



Mellanox Technologies

Mellanox WinOF VPI Readme

Rev. 2.1.1.1



© 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.

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

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.

Table of Contents

1	Revision History	4
2	Introduction.....	4
2.1	Mellanox VPI Package Contents	4
2.2	HW and SW Requirements.....	5
2.3	Managing Firmware.....	6
2.3.1	Downloading the Firmware Tools Package.....	6
2.3.2	Download the Firmware Image of the Adapter Card.....	7
2.3.3	Updating Adapter Card Firmware	7
3	Mellanox WinOF VPI Installation Process.....	7
4	IPoIB.....	7
4.1	IPoIB Setup.....	7
4.2	Performance Remarks.....	8
4.2.1	Tunable Performance Parameters	8
4.2.2	Performance Tuning.....	9
4.2.3	MAC Generation.....	9
4.2.4	IGMP Configuration.....	11
5	OpenSM	12
6	Ethernet Driver.....	12
6.1	Overview.....	12
6.2	Performance Remarks.....	13
6.2.1	Performance Tuning.....	13
6.2.2	Known Performance Issues.....	15
6.3	Booting Windows from an iSCSI Target.....	16
6.4	Known Issues and Limitations.....	16
6.5	Troubleshooting	17
7	SDP	20
7.1	SDP Limitations.....	20
7.2	SDP Installation	20
7.3	Running Applications over SDP.....	20
7.4	Running an Application over SDP and Ethernet	21
7.5	Available Programs.....	21
7.6	Troubleshooting	22
8	WSD.....	23
8.1	Running Applications over WSD	23
8.2	Performance	23
9	SRP	23
10	Starting and Verifying the IB Fabric	24
11	Low level Performance Tests.....	24
12	Debug options	25
13	Driver Update and Uninstall Process.....	26
14	Documentation.....	26

1 Revision History

- Rev 2.1.1 May 2010 – first release
- Rev 2.1.1.1 July 14, 2010 – the InfiniHost® adapter is not supported starting with WinOF VPI v2.1.1, therefore all references to this device were removed.

2 Introduction

This is a Readme for the Mellanox WinOF VPI driver v2.1.1 package, distributed for Windows Server 2003 (x86 and x64), Windows Server 2008 (x86 and x64) and Windows Server 2008 R2.

Mellanox WinOF VPI is composed of several software modules that can be used on a computer cluster configured as an InfiniBand and/or, 10Gb/s Ethernet network.

The Mellanox WinOF VPI driver can be used in one of the following modes: 2 InfiniBand ports, 2 Ethernet ports, or 1 InfiniBand and 1 Ethernet port (that is, VPI mode).

Please refer to the `MLNX_WinOF_IB_ReleaseNotes.txt` file to check for known issues and fixed bugs for IB driver.

Please refer to the `MLNX_WinOF_ETH_ReleaseNotes.txt` file to check for known issues and fixed bugs for Ethernet driver

Note: If you plan to upgrade any previous driver or SW component on your cluster, please uninstall the previous Mellanox WinOF VPI version and install the new one on all nodes.

2.1 Mellanox VPI Package Contents

The Mellanox WinOF for Windows package contains the following components:

- Core and ULPs
 - IB network adapter cards low-level drivers (mthca, mlx4)
 - IB Access Layer (IBAL)
 - Ethernet driver (ETH)
 - Upper Layer Protocols (ULPs):
 - IP over InfiniBand (IPoIB)
 - NetworkDirect (ND)

- Winsock Direct (WSD)
 - Beta: Sockets Direct Protocol (SDP)
 - Beta: SCSI RDMA Protocol (SRP)
 - Utilities
 - OpenSM (OSM): InfiniBand Subnet Manager
 - Low level performance tests
 - vstat - get the card status
 - SdpConnect - SDP\WSD test
 - IB Diagnostics tools
 - SW Development Kit (SDK)
 - Documentation
- Note:** Core drivers, IPoIB, WSD are at GA level. SDP and SRP are now at Beta stage.

2.2 HW and SW Requirements

- Administrator privileges on your machine(s)
- Disk Space for installation: 100MB
- Supported Mellanox Technologies network adapter cards -- VPI mode:
 - ConnectX®/ConnectX®-2 IB SDR/DDR/QDR (fw-25408 Rev 2.5.700 or later)
- Supported Mellanox Technologies network adapter cards -- IB only mode:
 - ConnectX/ConnectX-2 IB SDR/DDR/QDR (fw-25408 Rev 2.5.700 or later)
 - InfiniHost® III Ex
 - MemFree: fw-25218 Rev 5.3.000 or later;
 - with memory: fw-25208 Rev 4.8.200 or later
 - InfiniHost® III Lx (fw-25204 Rev 1.2.000 or later)
- Supported Mellanox Technologies network adapter cards -- Ethernet only mode:

- ConnectX/ConnectX-2 IB SDR/DDR/QDR (fw-25408 Rev 2.5.700 or later)
- Supported PCI Device IDs: 23108, 25204, 25208, 25218, 25408, 25418, 25448, 25458, 26418, 26428, 26448 or 26458.
- For official firmware versions please see:
http://www.mellanox.com/content/pages.php?pg=firmware_download
- Supported Operating Systems and Service Packs:
 - Windows Server 2003 SP1 and SP2 (x86, x64)
 - Windows Server 2003 CCS (x64)
 - Windows Server 2008 (x86, x64, x64 R2)
 - Windows HPC Server 2008 (x64)
- Supported CPU architectures:
 - x86
 - x64

2.3 Managing Firmware

The adapter card may not have been shipped with the latest firmware version. This section describes how to update firmware.

2.3.1 Downloading the Firmware Tools Package

1. Download Mellanox Firmware Tools
Please download the current firmware tools package (MFT) from http://www.mellanox.com/content/pages.php?pg=management_tools&menu_section=34. The tools package to download is "MFT_SW for Windows" (WinMFT).
2. Install and Run WinMFT
To install the WinMFT package, double click the MSI or run it from the command prompt.

Note: On a Windows 2008 server, install the WinMFT package from the command line with administrator privileges.

Enter:

`msiexec.exe /i WinMFT_<arch>_<version>.msi`
3. Check the Device Status
To start the mst service (required by the tools), run `> sc start mst`

To check device status run `> mst status`

If no card installation problems occur, the status command should produce the following output:

- mt<device id>_pciconf0
- mt<device id>_pci_cr0

where device ID will be one of the supported PCI device IDs.

2.3.2 Download the Firmware Image of the Adapter Card

To download the correct card firmware image, please visit

http://www.mellanox.com/content/pages.php?pg=firmware_download

For help in identifying your adapter card, please visit

http://www.mellanox.com/content/pages.php?pg=firmware_HCA_FW_identification

2.3.3 Updating Adapter Card Firmware

Using a card specific binary firmware image file, enter the following command:

```
> flint -d mt<device id>_pci_cr0 -i <image_name.bin> burn
```

Note: You may need to unzip the downloaded firmware image prior to burning.

For additional details, please check the MFT user's manual under

http://www.mellanox.com/content/pages.php?pg=management_tools&menu_section=34.

3 Mellanox WinOF VPI Installation Process

Please refer to the Mellanox WinOF VPI Installation Guide for installation instructions.

4 IPoIB

IPoIB is a network driver implementation that enables transmitting IP and ARP protocol packets over an InfiniBand UD channel. The implementation conforms to the relevant IETF working group's RFCs (<http://www.ietf.org>).

4.1 IPoIB Setup

Note: You may skip this section if you have configured one of the machines as a DHCP server for the IPoIB interface.

1. Go to Control Panel
2. Double-click Network Connections
3. Select the desired adapter (from Mellanox IPoIB Adapters), then right click and select Properties
4. Choose the General tab,
5. Select Internet Protocol (TCP/IP)
6. Click Properties
7. In the Internet Protocol (TCP/IP) Properties dialog box, click "Use the following IP address"
8. Enter the appropriate IP address and Subnet Mask. Use a different IP subnet for each IB port. IB ports IP subnet addresses must be different from Ethernet subnet addresses. In most cases the first number of an IP address is a constant, therefore it is common to assign new IPoIB addresses by changing the first number. For example:
 - Host Ethernet IP: 10.2.3.4
 - IPoIB IP address: 11.2.3.4

Note: OpenSM must be active continuously on at least one machine in the cluster to allow proper IPoIB functioning.

4.2 Performance Remarks

4.2.1 Tunable Performance Parameters

The file `IPoIB_registry_values.pdf` provides the complete list of registry entries that may be added/changed by the performance tuning procedure described in [Performance Tuning](#) below.

The following is a list of key parameters for performance tuning.

- Payload MTU

The maximum available size of IPoIB transfer unit. It should be decremented by the size of an IPoIB header (=4B). For example, if the network adapter card supports a 4K MTU, the upper threshold for payload MTU is 4092B and not 4096B. A 4K MTU size also improves performance for short messages, since NDIS can coalesce a small message into a larger one.

Note: 4K MTU support is considered at beta level in the 2.1.1 release. Therefore, it is not advisable to enable both a 4K MTU and the Large Send Offload feature simultaneously (see below).

- Send and Receive checksum offload

Possible values:

- Disabled - No hardware checksum

- Enabled - Try to offload if the device supports it (default)
- Bypass - Always report success (checksum bypass)
- Large Send Offload (LSO)
 - Disables/Enables the LSO feature (if supported by HW). This feature has a positive impact on overall performance.

Note: 4K MTU support is considered at beta level in the 2.1.1 release. Therefore, it is not advisable to enable both a 4K MTU and the Large Send Offload feature simultaneously.

4.2.2 Performance Tuning

To improve performance, activate the performance tuning tool as follows:

1. Start the "Device Manager" (open a command line window and enter: devmgmt.msc).
2. Open "Network Adapters".
3. Right click the relevant IPoIB adapter and select Properties.
4. Select the "Advanced" tab
5. Modify performance parameters (properties) as desired.

4.2.3 MAC Generation

IPoIB generates MAC addresses based on the GUID of the port. These MAC addresses are reported to Windows to enable normal communication. These addresses are replaced by the IPoIB driver before messages are sent on the wire, and are only for local usage. Mellanox cards are usually shipped with GUIDs of the form:

00-02-C9-02-00-XX-YY-ZZ or 00-02-C9-03-00-XX-YY-ZZ.

Since a GUID contains 8 bytes, the appropriate truncation should be done as illustrated in the following example:

Mellanox Port GUID = "0002c90200XXYYZZ" => MAC = "0002c9XXYYZZ".

Mellanox Port GUID = "0002c90300XXYYZZ" => MAC = "0002caXXYYZZ".

This release supports generic MAC address generation according to a user-defined bitwise GUID mask. A GUID mask is an 8-bit field that indicates which bytes of a GUID should be used in MAC address generation.

Since a MAC address has a fixed 6-byte length, the mask must contain exactly 6 non-zero bits.

- Examples of valid masks: 0xfc (binary: 1111 1100); 0x3f (binary: 0011 1111)

- Examples of invalid masks: 0xfd - contains 7 non-zero binary digits; 0x2d contains only 4 non-zero binary digits
- Example of MAC generation given a mask of 0xe7: Port GUID = "0002c90200112233" => (mask == 0xe7) => MAC = "0002c9112233".

To specify the mask, the user should change the appropriate registry value (GUIDMask) located under

```
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Class\
{4D36E972-E325-11CE-BFC1-08002bE10318}\<IPoIB interface id>
```

This value is also accessible via the adapter's Properties user interface box.

IPoIB supports other companies' GUIDs such as Cisco, HP, SuperMicro, SilverStorm and Voltaire. If the port GUID is not another company's GUID, or if it is not in one of the forms above, IPoIB will not be able to generate the correct MAC addresses from the HCA port GUID. In this case, the GUID that is generated will be an integer starting with the number 02-00-00-00-00-00.

Please use "ipconfig /all" to obtain the MAC that was reported to Windows.

If the installation was successful yet the DHCP did not assign an IP to the IPoIB interface, most likely the IPoIB driver did not recognize the port's GUID. You can run the utility 'guid2mac_checker.exe' which is available via www.mellanox.com > Products > InfiniBand SW/Drivers > Mellanox WinOF.

The utility checks whether the port's GUID is recognized by the driver, and performs one of the following actions:

1. If the IPoIB driver recognizes the GUID, then it prints a confirmation message;
2. If the driver does not recognize the GUID but guid2mac_checker.exe recognizes it, then the utility writes an appropriate GUID mask to the registry;
3. If neither the driver nor guid2mac_checker.exe recognize the GUID, then the utility instructs the user how to create an appropriate GUID mask.

Note: An invalid GUID mask will be rejected and IPoIB will return to its default flow.

Note: It is not possible to change MAC address generation for known vendors like Cisco, HP, DELL etc.

To change network adapter card GUIDs, add the following flags when burning firmware:

- "-guid <GUID> -mac <mac>" for ConnectX HCA devices, and
- "-guid <GUID>" for the other (InfiniHost III family) HCA device.

See Section [Updating Adapter Card Firmware](#) for details.

Using the specified <GUID>, the following four parameters will be assigned:

- node GUID=<GUID>>,
- port1 GUID=<GUID>+1
- port2 GUID=<GUID>+2
- system image GUID=<GUID>+3

For example, to burn firmware on a ConnectX network adapter, enter:

```
flint -d mt25418_pci_cr0 -i <image_name.bin> -guid 0002c90200123456  
-mac 0002c9123457 -burn
```

4.2.4 IGMP Configuration

Multicast traffic on IPoIB works only with IGMP v2 and not with IGMP v3 which is the default.

To configure your machine to use IGMP v2, please follow the instructions below.

- For Windows 2003 and Windows XP, run the following commands from the command line:
 - netsh routing ip igmp install
 - netsh routing ip igmp install add interface "interface name of IPoIB adapter" igmpprototype=igmpv2

Note: If after executing the commands above IGMP V3 remains in use, please follow the instructions on

<http://support.microsoft.com/default.aspx/kb/815752>

- For Windows 2008, run the following commands from the command line:
 - servermanagercmd.exe -install NPAS-RRAS-Services
 - netsh routing ip igmp install
 - netsh routing ip igmp install add interface "interface name of IPoIB adapter" igmpprototype=igmpv2

5 OpenSM

OpenSM is an InfiniBand Subnet Manager. For Mellanox WinOF VPI to operate, OpenSM must be running on at least one host machine in the InfiniBand cluster.

OpenSM can either run as a Windows service which starts automatically during boot or can be started manually from the following directory: <installation_directory>\tools.

Please configure at least one machine to start the service automatically:

1. Right click on "My computer" and select Manage
2. Go to "Services and Applications" and select Services
3. Right click "OpenSM" and select Properties
4. Change "Startup type" to Automatic
5. Change service to start mode

OpenSM as a service will use the first port which is not in "down" state.

To run OpenSM manually, enter on the command line: opensm.exe

For additional run options, enter: opensm.exe -h

Notes:

- For long term running, please avoid using the '-v' (verbosity) option to avoid exceeding disk quota.
- Running OpenSM on multiple servers may lead to incorrect OpenSM behavior.

Please do not run OpenSM on more than 2 machines in the subnet.

- IBDiagnet cannot run on the same IB port that OpenSM is running on.

6 Ethernet Driver

6.1 Overview

The Mellanox VPI WinOF driver release introduces the following capabilities:

- One or two ports
- Up to 16 Rx queues per port
- Rx steering mode (RSS)
- Hardware Tx/Rx checksum calculation
- Large Send Offload (i.e., TCP Segmentation Offload)

- Hardware multicast filtering
- Adaptive interrupt moderation
- Polling on send completion queue to decrease the number of interrupts (default: disabled)
- Polling on receive completion queue to decrease the number of interrupts (default: disabled)
- MSI-X support (only on Windows Server 2008 and higher)
- VLAN Tx/Rx acceleration (HW VLAN stripping/insertion)
- High Availability (HA) between ports and Mellanox NICs
- Load Balancing between ports and Mellanox NICs
- Quality of Service (QoS)
- HW VLAN filtering
- Tx arbitration mode: VLAN user-priority (off by default)

6.2 Performance Remarks

6.2.1 Performance Tuning

To improve performance, activate the performance tuning tool as follows:

1. Go to Control Panel.
2. Open Network Connections.
3. Right click on one of the entries "Mellanox ConnectX 10Gbit Ethernet Adapter" and select Properties.
4. Select the Performance tab.
5. Click General Tuning.

Clicking the "General Tuning" button will change several registry entries (described below), and will check for system services that may decrease performance. It will also generate a log of the changes made. Users can refer to this log to restore the previous values.

The log path is:

%HOMEDRIVE%\windows\system32\logfiles\performancetunning.log.

This tuning is needed on one adapter only, and only once after the installation (as long as these entries are not changed directly in the registry, or by some other install or script).

Note: You may need to reboot for the changes to take effect. You will be asked to reboot if necessary.

The registry entries that may be added/changed by this procedure are:

1. Windows 2003:

- Under
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters:
 - TcpWindowSize, type REG_DWORD, value set to 512K.
 - Tcp1323Opts, type REG_DWORD, value set to 1.
 - SackOpts, type REG_DWORD, value set to 0.
 - EnableRss, type REG_DWORD, value set to 1.
 - RssBaseCpu, type REG_DWORD, value set to 1.
 - MaxNumRssCpus, type REG_DWORD, value set to 2.
- Under
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\AFD\Parameters:
 - FastSendDatagramThreshold, type REG_DWORD, value set to 64K. The following service is disabled:
 - "Windows Firewall/Internet Connection Sharing (ICS)"

2. Windows 2008 and Windows 2008-R2:

- Under
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters:
 - SackOpts, type REG_DWORD, value set to 0.
- Under
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\AFD\Parameters:
 - FastSendDatagramThreshold, type REG_DWORD, value set to 64K.
- Under
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Ndis\Parameters:
 - RssBaseCpu, type REG_DWORD, value set to 1.
 - MaxNumRssCpus, type REG_DWORD, value set to 2.

Enabling Receive Side Scaling (RSS) is performed by means of the following command:

```
"netsh int tcp set global rss = enabled"
```

Disable the time stamps on both sides. This is performed by means of the following command:

```
"netsh int tcp set global timestamps=disabled"
```

The following services are disabled:

- "Base Filtering Engine (BFE)"
- "Windows Firewall (MpsSvc)"

On some machines the following change may provide additional performance:

- Change the send completion method from interrupt to polling as follows:
 1. Open Device Manager
 2. Right click the used Ethernet adapter (Mellanox ConnectX 10G Ethernet Adapter) and select Properties.
 3. Select the Advanced tab.
 4. Select Performance Options and then click Properties.
 5. Select Send Completion Method.
 6. Change the value to polling.
 7. Click OK twice.

6.2.2 Known Performance Issues

- On Intel I/OAT supported systems, it is highly recommended to install and enable the latest I/OAT driver (download from www.intel.com).
- With I/OAT enabled, sending 256-byte messages or larger will activate I/OAT. This will cause a significant latency increase due to I/OAT algorithms. On the other hand, throughput will increase significantly when using I/OAT.
- On some systems, reducing the receive ring size ("Receive Ring Size" value under the Advanced tab) may improve performance.
- On some systems, changing the send completion method to polling ("Send Completion method" value under the Advanced tab) may improve performance.

6.3 Booting Windows from an iSCSI Target

Booting Windows from an iSCSI Target is supported on Windows Server 2008 and 2003 with the following limitations:

- 2008: Installing Windows Server 2008 directly to an iSCSI Target is not supported.
- 2003: Windows Server 2003 must be configured using a static IP address (and not through DHCP).

For more details on how to boot from a SAN using a Mellanox adapter card, please refer to <http://www.etherboot.org/wiki/sanboot>.

Also note that Mellanox has also tested the adapter card with a Windows iSCSI Target from StarWind (build 4.1).

6.4 Known Issues and Limitations

1. After uninstalling the MLNX_EN for Windows package, you need to reboot the machine.
2. This release does not support installing MLNX_WinOF_VPI and other drivers such as MLNX_EN for Windows, Mellanox WinOF, WinOF, and Mellanox WinIB.
3. If your machine has WDF version 1.5 installed, you will need to reboot the machine after the installation completes.
4. In "Control Panel\Add or Remove Programs", the Repair option of MLNX_EN does not repair "Mellanox Virtual Miniport Adapter" drivers.
Workaround: Uninstall the MLNX_EN for Windows package and reinstall it.
5. VLAN creation: VLANs on the same machine cannot be assigned dynamic IPs (using DHCP) of the same subnet.
6. Bundle creation (LBFO):
 - a. After creating a bundle, all adapters in the bundle still appear in the Network Connections display.
 - b. Once an adapter is part of a bundle, the following parameters cannot be changed: task offloading, RSS mode or MTU. To workaround this issue you need to (a) disassemble the bundle, (b) set the parameters for all the bundle adapters to the SAME values, and (c) reassemble the bundle.
 - c. When creating a new bundle, it may take some time (up to one minute) until the OS presents the new bundle.
 - d. When VLAN is present over the network adapter, LBFO will exclude this network adapter from the adapters list in LBFO GUI.
7. Replacing an adapter card may require reconfiguring LBFO bundles and/or VLAN adapters.

8. Uninstalling "Mellanox Virtual Miniport Driver" from a "Network properties" page may not work, and VLANs and LBFO bundles may still exist.

Workaround:

- a. Open the device manager and select the Ethernet adapter that you are using (Mellanox ConnectX 10G Ethernet Adapter).
- b. Double-click on the adapter and select the VLAN tab, then remove all existing VLANs.
- c. Select the LBFO tab, then remove all existing bundles.

6.5 Troubleshooting

- Issue: The installation of MLNX_WinOF_VPI for Windows fails with the following (or a similar) error message:

"This installation package is not supported by this processor type. Contact your product vendor."

Suggestion: This message is printed if you have downloaded and attempted to install an incorrect MSI -- for example, if you are trying to install a 64-bit MSI on a 32-bit machine (or vice versa).

- Issue: The performance is low.

Suggestion: This can be due to non-optimal system configuration. See the section "[Performance Tuning](#)" to take advantage of Mellanox 10 GBit NIC performance.

- Issue: The driver doesn't start.

Suggestion: This can happen due to an RSS configuration mismatch between the TCP stack and the Mellanox adapter. To confirm this scenario, open the event log and look under "System" for the "mlx4eth5" or "mlx4eth6" source. If found, enable RSS as follows:

- a. For Windows 2003, in the TCP registry set the KEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\EnableRss registry value to 1.
- b. For windows 2008, run the following command: "netsh int tcp set global rss = enabled".

Another option, which is less recommended and will cause low performance, is to disable RSS on the adapter. To do this set RSS mode to "No Dynamic Rebalancing".

- Issue: The Ethernet driver fails to start. In the Event log, under the mlx4_bus source, the following error message appears: RUN_FW command failed with error -22

Suggestion: The error message indicates that the wrong firmware image has been programmed on the adapter card.

See http://www.mellanox.com/content/pages.php?pg=firmware_download

- Issue: The Ethernet driver fails to start. A yellow sign appears near the "Mellanox ConnectX 10Gb Ethernet Adapter" in the Device Manager display.

Suggestion: This can happen due to a hardware error. Try to disable and re-enable "Mellanox ConnectX Adapter" from the Device Manager display.

- Issue: No connectivity to a Fault Tolerance bundle while using network capture tools (e.g., Wireshark).

Suggestion: This can happen if the network capture tool captures the network traffic of the non-active adapter in the bundle. This is not allowed since the tool sets the packet filter to "promiscuous", thus causing traffic to be transferred on multiple interfaces. Close the network capture tool on the physical adapter card, and set it on the LBFO interface instead.

- Issue: No Ethernet connectivity on 1Gb/100Mb adapters after activating Performance Tuning (part of the installation).

Suggestion: This can happen due to adding a TcpWindowSize registry value. To resolve this issue, remove the value key under HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\TcpWindowSize or set its value to 0xFFFF.

- Issue: System reboots on an I/OAT capable system on Windows Server 2008.

Suggestion: This may occur if you have an Intel I/OAT capable system with Direct Cache Access enabled, and 9K jumbo frames enabled. To resolve this issue, disable 9K jumbo frames.

- Issue: Packets are being lost.

Suggestion: This may occur if the port MTU has been set to a value higher than the maximum MTU supported by the switch.

- Issue: Issue(s) not listed above.

Suggestion: The MLNX_EN for Windows driver records events in the system log of the Windows event system. Using the event log you'll be able to identify, diagnose, and predict sources of system problems.

To see the log of events, open System Event Viewer as follows:

- Right click on My Computer, click Manage, and then click Event Viewer.

OR

- Click start-->Run and enter "eventvwr.exe".

- In Event Viewer, select the system log.

The following events are recorded:

- Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> has been successfully initialized and enabled.
- Failed to initialize Mellanox ConnectX EN 10Gbit Ethernet Adapter.
- Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> has been successfully initialized and enabled. The port's network address is <MAC Address>
- The Mellanox ConnectX EN 10Gbit Ethernet was reset.
- Failed to reset the Mellanox ConnectX EN 10Gbit Ethernet NIC. Try disabling then re-enabling the "Mellanox Ethernet Bus Driver" device via the Windows device manager.
- Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> has been successfully stopped.
- Failed to initialize the Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> because it uses old firmware version (<old firmware version>). You need to burn firmware version <new firmware version> or higher, and to restart your computer.
- Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> device detected that the link connected to port <Y> is up, and has initiated normal operation.
- Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> device detected that the link connected to port <Y> is down. This can occur if the physical link is disconnected or damaged, or if the other end-port is down.
- Mismatch in the configurations between the two ports may affect the performance. When Using MSI-X, both ports should use the same RSS mode. To fix the problem, configure the RSS mode of both ports to be the same in the driver GUI.
- Mellanox ConnectX EN 10Gbit Ethernet Adapter <X> device failed to create enough MSI-X vectors. The Network interface will not use MSI-X interrupts. This may affects the performance. To fix the problem, configure the number of MSI-X vectors in the registry to be at least <Y>.

7 SDP

SDP is currently under development. This is a preliminary version of this ULP, and it supports a limited set of API functions.

7.1 SDP Limitations

A limited set of API functions (w/w major flags) is supported by this version. These are: socket, connect, bind, listen, accept, send, WSASend, receive, WSARecv, select, AcceptEx, WSPShutdown and closesocket.

WSASend and WSARecv currently support all types of completion methods, including synchronous, completion routine, event and completion ports. Non-blocking IO is also supported.

Additionally:

getsockopt supports SO_PROTOCOL_INFOW and SO_CONNECT_TIME; and setsockopt supports SO_LINGER and SO_DONTLINGER WSPIoctl supports FIONBIO.

7.2 SDP Installation

SDP should be installed and activated at Mellanox WinOF VPI install time. If SDP is not installed, then please uninstall the Mellanox WinOF VPI package and reinstall it with SDP.

See MLNX_WinOF_VPI_ReleaseNotes.txt for details.

7.3 Running Applications over SDP

- Run 'sc start sdp' to verify that the SDP service is running. This is needed after each reboot.
- Set the environment variable 'SdpApplications' with the name of the program to use SDP. If there is more than one program, separate the names using semi-colons.

Examples:

```
SdpApplications=telnet.exe
```

```
SdpApplications=telnet.exe;ftp.exe
```

Note: If this variable is not set, then only programs named SdpConnect.exe can use SDP to connect.

- Run the application using the IPoIB interface IP address.

7.4 Running an Application over SDP and Ethernet

In order to allow your program to run both SDP sockets and Ethernet sockets, perform the following:

1. Set the registry value MIXED_SDP_APPLICATIONS to 1. It is located under HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\sdp\Parameters
2. Restart the SDP driver.
3. Make sure that the SdpApplications is *NOT* set to the name of your application.
4. Your program will now use only TCP connections and not SDP. In the places that you do want to use SDP and not TCP replace the call `s = socket(AF_INET_FAMILY, SOCK_STREAM, IPPROTO_TCP)`; with the call `s = WSASocket(AF_INET_FAMILY, SOCK_STREAM, IPPROTO_TCP, NULL, 0, WSA_FLAG_OVERLAPPED | 0x40)`;
Only that socket will use SDP.

7.5 Available Programs

The following applications were verified to work over SDP:

- Iometer: To obtain the program please refer to <http://www.iometer.org>
- iperf-2.0.1, iperf-1.7.0: These are test programs for 32-bit and 64-bit systems. To download them visit <http://sourceforge.net/projects/iperf>. Instructions for usage are included in the download package.
- TTcp.exe: Testing was conducted using the TTcp.exe version shipped with Windows XP SP2. Both synchronous and overlapped operations can be used.
Note: Other TTcp.ext versions may also work.
- Ntttcp.exe: This is a benchmark developed by Microsoft. Please contact Microsoft to obtain the program.
- NetPipe: Used to measure latency. To download visit <http://na-net.jp/na/>
- Microsoft CCS MPI
- SdpConnect.exe: This is a simple test program located under the SDP example directory. The program has two modes: client and server. In the server mode the program listens for connection; in the client mode the program connects to the server. The program can be used to test SDP with synchronous and overlapped operations.

Example 1:

- At node 1: SdpConnect.exe server 2222
- At node 2: SdpConnect.exe client 11.4.8.63 2222 0 1 0 0 1 3000 16000

Example 2:

- At node 1: SdpConnect.exe server 2222
- At node 2: SdpConnect.exe pingpong 11.4.8.63 2222 10000 10

For more options, enter: SdpConnect.exe

Note: SdpConnect source code is included in the SDK component of Mellanox WinOF.

7.6 Troubleshooting

- How can I verify that SDP is being used?

Currently, there is no simple way to indicate SDP is being used. However, if you know that your program consumes a lot of bandwidth, then there is an indirect way to find out. Open the Task Manager and switch to the networking tab. If you see that network utilization is low, this means that SDP is being used. Alternatively, if the program is running (i.e., the two sides communicate), stop the SDP on one side (via "net stop sdp") then try to reconnect it. If it succeeds then SDP was NOT used; if it fails then SDP was used.

- My program does not seem to use SDP.

Suggestions:

- a. Ping the remote node (ping <IP address of IPoIB interface>) to verify IPoIB is up.
- b. Verify that the SDP driver is loaded (net start sdp).
- c. Verify that the SdpApplications environment variable is correctly set (see Section [Bootting Windows from an iSCSI Target](#) above).
- d. Verify that the SDP provider is installed by running \Program Files\Mellanox\MLNX_WinOF\SDP\InstallSdpProvider.exe -l The output of this command should include 'SDP provider'. Otherwise, install the SDP provider using <...>\InstallSdpProvider.exe -i

- My system is experiencing instability and/or no network connectivity.

Suggestions: Remove the SDP provider using \Program Files\Mellanox\MLNX_WinOF\SDP\InstallSdpProvider.exe -r then restart your computer.

- Interoperability with Linux SDP is broken on OFED 1.2.5, 1.3.0, and 1.3.1. A complete fix for the problem is only expected with the next OFED release. Until then, please use the following workaround:
 - Click Start->Run and enter regedit.
 - Go to
 - KEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\sdp\Parameters
 - Change the value for MaximumRecvBufferSize and MaximumSendBufferSize to 0x810. This will allow both stacks to work but with lower BW due to the small message size.

8 WSD

8.1 Running Applications over WSD

1. Install the WSD provider on both computers. Enter:
`\Program Files\Mellanox\MLNX_WinOF\IPoIB\installsp.exe -i`
2. Check which providers are installed. Enter:
`\Program Files\Mellanox\MLNX_WinOF\IPoIB\installsp.exe -l`
3. Run the application. Please note that WSD has a fall back option; thus, if the connection fails over WSD, the connection will be attempted over IPoIB.
4. Remove the WSD provider:
`\Program Files\Mellanox\MLNX_WinOF\IPoIB\installsp.exe -r`

8.2 Performance

WSD has its performance counters.

1. Open perfmon and select add counters.
2. Locate the performance object called "IB winsock direct" and select "total sent bytes" or "total received bytes" -- this will display how much traffic is going through WSD (if any).

9 SRP

The Mellanox WinOF VPI stack does not install the SRP driver by default. If SRP is selected in the custom installation window, it will only be copied during installation.

To complete the SRP driver installation, an SRP target must be detected. This requires a Subnet Manager to be running somewhere in the InfiniBand subnet.

When an SRP target is detected, the "New Hardware Found" Wizard pops up.

Select Install Automatically and click Next. This installs the I/O unit device.

Once completed, the "New Hardware Found" Wizard pops up again. Select Install Automatically and click Next. This installs the SRP driver.

10 Starting and Verifying the IB Fabric

- If you rebooted your machine after the installation process completed, then IB interfaces should be up.
- 1. Check that the IB driver is running on all nodes by using 'vstat'. The vstat utility located at <installation_directory>\tools, displays the status and capabilities of the network adaptor card(s).
On the command line, enter "vstat" (use -h for options) to retrieve information about one adapter port for PCI Device ID 25204 or two adapter ports for all other PCI Device IDs. The field port_state will be equal to
 - PORT_DOWN - when there is no InfiniBand cable ("no link");
 - PORT_INITIALIZED - when the port is connected to some other port ("physical link");
 - PORT_ACTIVE - when the port is connected and OpenSM is running ("logical link").
- 2. Run OpenSM - see OpenSM operation instructions in the [OpenSM](#) section above.
- 3. Verify the status of ports by using vstat: All connected ports should report "PORT_ACTIVE" state.

11 Low level Performance Tests

The following performance tests are provided with the Mellanox WinOF VPI release under <installation_directory>\tools:

- Latency tests
 - ib_write_lat: RDMA write
 - ib_read_lat: RDMA read
 - ib_send_lat: UD, UC and RC (default) send

- Bandwidth tests
 - `ib_write_bw`: RDMA write
 - `ib_read_bw`: RDMA read
 - `ib_send_bw`: UD, UC and RC (default) send

For usage information, run: `<test name> -h`

Note: Since the default MTU value is different per network adaptor card type, use `"-m MTU"` to set the MTU value on both the server and the client to the same value. This should be done only on heterogeneous systems (different network adaptor cards on different servers).

12 Debug options

- IBAL supports WPP tracing tools by using the following GUIDs:
 - `"B199CE55-F8BF-4147-B119-DACD1E5987A6"` for user debug
 - `"99DC84E3-B106-431e-88A6-4DD20C9BBDE3"` for kernel debug
- MTHCA supports WPP tracing tools by using the following GUIDs:
 - `"2C718E52-0D36-4bda-9E58-0FC601818D8F"` for user debug
 - `"8BF1F640-63FE-4743-B9EF-FA38C695BFDE"` for kernel debug
- MLX4_HCA supports WPP tracing tools by using the following GUIDs:
 - `"1752F07C-7E5C-402c-9C5F-AD21E572F852"` for user debug
 - `"F8C96A49-AE22-41e9-8025-D7E416884D89"` for kernel debug
- MLX4_BUS supports WPP tracing tools by using the following GUIDs:
 - `"E51BB6E2-914A-4e21-93C0-192F4801BBFF"` for kernel debug
- IPoIB supports WPP tracing tools by using the following GUID:
 - `"3F9BC73D-EB03-453a-B27B-20F9A664211A"`
- WSD supports WPP tracing tools by using the following GUID:

- "156A98A5-8FDC-4d00-A673-0638123DF336"
- SDP supports WPP tracing tools by using the following GUIDs:
 - "D6FA8A24-9457-455d-9B49-3C1E5D195558" for user debug
 - "2D4C03CC-E071-48e2-BDBD-526A0D69D6C9" for kernel debug
- SRP supports WPP tracing tools by using the following GUID:
 - "5AF07B3C-D119-4233-9C81-C07EF481CBE6"

The flags and level of debug can be controlled at load time or runtime.

13 Driver Update and Uninstall Process

1. Clean Uninstall
2. Uninstall the package using the "Add or Remove Programs" utility
3. Driver Update
4. Driver update is currently not supported
5. Uninstall the driver package and then install a new version
6. Reboot the server to complete the uninstall process

14 Documentation

- Under <installation_directory>\documents:
 - Release Notes for: core (IBAL), IPoIB, WSD
 - README and user manuals for: opensm, SDP
- Under <installation_directory>:
 - License file
 - This document
- Under <installation_directory>\SDK:
 - core (IBAL) API HTML documentation (in SDK package)
 - hello_world code example (in SDK package): This is a 'two-sided' code example built by the DDK environment

Activation:

- Side A: hello_world.exe -d [daemon options]
- Side B: hello_world.exe --ip=<daemon_host_ip> [client options]

For options, enter: `hello_world.exe --help`