



***InfiniBand OFED Driver for
VMware Virtual Infrastructure (VI) 3.5
and VMware vSphere 4.0***

***Installation Guide
Rev. 1.30***



NOTE:

THIS INFORMATION IS PROVIDED BY MELLANOX FOR INFORMATIONAL PURPOSES ONLY AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, 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 OUT OF THE USE OF THIS HARDWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.



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

© Copyright 2010. Mellanox Technologies, Inc. All Rights Reserved.

Mellanox, BridgeX, ConnectX, InfiniBlast, InfiniBridge, InfiniHost, InfiniRISC, InfiniScale, InfiniPCI, and Virtual Protocol Interconnect are registered trademarks of Mellanox Technologies, Ltd. CORE-Direct, FabricIT and PhyX are trademarks of Mellanox Technologies, Ltd.

All other marks and names mentioned herein may be trademarks of their respective companies.



Table of Contents

Revision History	4
1 Introduction.....	5
2 Hardware Support	11
3 Software Support	12
3.1 User Layer Protocols (ULPs) Support	12
3.2 Tools	12
4 Setup and Configuration	13
4.1 Installation on VMware VI 3	13
4.2 Installation on VMware vSphere 4.....	13
4.2.1 ESX.....	13
4.2.2 ESXi.....	14
4.2.3 VPI – Multi-protocol Support.....	15
4.3 Usage.....	16
4.4 Configuration	17
4.4.1 Subnet Manager	17
4.4.2 Networking	17
4.4.3 Storage	23
4.4.4 Performance	24
4.4.5 High Availability	24
5 Firmware Burning.....	25
6 Useful Links.....	26
7 Hardware Compatibility List	27

List of Tables

Table 1- History Table	4
Table 2- Supported Mellanox Technologies HCA Devices	11

Table of Figures

Figure 1 - Typical Deployment Configuration	5
Figure 2 – Configuration With InfiniBand Adapter Deployed.....	6
Figure 3 – InfiniBand Network Adapters as Listed in VI Client (of VI 3).....	7
Figure 4 – InfiniBand Storage Adapters as Listed in VI Client (of VI 3)	7
Figure 5 – Multiple InfiniBand Network Adapters Exposed by a Single HCA	8
Figure 6 – Usage of Multiple InfiniBand Network Adapters	9
Figure 7 – Multiple InfiniBand Storage Adapters Exposed by a Single HCA	10
Figure 8 - Usage of Multiple InfiniBand Storage Adapters.....	10



Revision History

Table 1- History Table

Revision	Date	Details
1.30	January 2010	- Added vSphere 4.0
1.20	October 2008	- ConnectX PCIe Gen1/2 support - New IPoIB and SRP features - Performance enhancements - Added <i>ibstat</i> utility to package
1.1	January 2008	First GA version

1 Introduction

This document is a user's guide for installing the InfiniBand OFED Driver for VMware® Infrastructure 3, version 3.5, and VMware® vSphere version 4.0.

The InfiniBand OFED software driver package adds to VMware support for the InfiniBand interconnect technology. The driver package is based on the OpenFabrics Enterprise Distribution, OFED, 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 environment provides consolidated networking and storage services over the same InfiniBand port.

In addition, InfiniBand services support VMware 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.

Figure 1 - Typical Deployment Configuration

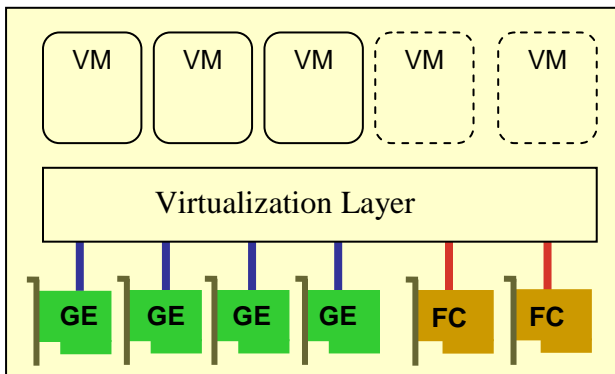
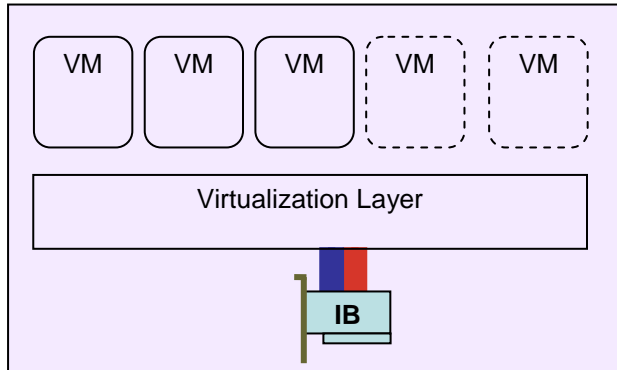


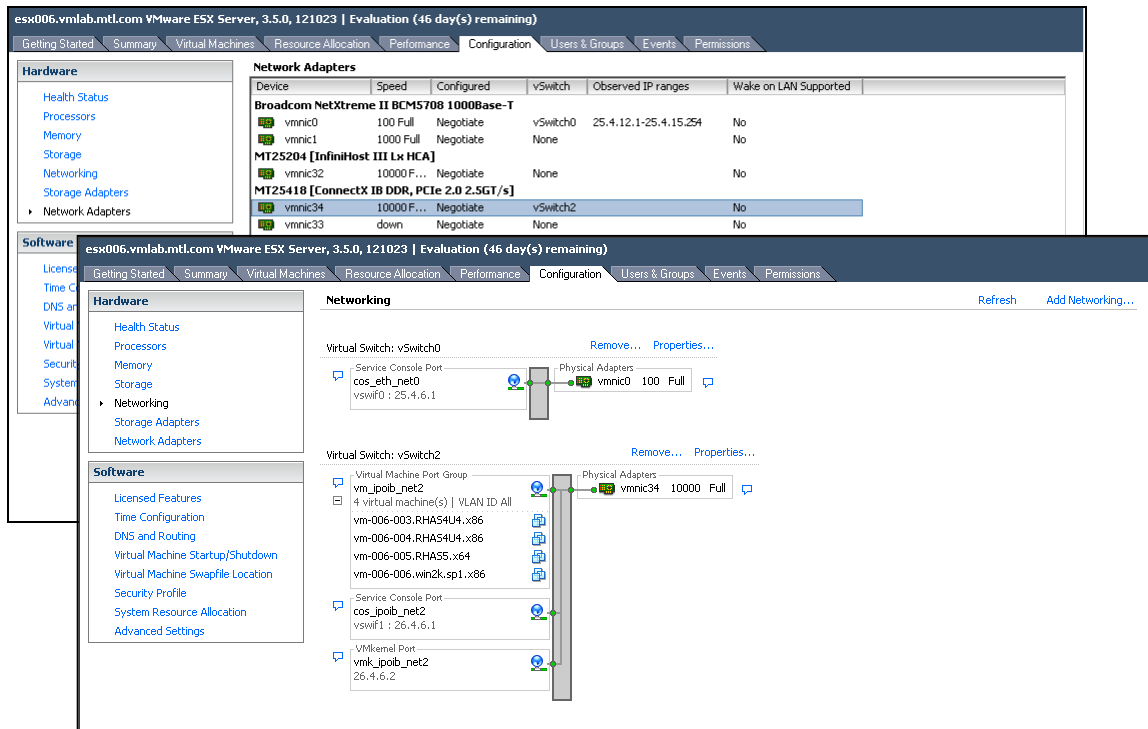
Figure 2 – Configuration With InfiniBand Adapter Deployed



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 (of VI 3)



esx006.vmlab.mtl.com VMware ESX Server, 3.5.0, 121023 | Evaluation (46 day(s) remaining)

Getting Started Summary Virtual Machines Resource Allocation Performance Configuration Users & Groups Events Permissions

Hardware

Health Status
Processors
Memory
Storage
Networking
Storage Adapters
Network Adapters

Network Adapters

Device	Speed	Configured	vSwitch	Observed IP ranges	Wake on LAN Supported
Broadcom NetXtreme II BCM5708 1000Base-T					
vmnic0	100 Full	Negotiate	vSwitch0	25.4.12.1-25.4.15.254	No
vmnic1	1000 Full	Negotiate	None		No
MT25204 [InfiniHost III Lx HCA]					
vmnic32	10000 F...	Negotiate	None		No
MT25418 [ConnectX IB DDR, PCIe 2.0 2.5GT/s]					
vmnic34	10000 F...	Negotiate	vSwitch2		No
vmnic33	down	Negotiate	None		No

esx006.vmlab.mtl.com VMware ESX Server, 3.5.0, 121023 | Evaluation (46 day(s) remaining)

Getting Started Summary Virtual Machines Resource Allocation Performance Configuration Users & Groups Events Permissions

Hardware

Health Status
Processors
Memory
Storage
Networking
Storage Adapters
Network Adapters

Software

Licensed Features
Time Configuration
DNS and Routing
Virtual Machine Startup/Shutdown
Virtual Machine Swapfile Location
Security Profile
System Resource Allocation
Advanced Settings

Networking

Virtual Switch: vSwitch0
Remove... Properties...

Service Console Port
cos_eth_net0
vswif0 : 25.4.6.1

Physical Adapters
vmnic0 100 Full

Virtual Switch: vSwitch2
Remove... Properties...

Virtual Machine Port Group
vm_jpoib_net2
4 virtual machine(s) | VLAN ID All

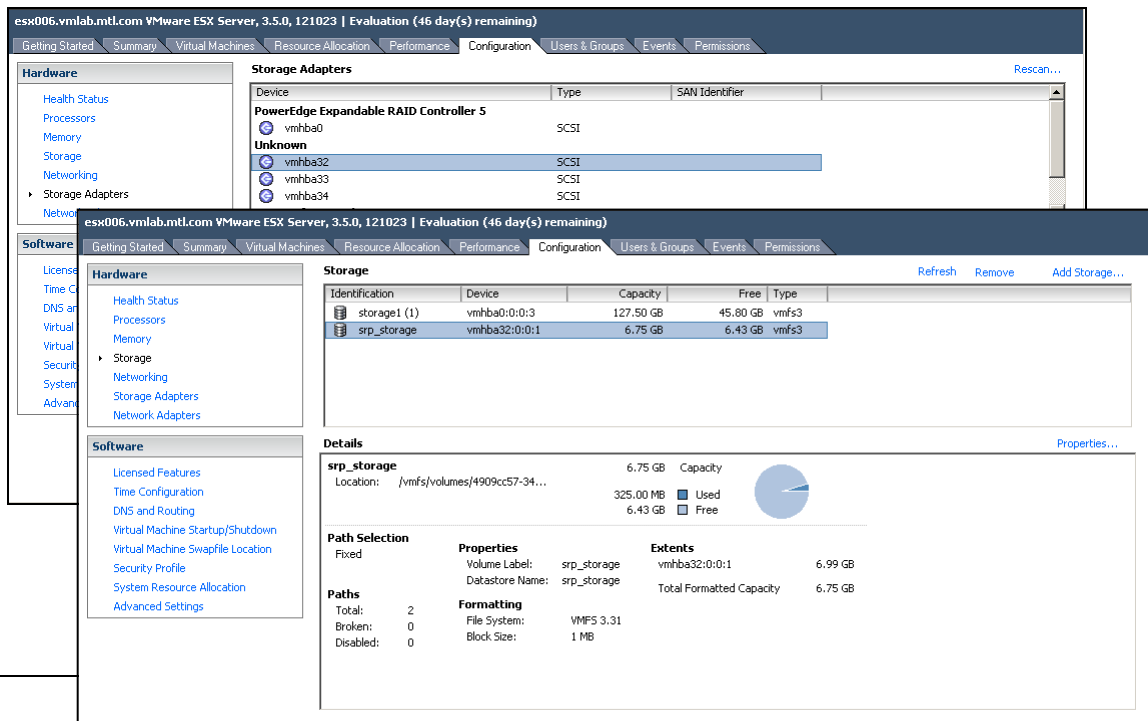
vm-006-003.RHAS4U4.x86
vm-006-004.RHAS4U4.x86
vm-006-005.RHAS5.x64
vm-006-006.win2k.sp1.x86

Service Console Port
cos_jpoib_net2
vswif1 : 26.4.6.1

VMkernel Port
vmk_jpoib_net2
26.4.6.2

Physical Adapters
vmnic34 10000 Full

Figure 4 – InfiniBand Storage Adapters as Listed in VI Client¹ (of VI 3)



esx006.vmlab.mtl.com VMware ESX Server, 3.5.0, 121023 | Evaluation (46 day(s) remaining)

Getting Started Summary Virtual Machines Resource Allocation Performance Configuration Users & Groups Events Permissions

Hardware

Health Status
Processors
Memory
Storage
Networking
Storage Adapters
Network Adapters

Storage Adapters

Device	Type	SAN Identifier
PowerEdge Expandable RAID Controller 5		
vmhba0	SCSI	
Unknown		
vmhba32	SCSI	
vmhba33	SCSI	
vmhba34	SCSI	

esx006.vmlab.mtl.com VMware ESX Server, 3.5.0, 121023 | Evaluation (46 day(s) remaining)

Getting Started Summary Virtual Machines Resource Allocation Performance Configuration Users & Groups Events Permissions

Hardware

Health Status
Processors
Memory
Storage
Networking
Storage Adapters
Network Adapters

Software

Licensed Features
Time Configuration
DNS and Routing
Virtual Machine Startup/Shutdown
Virtual Machine Swapfile Location
Security Profile
System Resource Allocation
Advanced Settings

Storage

Identification	Device	Capacity	Free	Type
storage1 (1)	vmhba0:0:0:3	127.50 GB	45.80 GB	vmfs3
srp_storage	vmhba32:0:0:1	6.75 GB	6.43 GB	vmfs3

Details

srp_storage
Location: /vmfs/volumes/4909cc57-34...
6.75 GB Capacity
325.00 MB Used
6.43 GB Free

Path Selection

Fixed

Properties
Volume Label: srp_storage
Datastore Name: srp_storage

Extents
vmhba32:0:0:1 6.99 GB
Total Formatted Capacity 6.75 GB

Paths
Total: 2
Broken: 0
Disabled: 0

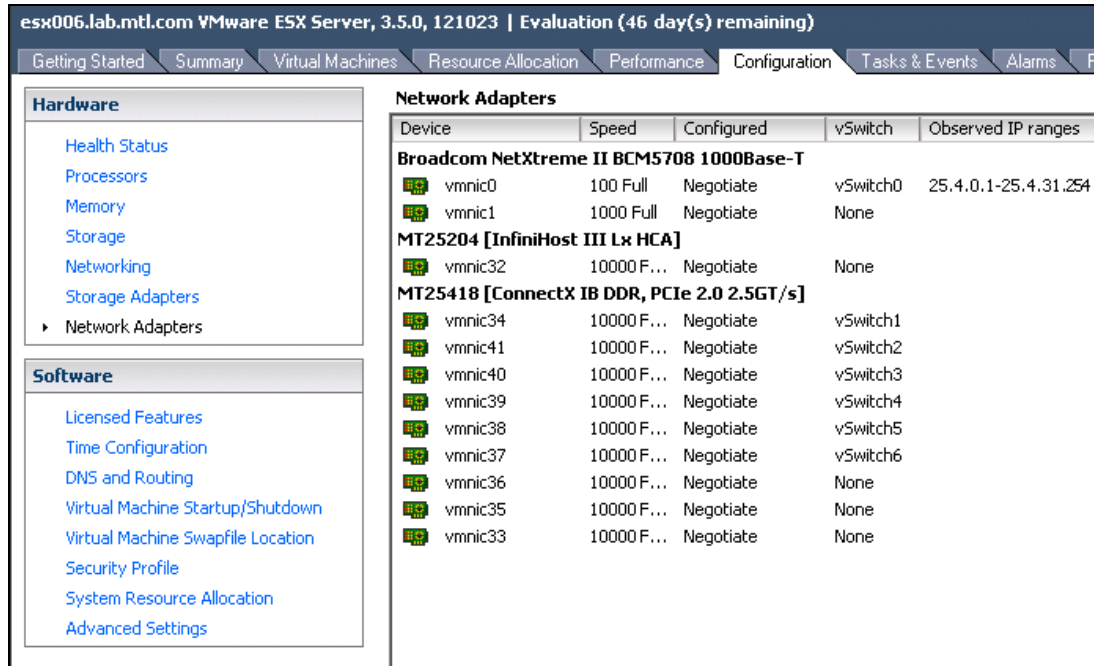
Formatting
File System: VMFS 3.31
Block Size: 1 MB

¹ The VI Client lists mistakenly InfiniBand storage adapters under “Unknown” family.

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



Device	Speed	Configured	vSwitch	Observed IP ranges
Broadcom NetXtreme II BCM5708 1000Base-T				
vmnic0	100 Full	Negotiate	vSwitch0	25.4.0.1-25.4.31.254
vmnic1	1000 Full	Negotiate	None	
MT25204 [InfiniHost III Lx HCA]				
vmnic32	10000 F...	Negotiate	None	
MT25418 [ConnectX IB DDR, PCIe 2.0 2.5GT/s]				
vmnic34	10000 F...	Negotiate	vSwitch1	
vmnic41	10000 F...	Negotiate	vSwitch2	
vmnic40	10000 F...	Negotiate	vSwitch3	
vmnic39	10000 F...	Negotiate	vSwitch4	
vmnic38	10000 F...	Negotiate	vSwitch5	
vmnic37	10000 F...	Negotiate	vSwitch6	
vmnic36	10000 F...	Negotiate	None	
vmnic35	10000 F...	Negotiate	None	
vmnic33	10000 F...	Negotiate	None	

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 Adapters²

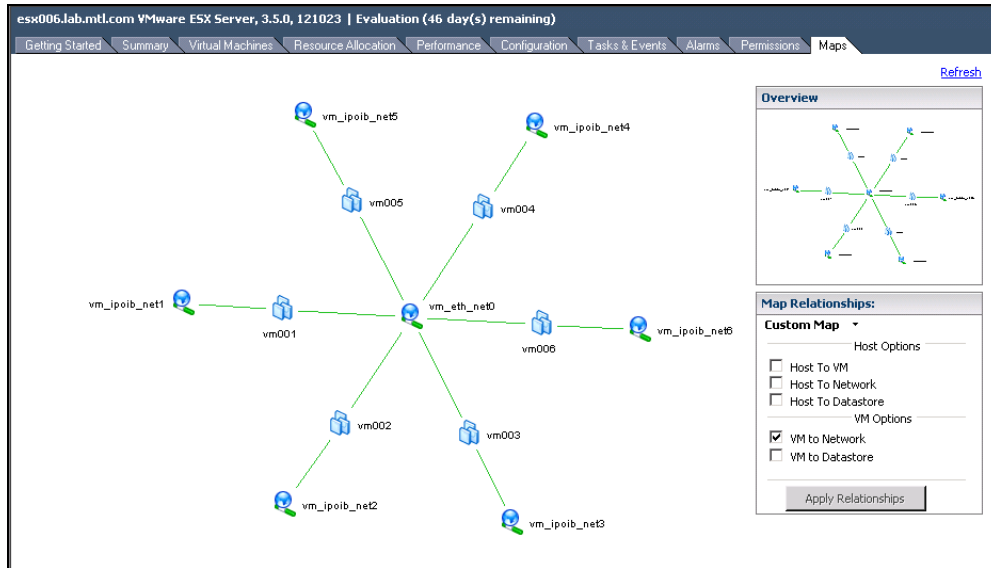
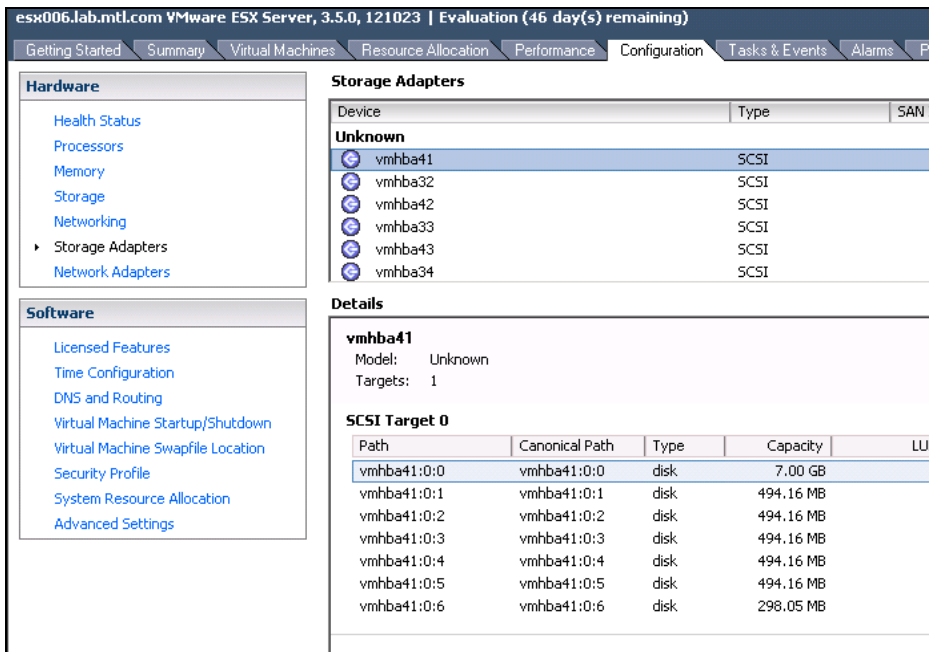


Figure 7 shows multiple InfiniBand storage adapters exposed by a single HCA (MT25418 ConnectX IB DDR Dual Port HCA).

² The maps tab is available in VI3 only

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



The screenshot shows the VMware vSphere Configuration page for a host. The left sidebar has a tree view with 'Storage Adapters' selected. The main area is divided into 'Storage Adapters' and 'Details'.

Storage Adapters Table:

Device	Type	SAN ID
Unknown		
vmhba41	SCSI	
vmhba32	SCSI	
vmhba42	SCSI	
vmhba33	SCSI	
vmhba43	SCSI	
vmhba34	SCSI	

Details for vmhba41:

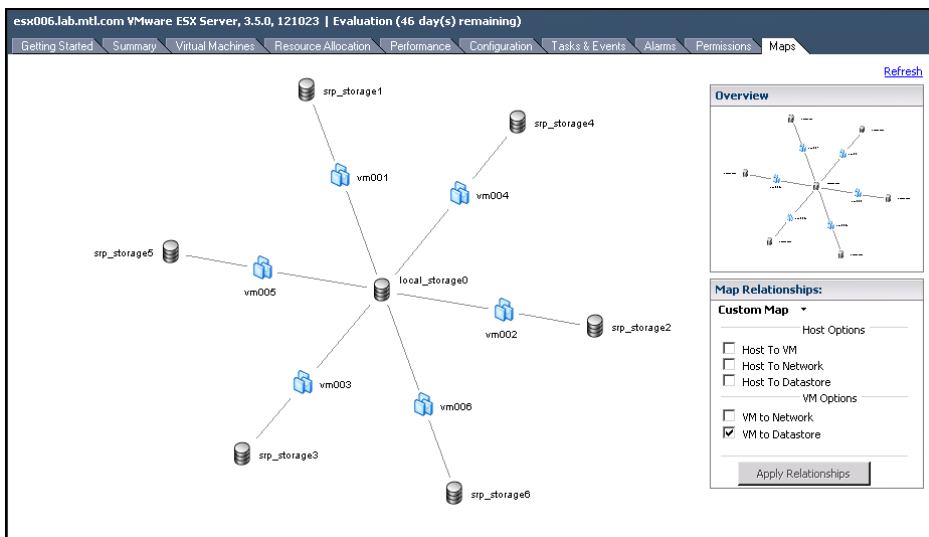
Model: Unknown
Targets: 1

SCSI Target 0 Table:

Path	Canonical Path	Type	Capacity	LUN
vmhba41:0:0	vmhba41:0:0	disk	7.00 GB	
vmhba41:0:1	vmhba41:0:1	disk	494.16 MB	
vmhba41:0:2	vmhba41:0:2	disk	494.16 MB	
vmhba41:0:3	vmhba41:0:3	disk	494.16 MB	
vmhba41:0:4	vmhba41:0:4	disk	494.16 MB	
vmhba41:0:5	vmhba41:0:5	disk	494.16 MB	
vmhba41:0:6	vmhba41:0:6	disk	298.05 MB	

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 Adapters³



Instructions on how to enable and configure this feature are available under the section Setup and Configuration on page 13.

³ The Maps tab is available in VI 3.5 only



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

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



3 Software Support

The InfiniBand OFED driver package for VMware Infrastructure 3 is based on the OpenFabrics Enterprise Distribution, OFED 1.3.1.

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>.

3.1 User Layer Protocols (ULPs) Support

The following ULPs are supported:

- IP over InfiniBand (IPoIB)⁴
- 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 4.3.

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

⁴ IPoIB driver supports Unreliable Datagram (UD) mode only.

4 Setup and Configuration

4.1 Installation on VMware VI 3

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

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 Configuration Guide* document (see "Useful Links" on page 26).

4.2 Installation on VMware vSphere 4

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

4.2.1 ESX

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

Note: The VIB bundles enforce package dependencies. This forces the following package installation sequence: mlx4_en -> ib_basic -> ib_ulps.

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 and to retrieve the bundle id, run:

```
cos# esxupdate query
```

To uninstall an old driver package, run:

```
cos# esxupdate -b <bundle id to remove> remove
```

2. Install the `mlx4_en` driver by running:

```
cos# esxupdate --bundle <mlx4_en VIB bundle full path> --  
maintenancemode --nosigcheck update
```

Note: When a certified driver CD is used for `mlx4_en`, the "--nosigcheck" parameter is not required.

3. Install the `ib_basic` driver package by running:

```
cos# esxupdate --bundle <ib_basic full path> --maintenancemode  
update
```

4. Install the ULP driver package by running:

```
cos# esxupdate --bundle <ib_ulp full path> --maintenancemode  
update
```

5. Enable / Disable ULP modules by running:

To enable: `cos# esxcfg-module -e <module name>`

To disable: `cos# esxcfg-module -d <module name>`

where module name is `ib_ipoib`, `ib_srp`

Note: If `mlx4_en` or `ib_basic` are disabled, `ib_ipoib` and `ib_srp` will not load.

6. Configure VPI functionality if needed.
7. Reboot ESX/ESXi server.

4.2.2 ESXi

Install vSphere remote CLI. The package is available stand-alone or as part of the vSphere Management Assistant (vMA) virtual appliance from <http://www.vmware.com>. Refer to VMware's document "vSphere Command-Line Interface Installation and Reference Guide" for detailed instructions on how to use them.

Use a remote CLI to run the following commands:

Note: The VIB bundles enforce package dependencies. This forces the following package installation sequence: `mlx4_en -> ib_basic -> ib_ulps`.

1. If you have a previous version of the driver package installed on your system, then you must uninstall it first. To see all packages installed on your machine and to retrieve the bundle id, run:

```
rcli# vihostupdate --server <server ip> --query
```

To uninstall an old driver package, run:

```
rcli# esxupdate --server <server ip> --bulletin <bulletin id>  
--remove  
reboot ESXi
```

2. Install the `mlx4_en` driver by running:

```
rcli# vihostupdate --server <server ip> --bundle <mlx4_en VIB>
--install --nosigcheck
```

3. Install the `ib_basic` driver package by running:

```
RCLI#> vihostupdate --server <server ip> --bundle <ib_basic
VIB> --install
reboot ESXi
```

4. Install the ULP driver package by running:

```
RCLI#> vihostupdate --server <server ip> --bundle <ib_ulp VIB>
--install
```

5. Configure VPI functionality if needed.

6. Reboot ESX/ESXi server.

4.2.3 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), (ETH,ETH).

Note: The configuration of port 1 ETH and port 2 IB 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 `/var/log/vmkernel`.

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

4.2.3.1 Interactive VPI port configuration

To configure the port types interactively, run `connectx_port_config` script.

`connectx_port_config` is installed by the `ib_basic` package, and is only available for ESX (and not ESXi).

```
cos# connectx_port_config
```

4.2.3.2 Manual port type configuration

1. Manual port configuration is useful for ESXi and for automated installations.
2. To configure all ports to IB, set `mlx4_en` module parameter “`port_type_default`” to 1 (IB). This will default all ports that are not specifically configured (as explained below) to IB. To set default port type to IB run:

On ESX:

```
cos# esxcfg-module -s "port_type_default=1" mlx4_en
```

On ESXi:

```
rcli# vicfg-module.pl --server <ip> -s  
"port_type_default=1" mlx4_en
```

3. If you have other module parameters in use with `mlx4_en`, make sure to include them in the configuration command.
4. To configure a specific port to IB or Eth, set the `mlx4_en` module parameter `"port_types"` to the required vector. The `"port_types"` vector is made of a set of triplets of the form: `pci_id,port_num,type`. For example:
 - a. To configure port 1 to IB (type 1) and port 2 to Eth (type 2) of a two port ConnectX on pci bus 00d:00.0 will require:

For ESX:

```
COS#> esxcfg-module -s "port_types=13,1,1,13,2,2" mlx4_en
```

For ESXi:

```
RCLI#> vicfg-module.pl --server <ip> -s  
"port_types=13,1,1,13,2,2" mlx4_en
```

- b. To configure ports 1 and 2 of a dual-port ConnectX device to IB (type 1) on pci bus 009:00.0 or port 1 to IB and port 2 to Eth on pci bus 011:00.0, run:

```
COS#> esxcfg-module -s  
"port_types=9,1,1,9,2,1,17,1,1,17,2,2" mlx4_en
```

5. The pci bus id of the ConnectX devices can be retrieved from the interface names of the uplinks. Go to configuration->network adapters in vSphere client. The interface name is of the format `vmnicX.pY`, where X is the pci bus device id, and Y is the port number. For example: `Vmnic11.p1` is installed on pci bus id 11 (meaning 00b:00.0). The value of 11 should be used as the first value in the VPI configuration triplet.

4.3 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:

On VI 3:

```
cos# rpm -qi <package name>
```

On vSphere 4:

To retrieve the package bundle id run:



```
cos# esxupdate query
cos# esxupdate -b <package bundle id> info
```

You can also retrieve information on InfiniBand ports available on your machine using *ibstat*⁵. For usage and help, log into the service console and run:

```
cos# ibstat -h
```

4.4 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.

This section includes instructions for configuring various module parameters. When using ESXi, use remote CLI `vicfg-module.pl` to configure the module parameters in a similar way to what is done in the COS for ESX.

4.4.1 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 machine⁶ in the subnet. You can download an InfiniBand SM such as OpenSM from www.openfabrics.org under the Downloads section.

4.4.2 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.

⁵ *ibstat* of this package produces similar output to *ibstat* of the Linux OFED package.

⁶ 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 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.

4.4.2.1 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.

On ESXi use remote CLI’s equivalent utility *vicfg-module* 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
```

For VI 3 only:

Note: 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 order⁷. For example, to change the receive ring size of the first two InfiniBand ports,

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

⁷ The PCI bus number is retrieved using the *lspci* command from the service console.



4.4.2.2 Multiple Network Adapters

VMware Infrastructure 3:

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 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
```

VMware vSphere 4:

IPoIB has the ability to expose multiple network adapters over a single InfiniBand port. Each port can expose up to 8 network adapters, and each ESX Server machine can expose up to 16 network adapters. By default, one network adapter is configured for each physical InfiniBand port. This setting is determined by a module parameter (called ipoib_uplink_num), which is passed to the IPoIB module. To change it, log into the service console and run:

```
cos# esxcfg-module ib_ipoib -s 'ipoib_uplink_num=n'  
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
```

4.4.2.3 NetQueue

VMware Infrastructure 3:

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 “vnicX”.

```
cos# cat /proc/net/ipoib/vnicX/NetQueue
```

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

```
cos# cat /proc/net/ipoib/vnic32/NetQueue  
cos# 1
```

VMware vSphere 4:

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.

4.4.2.4 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)

Note: 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
 - 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. See the *ESX Server 3 Configuration Guide* and *ESX Server 3 802.1Q VLAN Solutions* documents.

4.4.2.5 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 (VI3 / 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:

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 4K MTUs on a default PKey using the OpenSM SM8, log into the Linux machine (running OpenSM) and perform the following commands:

⁸ Version 3.1

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`

4.4.3 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.

4.4.3.1 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.

1. To disable the SRP module run:
`cos# esxcfg-module ib_srp -d`
2. 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`

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'
```

4.4.3.2 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
```

4.4.4 Performance

For best performance, it is recommended to use ConnectX InfiniBand cards⁹ since they have enhanced capabilities and offloading features.

To enhance networking performance, it is recommended to enable NetQueue as explained in Section 4.4.2.3 and to use Jumbo Frames as explained in Section 4.4.2.5.

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

4.4.5 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 Configuration Guide* document.

⁹ You can identify which InfiniBand card is installed on your machine using the `lspci` command from the service console.



5 Firmware Burning

Use the MFT for VMware ESX Server 3.5 / vSphere 4 tools package to burn a firmware image onto the Flash of an InfiniBand device. Download the binaries and the documents from:
http://www.mellanox.com/products/management_tools.php

6 Useful Links

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

- Mellanox Technologies InfiniBand products - <http://www.mellanox.com/products>
- OpenFabrics Alliance - <http://www.openfabrics.org>
- VMware product information - <http://www.vmware.com/products/>
- General documentation - <http://www.vmware.com/support/pubs>
- ESX Server 3 Configuration Guide - http://www.vmware.com/support/pubs/vi_pages/vi_pubs_35.html
- ESX Server 3: 802.1Q VLAN Solutions - <http://www.vmware.com/resources/techresources/412>
- 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 11