WHITE PAPER

Windows Server

Mellanox TECHNOLOGIES

Deploying Windows Server® 2012 with SMB Direct over Mellanox InfiniBand End-to-End Interconnect Solutions

1. Background

- Windows Server 2012 SMB Overview: http://technet.microsoft.com/en-us/library/hh831795.aspx
- High-Performance, Continuously Available File Share Storage for Server Applications Technical Preview: http://technet.microsoft.com/en-us/library/hh831399.aspx
- Deploying Fast and Efficient File Servers for Server Applications: http://technet.microsoft.com/en-us/library/hh831723.aspx
- Windows Server 2012 - Test cases for Hyper-V over SMB (includes PowerShell examples): http://blogs.technet.com/b/josebda/archive/2012/03/06/windows-server-quot-8-quot-beta-test-cases-for-hyper-v-over-smb.aspx
- Building Your Cloud Infrastructure: Converged Data Center with File Server Storage: http://technet.microsoft.com/en-us/library/hh831738.aspx

2. Hardware and Software

To implement and test this technology, you will need:

- Two or more computers running Windows Server 2012
- One or more of Mellanox's ConnectX-3 family of adapters for each server
- One or more Mellanox InfiniBand switches
- Two or more cables required for InfiniBand (typically using QSFP connectors)

Mellanox states support for Windows Server 2012 SMB Direct and Kernel-mode RDMA capabilities on the following adapter models:

- Mellanox ConnectX-3/ConnectX-3 Pro. These cards use Fourteen Data Rate (FDR) InfiniBand at 56 Gb/s data rate.

You can find more information about these adapters on Mellanox's website.

Important note: Support for SMB Direct is only provided starting with the ConnectX-3 family of adapters. SMB Direct is not supported by the ConnectX-2, ConnectX, or earlier adapter families.

There are many options in terms of adapters, cables and switches. At the Mellanox web site you can find more information about these InfiniBand adapters (http://www.mellanox.com/content/pages.php?pg=infiniband_cards_overview&menu_section=41) and InfiniBand switches (http://www.mellanox.com/content/pages.php?pg=switch_systems_overview&menu_section=49). Here are some examples of configurations you can use to try the Windows Server 2012:

2.3 - Two computers using FDR

You may also try the faster FDR speeds (56Gb/x data rate). The minimum setup in this case would again be two cards and a cable. Please note that the QDR and FDR cables are different, although they use similar connectors. Here's what you will need:

Quantity	Part Number	Description	Links
2	MCX353A-FCBT	ConnectX-3 Adapter, Single Port, QSFP, FDR InfiniBand	http://www.mellanox.com/related-docs/prod_adapter_cards/ConnectX3_VPI_Card.pdf
1	MC2207130-001	QSFP to QSFP cables (FDR), 1m (3ft)	http://www.mellanox.com/related-docs/prod_cables/DS_FDR_56Gbs_Passive_Copper_Cables.pdf

Please note that you will need a system with PCIe Gen3 slots to achieve the rated speed in this card. These slots are available on newer system like the ones equipped with an Intel Romley motherboard. If you use an older system, the card will be limited by the speed of the older PCIe Gen2 bus.

2.4 - Ten computers using dual FDR cards

If you're interested in experience great throughput in a private cloud setup, you could configure a two-node file server cluster plus an eight-node Hyper-V cluster. You could also use two InfiniBand cards for each system, for added performance and fault tolerance. In this setup, you would need 20 FDR cards and a 20-port FDR switch (Mellanox sells a model with 36 FDR ports). Here are the parts required:

Quantity	Part Number	Description	Links
20	MCX353A-FCBT	ConnectX-3 Adapter, Single port, QSFP, FDR InfiniBand	http://www.mellanox.com/related-docs/prod_adapter_cards/ConnectX3_VPI_Card.pdf
20	MC2207130-001	QSFP to QSFP cables (FDR), 1m (3ft)	http://www.mellanox.com/related-docs/prod_cables/DS_FDR_56Gbs_Passive_Copper_Cables.pdf
1	SX6036	InfiniBand Switch, 36 ports, QSFP, FDR	http://www.mellanox.com/content/pages.php?pg=products_dyn&product_family=132&menu_section=49

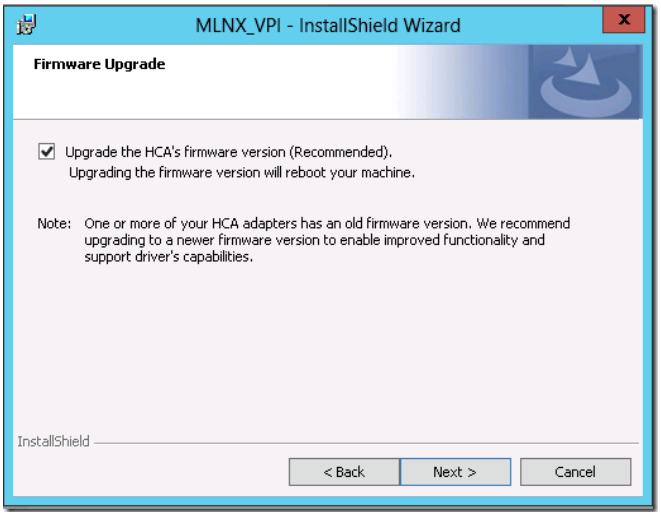
3. Download and Update the Drivers

Windows Server 2012 includes an inbox driver for the Mellanox ConnectX-3 cards. However, Mellanox provides updated firmware and drivers for download. You should be able to use the inbox driver to access the Internet to download the updated driver.

The latest Mellanox drivers for Windows Server 2012 can be downloaded from the Windows Server 2012 tab on this page on the Mellanox web site: http://www.mellanox.com/content/pages.php?pg=products_dyn&product_family=32&menu_section=34.

The package is provided to you as a single executable file. Simply run the EXE file to update the firmware and driver. This package will also install Mellanox tools on the server.

After the download, simply run the executable file and choose one of the installation options (complete or custom). The installer will automatically detect if you have at least one card with an old firmware, offering to update it. You should always update to the latest firmware provided.



Note 1: This package does not update firmware for OEM cards. If you are using this type of card, contact your OEM for an update.

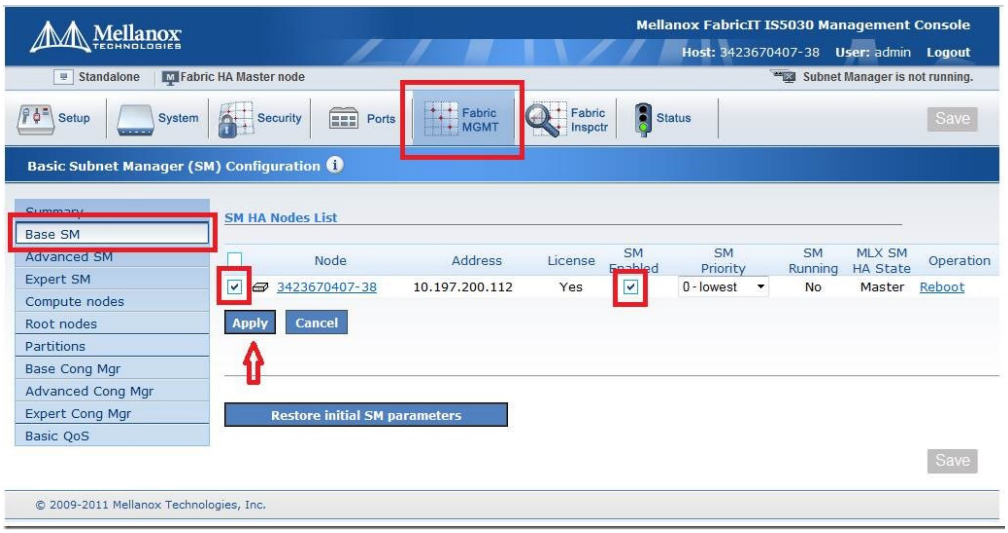
Note 2: Certain Intel Romley systems won't boot Windows Server 2012 when an old Mellanox firmware is present. It might be required for you to update the firmware of the Mellanox card using another system before you can use that Mellanox card on the Intel Romley system. That issue might also be addressed in certain cases by updating the firmware/BIOS of the Intel Romley system.

4. Configure a Subnet Manager

When using an InfiniBand network, you are required to have a subnet manager running. The best option is to use a managed InfiniBand switch (which runs a subnet manager), but you can also install a subnet manager on a computer connected to an unmanaged switch. Here are some details:

4.1 - Best option – Using a managed switches with a built-in subnet manager

For this option, make sure you use managed switches. These switches come ready to run their own subnet manager and all you have to do is enable that option using the switch's web interface. See the example below:



4.2 - Using OpenSM with a single unmanaged switch

If you don't have a managed switch, you can use one of the computers running Windows Server 2012 to run your subnet manager. When you installed the Mellanox tools on step 3, you also installed the OpenSM.EXE tool, which is a subnet manager that runs on Windows Server. You want to make sure you install it as an auto-starting service.

Although the installation program configures OpenSM to run as a service, it misses the parameter to limit the log size. Here are a few commands to remove the default service and add a new one that has all the right parameters and starts automatically. Run them from a PowerShell prompt running as Administrator:

SC.EXE delete OpenSM

New-Service –Name "OpenSM" –BinaryPathName "`"C:\Program Files\Mellanox\MLNX_VPI\IB\Tools\opensm.exe`" --service -L 128" -DisplayName "OpenSM" –Description "OpenSM" -StartupType Automatic

Start-Service OpenSM

Note 1: This assumes that you installed the tools to the default location: C:\Program Files\Mellanox\MLNX_VPI

Note 2: For fault tolerance, make sure you have two computers on your network configured to run OpenSM. It is not recommended to run OpenSM in more than two computers connected to a switch.

4.3 - Using OpenSM with two unmanaged switches

For complete fault tolerance, you want to have two switches and have two cards (or a dual-ported card) per computer, one going to each switch. With SMB Multichannel, you get fault tolerance in case a single card, cable or switch has a problem. However, each instance of OpenSM can only handle a single switch. In this case, you need two instances of OpenSM.EXE running on the computer, one for each card, working as a subnet manager for each of the two unmanaged switches.

In order to identify the two ports you have on the system (either on a single dual-ported card or in two single-ported cards). To do this, you need to run the IBSTAT tool from Mellanox, which will show you the identification for each InfiniBand port in your system (look for a line showing the port GUID). Here's a sample with the two port GUIDs highlighted:

PS C:\> ibstat

CA 'ibv_device0'

CA type:

Number of ports: 2

Firmware version: 0x20009209e

Hardware version: 0xb0

Node GUID: 0x0002c903000f9956

System image GUID: 0x0002c903000f9959

Port 1:

State: Active

Physical state: LinkUp

Rate: 40

Base lid: 1

LMC: 0

SM lid: 1

Capability mask: 0x90580000

Port GUID: 0x0002c903000f9957

Port 2:

State: Down

Physical state: Polling

Rate: 70

Base lid: 0

LMC: 0

SM lid: 0

Capability mask: 0x90580000

Port GUID: 0x0002c903000f9958

Once you have identified the two port GUIDs, you can run the following commands from a PowerShell prompt running as Administrator:

```
SC.EXE delete OpenSM

New-Service –Name "OpenSM1" –BinaryPathName "`"C:\Program Files\Mellanox\MLNX_VPI\
IB\Tools\opensm.exe`" --service -g 0x0002c903000f9957 -L 128" -DisplayName "OpenSM1" –
Description "OpenSM for the first IB subnet" -StartupType Automatic

New-Service –Name "OpenSM2" –BinaryPathName "`"C:\Program Files\Mellanox\MLNX_VPI\
IB\Tools\opensm.exe`"  --service -g 0x0002c903000f9958 -L 128" -DisplayName "OpenSM2" –
Description "OpenSM for the second IB subnet" -StartupType Automatic

Start-Service OpenSM1

Start-Service OpenSM2
```

Note 1: This assumes that you installed the tools to the default location: C:\Program Files\Mellanox\MLNX_VPI

Note 2: For fault tolerance, make sure you have two computers on your network, both configured to run two instances of OpenSM. It is not recommended to run OpenSM in more than two computers connected to a switch.

5. Configure IP Addresses

After you have the drivers in place, you should configure the IP address for your NIC. If you're using DHCP, that should automatically, so just skip to the next step.

For those doing manual configuration, assign an IP address to your interface using either the GUI or something similar to the PowerShell below. This assumes that the interface is called RDMA1, that you're assigning the IP address 192.168.1.10 to the interface and that your DNS server is at 192.168.1.2.

```
Set-NetIPInterface -InterfaceAlias RDMA1 -DHCP Disabled

Remove-NetIPAddress -InterfaceAlias RDMA1 -AddressFamily IPv4 -Confirm:$false

New-NetIPAddress -InterfaceAlias RDMA1 -AddressFamily IPv4 -IPv4Address 192.168.1.10
-PrefixLength 24 -Type Unicast

Set-DnsClientServerAddress -InterfaceAlias RDMA1 -ServerAddresses 192.168.1.2
```

6. Verify Everything is Working

Follow the steps below to confirm everything is working as expected:

6.1 - Verify network adapter configuration

Use the following PowerShell cmdlets to verify Network Direct is globally enabled and that you have NICs with the RDMA capability. Run on both the SMB server and the SMB client.

```
Get-NetOffloadGlobalSetting | Select NetworkDirect

Get-NetAdapterRDMA

Get-NetAdapterHardwareInfo
```

6.2 - Verify SMB configuration

Use the following PowerShell cmdlets to make sure SMB Multichannel is enabled, confirm the NICs are being properly recognized by SMB and that their RDMA capability is being properly identified.

On the SMB client, run the following PowerShell cmdlets:

```
Get-SmbClientConfiguration | Select EnableMultichannel

Get-SmbClientNetworkInterface
```

On the SMB server, run the following PowerShell cmdlets:

```
Get-SmbServerConfiguration | Select EnableMultichannel
Get-SmbServerNetworkInterface
netstat.exe -xan | ? {$_ -match "445"}
```

Note: The NETSTAT command confirms if the File Server is listening on the RDMA interfaces.

6.3 - Verify the SMB connection

On the SMB client, start a long-running file copy to create a lasting session with the SMB Server. While the copy is ongoing, open a PowerShell window and run the following cmdlets to verify the connection is using the right SMB dialect and that SMB Direct is working:

```
Get-SmbConnection
Get-SmbMultichannelConnection
```

```
netstat.exe -xan | ? {$_ -match "445"}
```

Note: If you have no activity while you run the commands above, it's possible you get an empty list. This is likely because your session has expired and there are no current connections.

7. Review Performance Counters

There are several performance counters that you can use to verify that the RDMA interfaces are being used and that the SMB Direct connections are being established. You can also use the regular SMB Server and SMB Client performance counters to verify the performance of SMB, including IOPs (data requests per second), Latency (average seconds per request) and Throughput (data bytes per second). Here's a short list of the relevant performance counters.

On the SMB Client, watch for the following performance counters:

- RDMA Activity - One instance per RDMA interface
- SMB Direct Connection - One instance per SMB Direct connection
- SMB Client Shares - One instance per SMB share the client is currently using

8. Review the Connection Log Details (optional)

On the SMB Server, watch for the following performance counters:

- RDMA Activity - One instance per RDMA interface
- SMB Direct Connection - One instance per SMB Direct connection
- SMB Server Shares - One instance per SMB share the server is currently sharing
- SMB Server Session - One instance per client SMB session established with the server

SMB 3.0 now offers a "Object State Diagnostic" event log that can be used to troubleshoot Multichannel (and therefore RDMA) connections. Keep in mind that this is a debug log, so it's very verbose and requires a special procedure for gathering the events. You can follow the steps below:

First, enable the log in Event Viewer:

- Open Event Viewer
- On the menu, select "View" then "Show Analytic and Debug Logs"
- Expand the tree on the left: Applications and Services Log, Microsoft, Windows, SMB Client, ObjectStateDiagnostic
- On the "Actions" pane on the right, select "Enable Log"
- Click OK to confirm the action

After the log is enabled, perform the operation that requires an RDMA connection. For instance, copy a file or run a specific operation.

If you're using mapped drives, be sure to map them after you enable the log, or else the connection events won't be properly captured.

Next, disable the log in Event Viewer:

- In Event Viewer, make sure you select Applications and Services Log, Microsoft, Windows, SMB Client, ObjectStateDiagnostic
- On the "Actions" page on the right, "Disable Log"

Finally, review the events on the log in Event Viewer. You can filter the log to include only the SMB events that confirm that you have an SMB Direct connection or only error events.

The "Smb_MultiChannel" keyword will filter for connection, disconnection and error events related to SMB. You can also filter by event numbers 30700 to 30706.

- Click on the "ObjectStateDiagnostic" item on the tree on the left.
- On the "Actions" pane on the right, select "Filter Current Log…"
- Select the appropriate filters

You can also use a PowerShell window and run the following cmdlets to view the events. If there are any RDMA-related connection errors, you can use the following:

Get-WinEvent -LogName Microsoft-Windows-SMBClient/ObjectStateDiagnostic -Oldest |? Message -match "RDMA"

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com