Configuring Mellanox Hardware for VPI Operation Application Note

Rev 1.2

www.mellanox.com

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

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

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

Table of Contents

Document Revision History

The following table presents the revision history of this document:

Table 1 - Document Revision History

Document Revision	Date	Changes
Rev 1.2	Oct. 2013	Added gateway functionality aligned with MLNX-OS 3.3.4100 release.
Rev 1.1	Mar 2013	Added gateway configurations.
Rev 1.0	Jun 2012	Initial release.

About this Manual

This manual provides information on basic configuration of the converged VPI networks.

Intended Audience

This manual is intended for network administrators who wish to build an infrastructure for converged VPI networks.

Related Documents

The following table provides a list of documents related to this application note.

Table 2 - Related Documents

Documents	Location
MLNX-OS® Management Software documentation suite (User Manual, Command Reference Guide, Release Notes)	support.mellanox.com > Software & Drivers > Management Software > MLNX-OS > Select the switch system in your possession and download its management manual
Mellanox OFED for Linux User Manual	www.mellanox.com > Products > InfiniBand/VPI drivers> Linux SW/Drivers
Mellanox OFED for Windows (WinOF) User Manual	www.mellanox.com > Products > InfiniBand/VPI drivers> Windows SW/Drivers
Mellanox Approved Cable List	Refer to: http://www.mellanox.com/related-docs/user_manuals/Mellanox_approved_cables.pdf

1 Introduction to VPI

Products based on the Mellanox ConnectX® family of adapters and Mellanox SwitchX® family of switches support Mellanox's proprietary Virtual Protocol Interconnect. Virtual Protocol Interconnect allows InfiniBand and Ethernet traffic to co-exist on one platform. Each port can operate independently as an InfiniBand link or as an Ethernet link. ConnectX® adapters can be configured to use InfinBand, Ethernet or to auto-sense the fab-ric through the port itself. In addition, a gateway can be added between Ethernet and InfiniBand subnets to link servers on the same subnet.

This application note describes how to configure Mellanox products for VPI operation and gate-way.

Figure 1: Hybrid Cluster - Ethernet and InfiniBand Links

1.1 VPI on Mellanox Adapters

ConnectX® adapter family ports can be configured with the following protocol link types:

- Ethernet (eth) – sets the port as an Ethernet link
- InfiniBand (ib) – sets the port as an InfiniBand link
- Auto Sensing (auto) – in this mode the port detects the port type based on the attached network type.

Setting the link protocol on ports of the ConnectX® adapter family can be performed using the Mellanox OFED (Mellanox OFED for Linux - MLNX_OFED) or Mellanox WinOF (Mellanox OFED for Windows) driver stack.

For easy migration from InfiniBand to Ethernet and vice versa, configure your adapter cards to Auto Sensing mode. Auto Sensing mode enables the adapter card to automatically sense the link type (InfiniBand or Ethernet) based on the cable connected to the port and load the appropriate driver stack (InfiniBand or Ethernet). If no link is detected, the driver retries link sensing every few seconds. For example, if the first port is connected to an InfiniBand switch and the second to an Ethernet switch, the adapter card automatically loads the first port as InfiniBand and the sec-ond as Ethernet.

In order to set the link protocol for adapters running in a Linux environment, refer to Section 3.2.1, "Setting the Link Protocol for Linux Adapters (Mellanox OFED)," on page 9.

In order to set the link protocol for adapters in a Windows environment, refer to Section 3.2.2, "Setting the Link Protocol for Windows Drivers (Mellanox WinOF)," on page 11.

1.2 VPI on Mellanox Switch Systems

VPI technology on Mellanox switch systems can be achieved in different levels:

- **System level VPI** – it can be decided, per system, whether to use InfiniBand or Ethernet for all the interfaces in the system. Either Ethernet switch or InfiniBand switch profile can be configured on the system in order to determine the running link protocol for all the system ports.
- **Interface level VPI** – it can be decided, per system port, whether to use InfiniBand or Ethernet as a link protocol. A single VPI system profile can be configured and, per port, the link protocol may be selected. Configuring the switch to VPI mode allows splitting the hardware into two separate switches (an Ethernet switch and an InfiniBand switch). Traffic does not pass between those switches. While configuring the VPI system profile, bridging (or gateway) capabilities can be added to pass traffic from the Ethernet to the InfiniBand hosts.

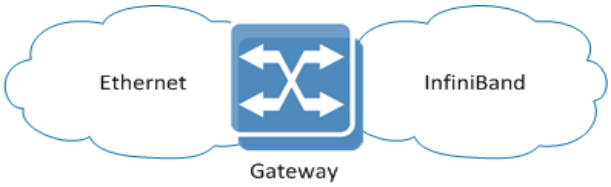
2 Gateway

In cases where the network consists of two types of link protocols (Ethernet and InfiniBand), Proxy-ARP can be used to forward IPv4 packets from the Ethernet network to the InfiniBand network and vice versa. Proxy-ARP is not an IP Router, but acts as a bridge that forwards the IPoETH packets to IPoIB in Unreliable Datagram (UD).

The Proxy-ARP forwards the traffic in a single subnet.

IP Routing, InfiniBand SM and IGMP snooping must be disabled in order to enable the Proxy- ARP.

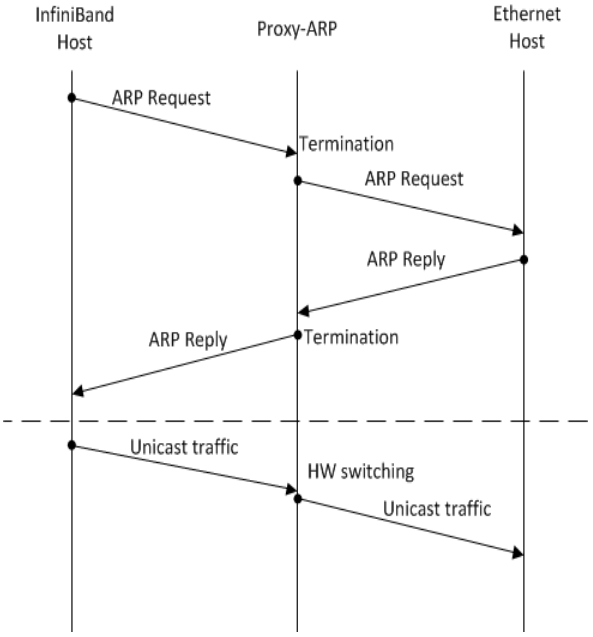
Figure 2: Gateway



2.1 ARP Flow

An ARP request sent from an InfiniBand host is terminated at the Proxy-ARP. Then a new ARP request is generated and sent on the VLAN interface to reach the Ethernet host. The Ethernet host responds with an ARP reply to the Proxy-ARP. The Proxy-ARP then terminates it and generates a new ARP reply to the InfiniBand host. Once the destination address has been resolved, unicast traffic is passed from the InfiniBand host to the Ethernet host. The process is similar in the opposite direction (Ethernet to InfiniBand).

Figure 3: ARP Flow



3 Configuring VPI

3.1 VPI Capable Network

The following issues should be considered when setting up your VPI network:

- For migration between Ethernet and InfiniBand switching, only SwitchX® based systems equipped with QSFP ports and the VPI capability should be used. For example, SX1036 and SX6036T/F can be selected as the switches in your network. The switch profile can be changed from Ethernet to InfiniBand, from InfiniBand to Ethernet, or to VPI through simple commands after a license upgrade.
- To obtain the 56GbE/FDR VPI capability, select the following systems in your network:
  - SX6012F, SX6012F, SX6036F, SX6036G
  - SX1012, SX1036
- ConnectX®-2 and ConnectX-3 VPI network adapter cards should be configured to Auto Sensing mode. After changing the switch profile from Ethernet to InfiniBand, the network adapters automatically change their link protocol type (may require reset).
- When connecting two network adapters configured to Auto Sensing which are connected back-to-back, their link protocol becomes InfiniBand by default.

3.2 Setting the Link Protocol for Adapters

Table 3 shows the supported port protocol configurations for a dual-port adapter.

Table 3 - Supported Port Protocol Configurations on Dual Port Adapters

Port 1 Configuration	Port 2 Configuration
ib	ib
ib	eth
eth	eth

The configuration “Port1 = eth, Port2 = ib” is not supported.

3.2.1 Setting the Link Protocol for Linux Adapters (Mellanox OFED)

Setting the link protocol on the ports of the ConnectX® adapter family can be performed using the MLNX_OFED driver stack. By default, ConnectX® adapter ports are initialized as InfiniBand ports. In order to change the link protocol type use the `connectx_port_config` script after the driver is loaded.

Step 1. Display the current port configuration for all adapter devices' ports by running `/sbin/connectx_port_config --show`.

```
(host)# /sbin/connectx_port_config --show
-------------------------------
Port configuration for PCI device: 0000:1f:00.0 is:
ib
ib
-------------------------------
(host)#
```

Step 2. Change the link protocol type by running `/sbin/connectx_port_config`.

```
(host)# connectx_port_config

ConnectX PCI devices :
|----------------------------|
| 1            0000:1f:00.0 |
|----------------------------|
t
Before port change:
ib
ib

|----------------------------|
| Possible port modes:       |
| 1: Infiniband              |
| 2: Ethernet                |
| 3: AutoSense               |
|----------------------------|
Select mode for port 1 (1,2,3): 3
Select mode for port 2 (1,2,3): 3

After port change:
auto (ib)
auto (ib)
(host)#
```

Step 3. Display the new port configuration for all adapter devices' ports by running `/sbin/connectx_port_config --show`.

In addition, port configuration is saved in the file: `/etc/infiniband/connectx.conf`. This saved configuration is restored at driver restart only if restarting via the command `/etc/init.d/openibd restart`.

Step 4. Use the `--help` flag to get additional script options.

```
(host)# connectx_port_config --help
Usage:
/sbin/connectx_port_config
/sbin/connectx_port_config -s|--show
/sbin/connectx_port_config -h|--help
/sbin/connectx_port_config [-d|--device <PCI device id>] -c|--conf <port1,port2>
Possible port configurations:
        eth,eth
        eth,ib
        eth,auto
        ib,ib
        ib,eth
        ib,auto
        auto,auto
        auto,eth
(host)#
```

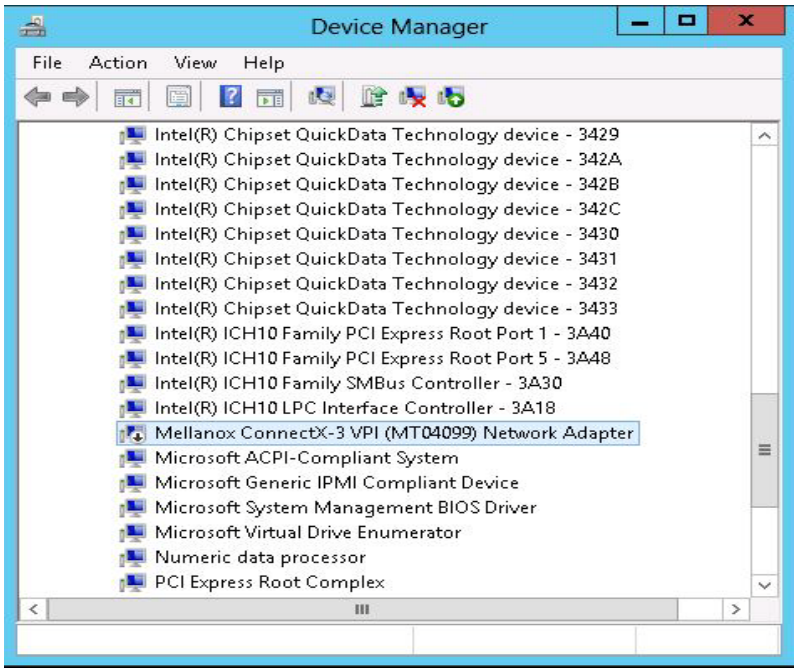
3.2.2 Setting the Link Protocol for Windows Drivers (Mellanox WinOF)

Setting the link protocol on the ports of the ConnectX® adapter family may be performed using the WinOF driver stack.

For Mellanox WinOF, Auto Sensing is performed only when rebooting the machine or after disabling or enabling the mlx4_bus interface from the Device Manager. Hence, if you replace cables during the runtime or change the link protocol of the switch, the adapter does not perform Auto Sensing.

➢ ***To configure ports:***

Step 1. Right-click the Mellanox ConnectX® VPI network adapter under System devices and left-click Properties.



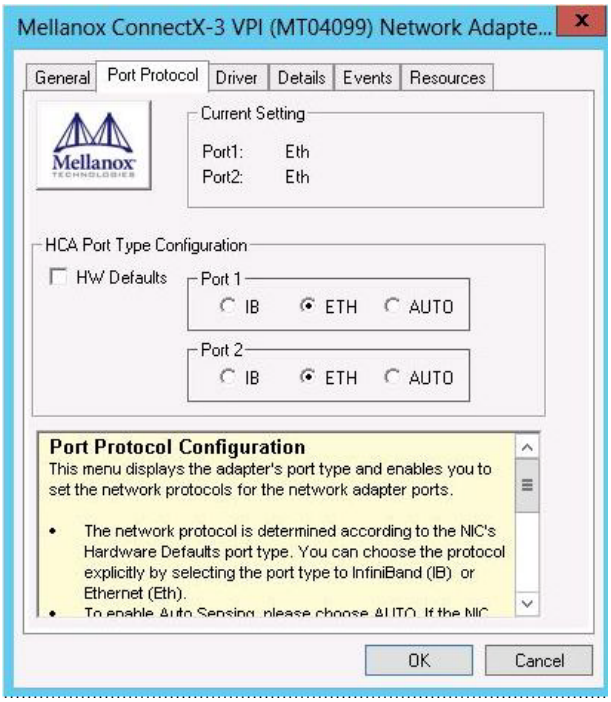
Step 2. Select the Port Protocol tab from the Properties sheet and enable Auto Sensing.

•Uncheck the hardware defaults checkbox.

•Enable Auto Sensing by checking the AUTO check-box for the desired port.

The "Port Protocol" tab is displayed only if the adapter is a VPI.

The figure below is an example of the displayed Port Protocol sheet for a dual port VPI adapter card.



3.3 Configuring VPI on Mellanox Switch Systems

Configuring your system to VPI single-switch mode splits your network interfaces to two groups:

- The Ethernet set of ports, which are connected to the Ethernet switch
- The InfiniBand set of ports, which are connected to the InfiniBand switch.

VPI single switch profile is not a gateway. Ethernet traffic does not pass to the Infini-Band ports and vice versa.

VPI mode requires using either a SX6036G system, or installing a license (UPGR-XXXX-GW) on SX1012, SX1036, SX6012, SX6018, and SX6036.
Refer MLNX-OS User Manual for more details on the licenses.

In order to set your system to work with VPI, the system profile should be changed to “vpi-single-switch”. In addition, the required set of ports should be changed from InfiniBand to Ethernet or vice versa.

The following systems can be configured as VPI switches:

- SX1012, SX1036
- SX6012, SX6018, SX6036, SX6036G

The SX6036G system supports VPI by default, with the port configured as follows:
- Interfaces 1/1-1/8 Ethernet
- Interfaces 1/9-1/36 InfiniBand

➢ ***To make the SX1012 or SX1036 system support VPI in a single-switch mode:***

Step 1. Make sure you have the latest software version installed.

Step 2. Install a gateway license.

Step 3. Set the system profile to be “vpi-single-switch”.

Step 4. Use the `port type force` command to change the disabled ports from Ethernet to InfiniBand.

This step may take several minutes.

```
switch (config)# license install <license>
switch (config)# system profile vpi-single-switch
...
switch (config)# port 1/9-1/36 type infiniband force
switch (config)# show ports type
Ethernet: 1/1, 1/2, … 1/8
Infiniband: 1/9, 1/10 … 1/36
switch (config) #
```

➢ ***To make the SX6012, SX6018 or SX6036 systems support VPI in a single-switch mode:***

Step 1. Make sure you have the latest software version installed

Step 2. Install a gateway license.

Step 3. Set the system profile to be “vpi-single-switch”.

Step 4. Use the `port type force` command to change the disabled ports from InfiniBand to Ethernet.

```
switch (config)# license install <license>
switch (config)# system profile vpi-single-switch
...
switch (config)# port 1/1-1/8 type ethernet force
switch (config)# interface ethernet 1/1-1/8 no shutdown
switch (config)# show ports type
Ethernet: 1/1, 1/2, … 1/8
Infiniband: 1/9, 1/10 … 1/36
switch (config) #
```

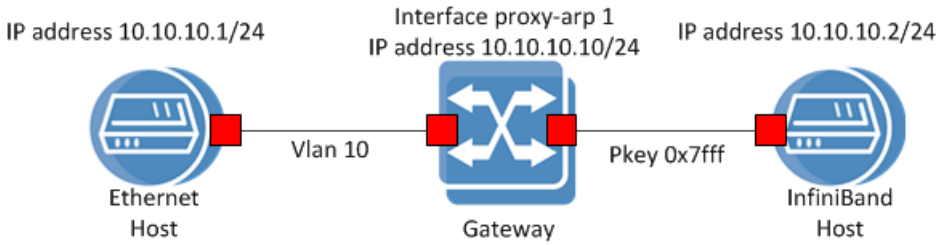
Changing the system profile will delete all the existing switch configurations and reboot the system. Management connectivity, however, will be kept.

4 Configuring Gateway

This section provides a basic example setup for a gateway configuration.

4.1 Proxy-ARP Configuration

Figure 4: Basic Gateway Setup



4.1.1 Prerequisites

Before trying to configure a Proxy-ARP in the system make sure the following conditions are met:

- Gateway license is installed (UPGR-XXXX-GW) on the switch. Run the command `show system capabilities` to verify that.:

```
switch (config)# show system capabilities
IB: Supported
Ethernet: L3
GW: Supported
Max SM nodes:648
Ethernet Max licensed speed: 40Gbps
IB max licensed speed: FDR
switch (config)#
```

SX6036G does not require a license.

- The system profile is vpi-single-switch. Run the command `show system profile` to verify that.
- InfiniBand and Ethernet interfaces are mapped on the system. Run the command `show ports type` to verify that.
- IP routing is disabled. To disable it run:

```
switch (config)# no ip routing
```

- IGMP snooping is disabled. To disable it run:

```
switch (config)# no ip igmp snooping
```

- InfiniBand SM is disabled. To disable it run:

```
switch (config)# no ib sm
```

4.1.2 Configuring Proxy-ARP

➢ ***To configure Proxy-ARP in the system:***

Step 1. Make sure the prerequisites conditions are met. Verify that gateway is supported as part of the system capabilities. Run:

```
switch (config)# show system capabilities
IB: Supported
Ethernet: L3
GW: Supported
Max SM nodes:648
Ethernet Max licensed speed: 40Gbps
IB max licensed speed: FDR
switch (config)#
```

Step 2. Enable Proxy-ARP. Run:

```
switch (config)# ip proxy-arp
switch (config)# show ip proxy-arp
Proxy-app: enabled
switch (config)#
```

Step 3. Create a Proxy-ARP interface. Run:

```
switch (config)# interface proxy-arp 1
switch (config interface proxy-arp 1)#
```

Step 4. Set an IP address and network mask to the Proxy-ARP interface. Run:

```
switch (config interface proxy-arp 1)# ip address 10.10.10.10
switch (config interface proxy-arp 1)# ip netmask /24
```

Step 5. Create a VLAN. Run:

```
switch (config)# vlan 10
switch (config vlan 10)#
```

Step 6. Add a VLAN to the interface. Run:

```
switch (config interface proxy-arp 1)# ip vlan 10
switch (config interface proxy-arp 1)#
```

Step 7. Add a PKEY to the interface. Run:

```
switch (config interface proxy-arp 1)# ip pkey 0x7fff
switch (config interface proxy-arp 1)#
```

Step 8. Enable the Proxy-ARP interface. Run:

```
switch (config interface proxy-arp 1)# no shutdown
```

Make sure one of the Ethernet or port-channel ports are configured with VLAN 10. For example:

```
switch (config interface ethernet 1/1)# switchport access vlan 10
switch (config interface ethernet 1/1)#
```

Step 9. (Optional) Configure a route to the default gateway in the subnet. Run:

```
switch (config interface proxy-arp 1)# ip route default 10.10.10.254
```

The default gateway configuration is not used for management purposes (mgmt0).

4.1.3 Verifying Proxy-ARP Configuration

➢ ***To verify the Proxy-ARP configuration:***

Step 1. Display the Proxy-ARP interface configuration. Run:

```
switch (config)# show interfaces proxy-arp 1
Proxy-arp 1
  Admin state: Enabled
  Operational state: Up
  GUID: 00:02:C9:03:00:66:08:63
  Internet Address: 10.10.10.10/24
  Broadcast Address: 10.10.10.255
  Description: N/A
  MTU: 1500
  Slowpath: Disabled
  Counters: Disabled
  Member interfaces: vlan 10, pkey 0x7fff
switch (config)#
```

Step 2. Display the Proxy-ARP brief status. Run:

```
switch (config)# show interfaces proxy-arp brief
Interface   Description       State Member interfaces
---------   -----------       ----- -----------------
Proxy-arp 1 N/A               Up    vlan 10, pkey 0x7fff
switch (config)#
```

Step 3. Display the routing table. Run:

```
switch (config) # show ip route interface proxy-arp 1
Destination Mask          Gateway      Interface   Source   Distance/Metric
10.10.10.0  255.255.255.0 0.0.0.0      proxy-arp 1 kernel   0/0
default     0.0.0.0       10.10.10.254 proxy-arp 1 static   0/0
switch (config) #
```

4.2 Advanced Settings

4.2.1 Default Gateway

It is recommended to configure a route to the default gateway in the subnet. If the default gateway is not configured, unregistered unicast traffic is dropped.

4.2.2 vTCA Interface

A virtual Target Channel Adapter, or vTCA, is an end-point of InfiniBand fabric. The gateway needs a vTCA enabled on the switch in order to function (SMA port #37).

The vTCA is active only in case the VPI single switch is configured and the Proxy ARP is enabled.

The vTCA interface is enabled by default. However, if the SM disables this interface, it can be re-enabled by running the following command:

```
switch (config)# no sma port 1 shutdown
switch (config)# show sma port 1
Enabled
switch (config)#
```

When using InfiniBand tools such as iblinkinfo, smpquery, or ibnetdiscover the user is able to see the status of the vTCA interface.

```
# iblinkinfo
...
6 37[]==(4X 14.0625 Gbps Active/LinkUp)==> 7 1[] "Mellanox vTCA switch-626a54" )...
#
```

```
# smpquery -D pi 0 1 37
Port info: DR path slid 65535; dlid 65535; 0 port 1
...
CapMask:.......................0x251486a
                               IsSM
                               IsTrapSupported
                               IsAutomaticMigrationSupported
                               IsSLMappingSupported
                               IsSystemImageGUIDsupported
                               IsExtendedSpeedsSupported
                               IsCommunicatonManagementSupported
                               IsVendorClassSupported
                               IsCapabilityMaskNoticeSupported
                               IsClientRegistrationSupported


...
...
LinkState:.....................Active
PhysLinkState:.................LinkUp
...
...
#
```

```
# ibnetdiscover
#
# Topology file: generated on Tue Jan 29 15:08:32 2013
#
# Initiated from node 0002c903003531b0 port 0002c903003531b1

...
...
Ca      1 "H-0002c903006cc4f2"          # "Mellanox vTCA switch-626a54"
[1](2c903006cc4f2)      "S-0002c903006cc4f1"[37]               # lid 7 lmc 0 "MF0;switch-
626a54:SX1036/U1" lid 6 4xFDR

vendid=0x2c9
devid=0x1003
sysimgguid=0x2c90300431cd3
caguid=0x2c90300431cd0
...
...

#
```

4.2.3 MTU

Make sure that the InfiniBand subnet MTU is similar to the Ethernet subnet MTU. In most cases the default MTU is 1500 bytes for Ethernet subnets while 4K in InfiniBand.

```
switch (config)# interface ethernet 1/1 mtu 4000
switch (config)# interface ib 1/10 mtu 4000
switch (config)# interface proxy-arp 1 mtu 4000
```

4.2.4 Slow-Path

Slow-path configuration can be used for debugging and bring-up troubleshooting. When slow-path is enabled all Proxy-ARP traffic is directed to the CPU by the switch. Then, the command `tcpdump` may be used to print the incoming packets to the terminal, allowing to see traffic types and timing.

If slow-path is enabled and traffic to CPU reaches a certain limit, packets over that limit are dropped.

```
switch (config interface proxy-arp 1)# slowpath
switch (config interface proxy-arp 1)# exit
switch (config)# tcpdump
...
switch (config)#
```