4.1  Installation ........................................................................................................................................ 10
    4.2  Usage .................................................................................................................................................. 10
    4.3  Configuration ..................................................................................................................................... 11
        4.3.1  Subnet Manager ......................................................................................................................... 11
        4.3.2  Networking ................................................................................................................................. 11
        4.3.3  Storage ....................................................................................................................................... 16
        4.3.4  Performance ............................................................................................................................... 18
        4.3.5  High Availability ....................................................................................................................... 18

5  Firmware Burning ..................................................................................................................................... 18

6  Useful Links .............................................................................................................................................. 18

7  Hardware Compatibility List ...................................................................................................................... 19
Revision History

Table 1- History Table

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>October 2008</td>
<td>- ConnectX PCIe Gen1/2 support</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- New IPoIB and SRP features</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Performance enhancements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Added <em>ibstat</em> utility to package</td>
</tr>
<tr>
<td>1.1</td>
<td>January 2008</td>
<td>First GA version</td>
</tr>
</tbody>
</table>

1 Introduction

This document is a user’s guide for installing the *InfiniBand OFED Driver for VMware® Infrastructure 3, version 3.5*. VMware Infrastructure 3 version 3.5 is comprised of:

- VMware® VirtualCenter 2 version 2.5
- VMware® ESX Server 3 version 3.5

The InfiniBand OFED software driver package adds to VMware Infrastructure 3 support for the InfiniBand interconnect technology. The driver package is based on the OpenFabrics Enterprise Distribution, OFED\(^1\), and provides support for Mellanox Technologies InfiniBand (IB) products along with the upper layer protocols IP-over-InfiniBand (IPoIB) and SCSI RDMA Protocol (SRP).

Introducing the InfiniBand OFED driver into a VMware Infrastructure 3 environment provides consolidated networking and storage services over the same InfiniBand port.

In addition, InfiniBand services support VMware Infrastructure 3 features like port failover and Virtual LANs. These features allow for the replacement of several storage and network adapters installed on an ESX Server machine by a significantly lower number of InfiniBand adapters, saving cost and power, and reducing cabling complexity. For example, a system with several Gigabit Ethernet adapters and Fiber Channel adapters can be replaced by one InfiniBand adapter without loss of performance. Furthermore, each virtual machine (VM) or ESX Server connection (e.g., VMware® VMotion, service console) can reside over a different InfiniBand Virtual LAN, providing for strict isolation. See Figure 1 and Figure 2 for illustration.

\(^1\) Check the *InfiniBand OFED Driver for VMware(R) Infrastructure 3 version 3.5 Release Notes* for the OFED revision.
When VMware ESX Server is used with the InfiniBand OFED driver, the maintenance of traditional network and storage interfaces available in virtual machines still enables the seamless run of existing applications qualified for the virtual machines.

Similarly, the configuration, usage, and allocation of InfiniBand LAN and SAN I/O resources to virtual machines remain seamless. This is because VMware® Virtual Infrastructure (VI) Client exposes the InfiniBand network and storage interfaces as traditional network (NIC) and storage (HBA) interfaces, see Figure 3 and Figure 4.
Figure 3 – InfiniBand Network Adapters as Listed in VI Client

Figure 4 – InfiniBand Storage Adapters as Listed in VI Client

2 The VI Client lists mistakenly InfiniBand storage adapters under “Unknown” family. See the InfiniBand OFED Driver for VMware(R) Infrastructure 3 version 3.5 Release Notes.
In addition, the InfiniBand driver has the ability to expose each InfiniBand port as multiple network and storage adapters. This provides the administrator the same look and feel of multiple traditional adapters installed on a VMware ESX Server machine, when, in reality, they all reside on a single high performance InfiniBand port.

Figure 5 shows multiple InfiniBand network adapters exposed by a single HCA (MT25418 ConnectX IB DDR Dual Port HCA).

**Figure 5 – Multiple InfiniBand Network Adapters Exposed by a Single HCA**

<table>
<thead>
<tr>
<th>Device</th>
<th>Speed</th>
<th>Configured</th>
<th>vSwitch</th>
<th>Observed IP address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcom NetXtreme II BCM5781 1000Base-T</td>
<td>100 Full</td>
<td>N/A</td>
<td>vs05h0e</td>
<td>25.4.0.1-25.4.31.254</td>
</tr>
<tr>
<td>vmnet0</td>
<td>1000 Full</td>
<td>N/A</td>
<td>vs05h0e</td>
<td>None</td>
</tr>
<tr>
<td>vmnet1</td>
<td>1000 Full</td>
<td>N/A</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>MT25204</td>
<td>N/A</td>
<td>N/A</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

**Figure 6 shows six virtual machines, where each machine uses a different InfiniBand network adapter and one shared regular Ethernet adapter (center).**

**Figure 6 – Usage of Multiple InfiniBand Network Adapters**
Figure 7 shows multiple InfiniBand storage adapters exposed by a single HCA (MT25418 ConnectX IB DDR Dual Port HCA).

**Figure 7 – Multiple InfiniBand Storage Adapters Exposed by a Single HCA**

Instructions on how to enable and configure this feature are available under the section Configuration on page 10.

Figure 8 shows six virtual machines, where each machine uses a different InfiniBand storage adapter and one shared regular storage adapter (center).

**Figure 8 - Usage of Multiple InfiniBand Storage Adapters**
2 Hardware Support

The InfiniBand OFED driver package supports all Mellanox Technologies Host Channel Adapter (HCA) cards based on the HCA devices listed in Table 2.

Table 2- Supported Mellanox Technologies HCA Devices

<table>
<thead>
<tr>
<th>Device Name</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT25408 [ConnectX® IB SDR]</td>
<td>ConnectX IB HCA/TCA IC, dual-port, SDR, PCIe 2.0 2.5 GT/s, x8, mem-free, RoHS</td>
</tr>
<tr>
<td>MT25418 [ConnectX® IB DDR, PCIe 2.0 2.5GT/s]</td>
<td>ConnectX IB HCA/TCA IC, dual-port, DDR, PCIe 2.0 2.5 GT/s, x8, mem-free, RoHS</td>
</tr>
<tr>
<td>MT26418 [ConnectX® IB DDR, PCIe 2.0 5.0GT/s]</td>
<td>ConnectX IB HCA/TCA IC, dual-port, DDR, PCIe 2.0 5.0 GT/s, x8, mem-free, RoHS</td>
</tr>
<tr>
<td>MT26428 [ConnectX® IB QDR, PCIe 2.0 5.0GT/s]</td>
<td>ConnectX IB HCA/TCA IC, dual-port, QDR, PCIe 2.0 5.0 GT/s, x8, mem-free, RoHS</td>
</tr>
<tr>
<td>MT25204 [InfiniHost® III Lx HCA]</td>
<td>InfiniHost III HCA/TCA IC, single-port, SDR/DDR, PCIe x8, mem-free, A0, RoHS</td>
</tr>
<tr>
<td>MT25208 InfiniHost® III Ex (InfiniHost compatibility mode)</td>
<td>InfiniHost III HCA/TCA IC, dual-port, SDR/DDR, PCIe x8, DDR memory interface, RoHS</td>
</tr>
<tr>
<td>MT25208 InfiniHost® III Ex (Mem-free mode)</td>
<td>InfiniHost III HCA/TCA IC, dual-port, SDR/DDR, PCIe x8, mem-free, RoHS</td>
</tr>
</tbody>
</table>

3 Software Support

The InfiniBand OFED driver package for VMware Infrastructure 3 is based on the OpenFabrics Enterprise Distribution, OFED 1.3.1. See http://www.openfabrics.org.

3.1 User Layer Protocols (ULPs) Support

The following ULPs are supported:

- IP over InfiniBand (IPoIB)
  - Unreliable Datagram (UD) mode only
- SCSI RDMA Protocol (SRP)
3.2 Tools

The package includes (and installs) the `ibstat` utility which allows the user to retrieve information on the InfiniBand devices and ports installed on an ESX Server machine - see instructions on section 4.2.

4 Setup and Configuration

The InfiniBand OFED driver installation on VMware ESX Server is done using a Red Hat package manager (RPM).

4.1 Installation

To install the driver package on a VMware ESX Server machine, log into the service console as root and execute the following commands:

1. If you have a previous version of the driver package installed on your system, you must uninstall it first. To see all packages installed on your machine, run:
   
   ```
   cos# rpm -qa
   ```

   To uninstall an old driver package, run:
   
   ```
   cos# rpm -e <IB driver package name>
   ```

2. To identify your ESX Server build number, run:
   
   ```
   cos# vmware -v
   ```

3. Download the suitable RPM based on the build number of your machine. RPMs for all ESX 3.5 versions are available under:
   
   ```
   http://www.mellanox.com/products/ciov_ib_drivers_vi3-1.php
   ```

4. Install the RPM. Run:
   
   ```
   cos# rpm -i <RPM file>
   ```

5. Reboot your machine. Run:
   
   ```
   cos# reboot
   ```

For more information on how to remotely connect to the service console, please refer to the ESX Server 3 Configuration Guide document (see "Useful Links" on page 18).

4.2 Usage

To manage the InfiniBand OFED driver and ULPs on VMware ESX Server, use the management script located at `/etc/init.d/mgmt-mlx`.

For example, to check the status of the InfiniBand OFED driver and the ULPs, run the following command:
cos# /etc/init.d/mgmt-mlx status

For help, run the script without flags.

To display the InfiniBand OFED driver details, run:

   cos# rpm -qi <package name>

You can also retrieve information on InfiniBand ports available on your machine using `ibstat`. For usage and help, log into the service console and run:

   cos# ibstat -h

4.3 Configuration

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

4.3.1 Subnet Manager

The driver package requires an InfiniBand Subnet Manager (SM) to be running in 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-ESX Server machine in the subnet. You can download an InfiniBand SM such as OpenSM from www.openfabrics.org under the Downloads section.

4.3.2 Networking

The InfiniBand package includes a networking module called IPoIB, which causes each InfiniBand port on the VMware ESX 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 ESX Server machine using the interface of VMware VI Client.
2. Select the “Configuration” tab.
3. Click the “Network Adapters” entry which appears in the “Hardware” list.

---

3 `ibstat` of this package produces similar output to `ibstat` of the Linux OFED package.
4 VMware ESX Server does not support any InfiniBand SM.
4. 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 *ESX Server 3 Configuration Guide* document.

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

### 4.3.2.1 Module Configuration

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

To disable the IPoIB module run:

```
cos# esxcfg-module ib_ipoib -d
```

For usage and help, run the utility with the `--help` flag.

In addition, you can modify the default parameters of IPoIB such as receive and transmit rings sizes and numbers. To retrieve the list and description of the parameters supported by the IPoIB module, run:

```
cos# vmkload_mod ib_ipoib -s
```

To check which parameters are currently used by the IPoIB module, run

```
cos# esxcfg-module ib_ipoib -g
```

To set a new parameter, run:

```
cos# esxcfg-module ib_ipoib -s <param=value>
```

To apply your changes, reboot the machine:

```
cos# reboot
```

Note that some module parameters must be passed as a comma-separated list of integers, where each integer corresponds to an InfiniBand port on the machine according to its PCI bus order\(^5\). For example, to change the receive ring size of the first two InfiniBand ports, run:

```
cos# esxcfg-module ib_ipoib -s 'RxQueueSize=4096,4096'
```

### 4.3.2.2 Multiple Network Adapters

IPoIB has the ability to expose multiple network adapters (also known as vmnics or uplinks) over a single InfiniBand port. Each port can expose up to 16 network adapters, and each ESX Server machine can expose up to 32 network adapters. By

\(^5\) The PCI bus number is retrieved using the `lspci` command from the service console.
default, one network adapter is configured for each physical InfiniBand port. This setting is determined by a module parameter (called UplinkNum), which is passed to the IPoIB module. To change it, log into the service console and run:

```
cos# esxcfg-module ib_ipoib -s 'UplinkNum=n1,n2,..'
```

where n1,n2,.. is list of uplinks number for InfiniBand ports p1,p2,..

For example, on a machine with two InfiniBand ports, you can set four uplinks for the first port and six for the second one. To do so, run:

```
cos# esxcfg-module ib_ipoib -s 'UplinkNum=4,6'
cos# reboot
```

As a result, ten uplinks will be registered by the IPoIB module on your machine.

To display the list of network adapters run:

```
cos# esxcfg-nics -l
```

4.3.2.3 NetQueue

VMware ESX Server supports NetQueue, a performance technology that significantly improves networking performance in virtualized environments. NetQueue requires MSI-X support from the server platform. The MSI-X interrupt mechanism provides better performance and load balancing.

MSI-X support is currently limited to specific systems - see System Compatibility Guide For ESX Server 3.5 for details.

Note that NetQueue is supported only on ConnectX InfiniBand cards. To enable it, log into the service console and run the following commands:

1. Make sure that the InfiniBand device installed on your machine is a ConnectX product. Run the following command and examine the output.

```
cos# lspci
```

2. Enable NetQueue in the ESX Server kernel.

```
cos# esxcfg-advcfg -k TRUE netNetqueueEnabled
```

3. Enable MSI-X in the InfiniBand basic module.

```
cos# esxcfg-module ib_basic -s
  '--export-ksyms=linux mlx4_msi_x=1'
```

4. Enable NetQueue in the IPoIB module.

```
cos# esxcfg-module ib_ipoib -s
  'NetQueue=<ib-ports-list>'
```

For example, to enable NetQueue on the first two InfiniBand ports installed on the machine, run:
cos# esxcfg-module ib_ipoib -s 'NetQueue=1,1'

5. Reboot your machine for the changes to take effect.

   cos# reboot

6. After reboot, make sure that NetQueue was successfully enabled on uplink “vmnicX”.

   cos# cat /proc/net/ipoib/vmnicX/NetQueue

   If the output is “1”, the NetQueue is enabled. For example:

   cos# cat /proc/net/ipoib/vmnic32/NetQueue
   cos# 1

4.3.2.4 Virtual Local Area Network (VLAN) Support

To support VLAN for VMware ESX 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. ESX Server Virtual Switch Tagging (VST Mode)
3. External Switch Tagging (EST Mode)

Note that EST is supported for Ethernet switches and can be used beyond IB/Eth Gateways transparently to VMware ESX 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.

  See the Subnet Manager manual installed in your subnet for InfiniBand Partition Keys configuration for IPoIB.

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

  - 64 for the InfiniHost™ III family
  - 128 for ConnectX™ IB family

- 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, run:

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

**Virtual Switch Configuration**

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

### 4.3.2.5 Maximum Transmit Unit (MTU) Configuration

On VMware ESX 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 in size. The maximum value of JF supported by the InfiniBand device is:

- 2044 bytes for the InfiniHost III family
- 4052 bytes for ConnectX IB family

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 ESX 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, run:
  
  ```
  ifconfig ib0 mtu 1500
  ```

- **On Microsoft® Windows** - execute the following steps:
  1. Open "Network Connections"
  2. Select the IPoIB adapter and right click on it
  3. Select "Properties"
  4. Press "Configure" and then go to the "Advanced" tab
  5. Select the payload MTU size and change it to 1500

- **On VMware ESX Server 3.5** - follow the instructions under the section *Enabling Jumbo Frames* in *ESX Server 3 Configuration Guide* document.
Make sure that the firmware of the HCAs and the switches supports the MTU you wish to set.

Also 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 4052-byte MTUs on a default PKey using the OpenSM SM\(^6\), log into the Linux machine (running OpenSM) and perform the following commands:

1. Edit the file:
   
   ```
   /usr/local/ofed/etc/opensm/partitions.conf
   ```
   
   and include the line:
   
   ```
   key0=0x7fff,ipoib,mtu=5 : ALL=full;
   ```

2. Restart OpenSM:
   
   ```
   /etc/init.d/opensmd restart
   ```

4.3.3 Storage

The InfiniBand package includes a storage module called SRP, which causes each InfiniBand port on the VMware ESX 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 VI 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 mistakenly under “Unknown” family label. To make sure what storage adapters (vmhbas) are associated with your InfiniBand device, log into the service console and run:

   ```
   cos# ibstat -vmhba
   ```

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).

To allocate new storage for the InfiniBand target LUN of the InfiniBand storage adapter, follow the instructions in the *ESX Server 3 Configuration Guide* document.

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

---

\(^6\) Version 3.1
4.3.3.1 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.

To disable the SRP module run:

```
cos# esxcfg-module ib_srp -d
```

In addition, 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
```

To check which parameters are currently used by SRP module, run:

```
cos# esxcfg-module ib_srp -g
```

To set a new parameter, run:

```
cos# esxcfg-module ib_srp -s <param=value>
```

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'
```

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

```
cos# esxcfg-mpath -l
```
4.3.4 Performance

For best performance, it is recommended to use ConnectX InfiniBand cards since they have enhanced capabilities and offloading features.

To enhance networking performance, it is recommended to enable NetQueue as explained in Section 4.3.2.3 and to use Jumbo Frames as explained in Section 4.3.2.5.

In addition, please read and follow the instructions in the Performance Tuning Best Practices for ESX Server 3 document.

4.3.5 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 ESX Server 3 Configuration Guide document.

5 Firmware Burning

Use the MFT for VMware ESX Server 3.5 tools package to burn a firmware image onto the Flash of an InfiniBand device. Download the binaries and the documents from:


6 Useful Links

Use the following links to obtain more information on InfiniBand and OFED:

- OpenFabrics Alliance - http://www.openfabrics.org
- General documentation - http://www.vmware.com/support/pubs

---

7 You can identify which InfiniBand card is installed on your machine using the `lspci` command from the service console.
Performance Tuning Best Practices for ESX Server 3 -
http://www.vmware.com/resources/techresources/707

Compatibility guides and HCL -
http://www.vmware.com/resources/techresources/cat/119

Discussion forums - http://www.vmware.com/community

About Community Source - http://www.vmware.com/communitysource

7 Hardware Compatibility List

All adapter cards based on Mellanox Technologies’ HCA devices listed in Table 2 on page 9