Mellanox InfiniBand OFED Driver for VMware vSphere 4.X

User Manual

Rev 1.4.1-2.0.000

Last Updated: June 28, 2011

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
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies, Ltd.
PO Box 586 Hermon Building
Yokneam 20692
Israel
Tel: +972-4-909-7200
Fax: +972-4-959-3245

Table of Contents

List of Tables

Revision History

Rev 4.1 (May, 2011)

- Initial version of the document

Preface

This Preface provides general information concerning the scope and organization of this User's Manual. It includes the following sections:

Intended Audience

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.

Documentation Conventions

Typographical Conventions

Table 1 - Typographical Conventions

Description	Convention	Example
File names	`file.extension`	
Directory names	`directory`	
Commands and their parameters	**`command param1`**	
Optional items	`[ ]`	
Mutually exclusive parameters	`{ p1 | p2 | p3 }`	
Optional mutually exclusive parameters	`[ p1 | p2 | p3 ]`	
Prompt of a *user* command under bash shell	`hostname$`	
Prompt of a *root* command under bash shell	`hostname#`	
Prompt of a *user* command under tcsh shell	`tcsh$`	
Environment variables	VARIABLE	
Code example	`if (a==b){};`	
Comment at the beginning of a code line	!, #	
Characters to be typed by users as-is	**bold font**	
Keywords	**bold font**	
Variables for which users supply specific values	*Italic font*	
Emphasized words	*Italic font*	*These are emphasized words*
Pop-up menu sequences	menu1 --> menu2 -->... --> item	
Note	**Note:**	
Warning	**Warning!**	

Common Abbreviations and Acronyms

Table 2 - Abbreviations and Acronyms (Sheet 1 of 2)

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

Table 2 - Abbreviations and Acronyms (Sheet 2 of 2)

Abbreviation / Acronym	Whole Word / Description
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
VPI	Virtual Protocol Interconnect
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 3 - 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.
Virtual Protocol Interconnet (VPI)	A Mellanox Technologies technology that allows Mellanox channel adapter devices (ConnectX®) to simultaneously connect to an InfiniBand subnet and a 10GigE subnet (each subnet connects to one of the adpater ports)

Related Documentation

Table 4 - Reference Documents

Document Name	Description
MFT User's Manual	Mellanox Firmware Tools User's Manual. See under `docs/` folder of installed package.
MFT Release Notes	Release Notes for the Mellanox Firmware Tools. See under `docs/` folder of installed package.

Support and Updates Webpage

Please visit http://www.mellanox.com > Products > Adapter IB/VPI SW/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 Internconnect (VPI) software stack based on the OpenFabrics (OFED) Linux stack adapted for VMware, and operates across all Mellanox network adapter solutions supporting 10, 20 and 40Gb/s InfiniBand (IB); 10Gb/s Ethernet (10GigE); and 2.5 or 5.0 GT/s PCI Express 2.0 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 VPI Adapters

Mellanox VPI adapters, which are based on Mellanox ConnectX® and ConnectX®-2 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, VPI-enabled adapters can support:

- Connectivity to 10, 20 and 40Gb/s InfiniBand switches, Ethernet switches, emerging Data Center Ethernet switches and InfiniBand to Ethernet
- A single firmware image for dual-port ConnectX/ConnectX-2 adapters that supports independent access to different convergence networks (InfiniBand, Ethernet or Data Center Ethernet) per port
- 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

MLNX_OFED_VMware contains the following software components:

- Mellanox Host Channel Adapter Drivers
  - mlx4 (VPI), which is split into multiple modules:
    - ib_basic (low level IB helper)
    - mlx4_en (Ethernet)
- Mid-layer core
  - Verbs, MADs, SA, CM, CMA, uVerbs, uMADs
- Upper Layer Protocols (ULPs)

  - IPoIB, SRP Initiator
- Utilities
  - Diagnostic tools
  - Performance tests
- Firmware tools (MFT)
- Documentation

1.4 mlx4 VPI Driver

`mlx4` is the low level driver implementation for the ConnectX® and ConnectX®-2 adapters designed by Mellanox Technologies. ConnectX/ConnectX-2 can operate as an InfiniBand adapter, as an Ethernet NIC. The Mellanox OFED driver for VMware supports InfiniBand and Ethernet NIC configurations. To accommodate the supported configurations, the driver is split into four modules:

ib_basic

Handles low-level functions like device initialization and firmware commands processing. Also controls resource allocation so that the InfiniBand and Ethernet functions can share the device without interfering with each other.

Handles InfiniBand-specific functions and plugs into the InfiniBand midlayer

mlx4_en

A 10GigE driver under drivers/net/mlx4 that handles Ethernet specific functions and plugs into the netdev mid-layer

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

SRP

SRP (SCSI RDMA Protocol) 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 Chapter 3.1, “SCSI RDMA Protocol”.

1.4.2 Mellanox Firmware Tools

Mellanox Firmware Tools solution applies to VMware ESX Servers only.

The Mellanox Firmware Tools (MFT) package is a set of firmware management tools for a single InfiniBand node. MFT can be used for:

- Generating a standard or customized Mellanox firmware image
- Querying for firmware information
- Burning a firmware image to a single InfiniBand node

The following are the MFT package content:

- mstflint - firmware burning and diagnostic tools for Mellanox manufactured HCA/NIC cards.

  Please note, this burning tool should be used only with Mellanox-manufactured HCA/NIC cards. Using it with cards manufactured by other vendors may be harmful to the cards (due to different configurations). Using the diagnostic tools is normally safe for all HCAs/NICs.

  For details, see: http://www.mellanox.com/pdf/firmware/mstflint_README.txt

  For details on tools for ESX-4.x, see: http://mellanox.com/pdf/MFT/mlnx_mini_tools_for_vmesx40_release_notes.txt

  For details on tools for ESXi-4.x, see: http://mellanox.com/pdf/MFT/ESXi_bootable_mst_README.txt

2 Installation

This chapter describes how to install and test the Mellanox InfiniBand OFED Driver for VMware vSphere package on a single host machine with Mellanox InfiniBand and/or Ethernet adapter hardware installed. The chapter includes the following sections:

2.1 Hardware and Software Requirements

2.1.1 Hardware Requirements

For the supported hardware compatibility list (HCL), please refer to:

http://communities.vmware.com/cshwsw.jspa?vendor=mellanox

2.1.2 Software Requirements

The InfiniBand OFED driver package for VMware vSphere 4 is based on the OpenFabrics Enterprise Distribution, OFED 1.4.1.

See http://www.openfabrics.org

For the supported hardware compatibility list (HCL), please refer to:

- InfiniBand cards:

  http://communities.vmware.com/cshwsw.jspa?vendor=mellanox

- Ethernet cards:

  http://www.vmware.com/resources/compatibility/search?action=search&deviceCategory=io&key=mellanox

2.1.3 Upper Layer Protocols (ULPs) Support

InfiniBand OFED driver for VMware vSphere 4.x supports the following ULPs:

- IP over InfiniBand (IPoIB)
- SCSI RDMA Protocol (SRP)

2.1.4 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.

To use **`ibstat`**, the IPoIB driver must be loaded and at least one IPoIB interface must be available.

2.2 InfiniBand OFED Driver Installation on VMware ESX/ESXi Server

The InfiniBand OFED driver installation on VMware ESX Server 4.x is done using VMware's VIB bundles.

Please uninstall any previous versions on ESX/ESXi before installing the new version.

2.2.1 Installing InfiniBand OFED Driver on a VMware ESXi Server

ESXi package is available as a standalone or as part of the vSphere Management Assistant (vMA) virtual appliance from http://www.vmware.com.

For further information on how to use is, please refer to VMware's document "vSphere Command-Line Interface Installation and Reference Guide".

To install vSphere, perform the following steps:

1. Set the machine to maintenance mode
2. Log into the service console as root and execute the following steps:
   - Before installing, please verify if you have on your machine of the following ESXi builds. Run:
     /usr/bin/vmware -v
     - VMware ESXi 4.1.0 build-348481' - ESX 4.1 Update 1
     - VMware ESXi 4.1.0 build-260247' - ESX 4.1
     - VMware ESXi 4.0.0 build-164009' - ESX 4.0
     - VMware ESXi 4.0.0 build-241301' - ESX 4.0 Update 1
     - VMware ESXi 4.0.0 build-261974' - ESX 4.0 Update 2

     For further information on how to run the command, please go to:

     http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1003677

     If you have a different version, create an empty file **/NO_VER_CHECK**.
3. Install vSphere remote CLI.
4. Check the packages installed on your machine to retrieve the bulletin ID. Run:
   ```
   rcli#  vihostupdate --server <server ip> --query
   ```
5. Uninstall any previous version of the driver package installed on your system. Run:
   ```
   rcli#  esxupdate --server <server ip> --bulletin <bulletin id> --removereboot
   ESXi
   ```
6. Verify your machine is in maintenance mode.
7. Install the mlx4_en driver. Run:
   ```
   rcli# vihostupdate --server <server ip> --bundle <MLX-OFED VIB> --install
   ```
8. Configure VPI functionality if needed.
9. Reboot ESXi server.

2.2.2 Installing InfiniBand OFED Driver on a VMware ESX Server

To install the driver package on a VMware ESX Server machine:

1. Set the machine to maintenance mode
2. Log into the service console as root and execute the following steps:
   - Before installing, please verify if you have on your machine of the following ESX builds. Run :
     /usr/bin/vmware -v
     - VMware ESX 4.1.0 build-348481' - ESX 4.1 Update 1
     - VMware ESX 4.1.0 build-260247' - ESX 4.1
     - VMware ESX 4.0.0 build-164009' - ESX 4.0
     - VMware ESX 4.0.0 build-241301' - ESX 4.0 Update 1
     - VMware ESX 4.0.0 build-261974' - ESX 4.0 Update 2

   If you have a different version, create an empty file **/NO_VER_CHECK**.
3. Check the packages installed on your machine and to retrieve the bulletin ID. Run:
   cos# esxupdate query
4. Uninstall any previous version of the driver package installed on your system. Run:
   cos# esxupdate -b <Bulletin ID to remove> remove
5. Verify your machine is in maintenance mode.
6. Install the mlx4_en driver. Run:
   cos# esxupdate --bundle <MLX-OFED VIB bundle full path> update
7. Configure VPI functionality if needed.
8. Reboot ESX server.

2.3 Retrieving Information on InfiniBand OFED Drivers

1. Display the InfiniBand OFED driver details. Run:
   cos# esxupdate query
   cos# esxupdate -b <package bulletin id> info
2. Retrieve information on InfiniBand ports available on your machine. Run:
   ibstat
   For usage and help, log into the service console and run:

   cos# ibstat -h

2.4 Installing Mellanox IB OFED Driver on Non Major ESX/ESXi 4.x Release

1. Download the **ESX_4.x_patch.sh** script from www.mellanox.com > Products > Adapter IB/VPI SW > VMware Drivers.
2. Run the script.
3. Install the ESX/ESXi driver according to the procedures described in Section 2.2.1 (page 15) and Section 2.2.2 (page 16).

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 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 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

   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).

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/ConnectX-2 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 drivers are installed correctly and linked up. Run:

```
esxcfg-nics -l
```

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

4 Working With VPI

4.1 VPI - Multi-Protocol Support

This driver package supports Mellanox's multi-protocol VPI technology. VPI means the driver supports the coexistence of 10GigE NICs and IB HCAs on the same host (ESX server), and depending on the ConnectX device type also the coexistence of 10GigE and IB ports on the same HCA device.

The following port configurations are supported in VPI: (IB,IB), (IB,ETH) and (ETH,ETH).

Port 1 ETH and port 2 IB configuration is not allowed.

InfiniHost® III and some of the ConnectX® HCA cards are not VPI capable. If your device does not support the configured port types, it will start with the device's supported configuration instead.

Look for an error message in:

- ESX, the errors are printed to **`/var/log/vmkernel`** and older messages to **`/var/log/vmkernel [1...]`**
- ESXi, the errors from several logs are all printed to **`/var/log/messages`** and older messages to **`/var/log/messages [1...]`**

  The vmkernel messages are printed with a vmkernel prefix.

After configuring VPI, reboot the ESX for the changes to take effect.

4.1.1 Configuring VPI Ports using Mellanox Scripts

All port types in VMware v1.4.1-2.0.000 are InfiniBand (IB) by default. To manually configure non IB ports, please see Section 4.1.2, “Removing Corrupted Configuration Files,” on page 21.

If the configuration file is corrupted, the scripts below will not repair the corrupted sections. To delete the bogus sections, see Section 4.1.2, “Removing Corrupted Configuration Files,” on page 21.

- Configure the port types interactively. Run:

  `cos# connectx_port_config`

  `connectx_port_config` is installed by the **`ib_basic`** package, and is available only for ESX.

  Both the ESX/ESXi have a none interactive mode

```
ESX cos> connectx_port_config -h for detailed info
ESX cos> connectx_port_config -d <PCI_ID> -c< configuration>
ESXi cos> connectx_port_config for detailed info
ESXi cos> connectx_port_config <PCI_ID> < configuration>
```

4.1.2 Removing Corrupted Configuration Files

In certain cases the old **connectx_port_config** script may leave behind some bogus port configuration. To delete them, follow the steps below:

1. Retrieve the current configuration string. Run:
   - On ESX:

     cos# esxcfg-module -g mlx4_en
   - On ESXi:

     rcli# vicfg-module.pl --server <ip> -g mlx4_en
2. Configure the module without the port configuration. Run:

   esxcfg-module -s <config line from previous step> mlx4_en
3. Configure the port types if needed, using the **connectx_port_config** script.

4.2 Configuring VMware ESX Server Settings

VMware ESX 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 ESX 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 ESX 4.1 use **esxcfg-module -i <module name>** for viewing all the available module parameters and default settings.

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 ESX.

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-ESX 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 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 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 ESX 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 NetQueue

In VMware vSphere 4, NetQueue is enabled by default. It is also enabled by default in IPoIB driver (and mlx4_en 10GigE driver).

There is no need to configure the driver or the ESX/ESXi in order to take advantage of NetQueue technology.

4.2.4 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)

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.

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:

- 64 for the InfiniHost™ III family
- 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, run:

  ```
  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 *ESX Server 3 Configuration Guide* and *ESX Server 3 802.1Q VLAN Solutions* documents.

4.2.5 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 on VI3 and 4092 bytes on vSphere 4. 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 4)

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:

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 **esxcfg-module -s 'set_4k_mtu=1' mlx4_en**. Changes will take effect after the reboot.

4.2.6 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 “NetQueue” on page 22. and to use Jumbo Frames as explained in section “Maximum Transmit Unit (MTU) Configuration” on page 23.

IPoIB can create up to 8 nics over each physical port. To create 8 uplinks over each physical port run:

```
esxcfg-module -s 'ipoib_uplink_num=8 ipoib_mac_type=1' ib_ipoib
```

The change will take effect after the reboot

You must change the mac type, to add **ipoib_uplink_num.**

For further information on VMware performance, see http://www.vmware.com/resources/techresources/10161

4.2.7 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 Configuration Guide* document.