

INSTALLING (MELLANOX)INFINIBAND ON HPC SERVER 2008

One would assume that Infiniband on Windows is just going to be as easy as any other plug and play device installation. Well, in some cases it is. When you have some old cards and an old switch, no documentation, both not supported any longer and in an unknown state, it may not be! I am sure this experience is rather common, so I've decided to document what I did to make them work. In this case, I will focus on the Mellanox Infiniband stack, as it is the only one I could find that has public beta support for HPC Server 2008. Besides, Mellanox is the dominant provider of Infiniband hardware, even if it is then re-branded and re-sold by others.

DISCLAIMER: This is not official guidance, just my notes. No guarantee is offered or implied. Your experience may differ, mileage may vary etc... etc...

1. Identify the Infiniband card

First - it seems stupid - look at the card! Somewhere on the board there will be a model number (e.g. mt25204) and a revision number (e.g. a2). You may also find a GUID hexadecimal number. Note it down for later use. With the model and revision number, you can go to the manufacturer's (Mellanox) web site and download the latest [firmware](#). If you have a re-branded Mellanox card, check on the re-seller's site for their model number and the relative firmware. If you are out of luck (as I was) and the card is no longer supported, you can always try and identify the equivalent Mellanox model and download its firmware. Note: I do not recommend this. Go via official support channels if you can.

2. Download the software suite

Second, download the appropriate revision of the Infiniband software suite from your manufacturer's web site. Mellanox has announced officially that they will support Windows Server 2008, including the new network direct interface. At the time of writing this (February 25th 2008), you can download a beta version of their [WinIB 1.4](#) stack, which includes the tools you need to burn the firmware, card drivers, fabric drivers, protocol drivers, open subnet manager, etc...

3. Install the Infiniband tools

The fun begins! Let us assume that you have no Server 2008 cluster yet:

- Install Windows Server 2008 (HPC edition or any other x64 edition except web) on your intended cluster head node. Do NOT install the HPC pack yet.
- Install WinIB 1.4 from the downloaded msi package. Simply follow the installation wizard's on-screen instructions.

- Open the 2008 Server Manager application. Click on Diagnostics -> device manager. If everything went smoothly, an Infiniband adapter should be listed there - e.g. Mellanox Infinihost Lx III - and it should be active. You should also be able to see an extra Open IB Ethernet adapter, which is a way to implement the IP over Infiniband protocol. If this is the case, you've been lucky!

- You may want to verify the installed protocols by typing at the command prompt:

```
>cd c:\program files\mellanox\winib\ipoib
```

```
>dir
```

This will list the contents of the directory, which include the ipoib drivers, Winsock and NDI providers:

```
C:\Program Files\Mellanox\WinIB\IPoIB>dir
```

```
Volume in drive C has no label.  
Volume Serial Number is 9E44-FE08
```

```
Directory of C:\Program Files\Mellanox\WinIB\IPoIB
```

```
02/20/2008  01:41 PM    <DIR>          .  
02/20/2008  01:41 PM    <DIR>          ..  
12/16/2007  04:02 PM              94,720  ibwsd.dll  
12/16/2007  04:02 PM              62,464  ibwsd32.dll  
12/16/2007  04:02 PM              16,896  installsp.exe  
12/16/2007  04:02 PM               7,897  ipoib.cat  
12/16/2007  04:02 PM             136,704  ipoib.sys  
02/20/2008  01:41 PM    <DIR>          NDI  
12/16/2007  04:02 PM              6,385  netipoib.inf  
          6 File(s)              325,066 bytes  
          3 Dir(s)  101,674,553,344 bytes free
```

If the output looks like this, all the bits should be in the right place. To check whether they are actually installed, type:

```
>installsp -l
```

- You should be able to see a number of protocol providers, including Winsock direct and OpenIB Network Direct:

```
C:\Program Files\Mellanox\WinIB\IPoIB>installsp -l  
0000001001 - MSAFD Tcpip [TCP/IP]  
0000001002 - MSAFD Tcpip [UDP/IP]  
0000001003 - MSAFD Tcpip [RAW/IP]  
0000001004 - MSAFD Tcpip [TCP/IPv6]  
0000001005 - MSAFD Tcpip [UDP/IPv6]
```

```
0000001006 - MSAFD Tcpip [RAW/IPv6]
0000001007 - RSVP TCPv6 Service Provider
0000001008 - RSVP TCP Service Provider
0000001009 - RSVP UDPv6 Service Provider
0000001010 - RSVP UDP Service Provider
0000001011 - OpenIB Network Direct Provider
0000001012 - Mellanox Winsock Direct for InfiniBand
```

If you do, everything is probably working fine (this is still beta software, after all). Just make sure that you have a (compatible) subnet manager (e.g. OpenSM, provided with the Mellanox stack) running on the infiniband network and you can skip to step 4.

If you don't see the output above, the system reports "cables disconnected" even when they are connected, or IPoIB does not seem to work, a number of things could be wrong. These are the problems I encountered:

1. The providers are not (all) installed. This is easily fixed by typing

```
> installsp -i
```

2. The card firmware is in an inconsistent state. To check that, type:

```
> vstat
```

If you see that the card is in "flash recovery mode", you probably have an old, corrupt or incompatible firmware on it. *Vstat* may also return information that is useful to identify the correct firmware to restore, if the existing one is not completely unusable. Here's an example for a working device:

```
>vstat
```

```
hca_idx=0
uplink={BUS=UNRECOGNIZED (33)}
vendor_id=0x02c9
vendor_part_id=0x6274
hw_ver=0xa0
fw_ver=1.02.0000
PSID=_03B0120001
node_guid=7f7f:7f7f:7f7f:7f7f
num_phys_ports=1
    port=1
    port_state=PORT_ACTIVE (4)
    link_speed=2.5Gbps (1)
    link_width=4x (2)
    rate=10
    sm_lid=0x0001
    port_lid=0x0001
```

```
port_lmc=0x0
max_mtu=2048 (4)
```

Note the firmware version *fw_ver*, the hardware release *hw_ver* and the parameter set *psid*. Type *mst status* to get the name of the devices identified by the WinIB stack:

```
>mst status
Found 2 devices:
  mt25204_pciconf0
  mt25204_pci_cr0
```

The two entries above identify 2 ways of accessing the same device; *pci_cr0* allows you to write directly to it. Check that the device name (*mt25204* in my case) matches the one you found on the card and that you downloaded the correct firmware from the Mellanox website for the combination of device, hardware revision and *psid*.

Alas, if your firmware is seriously compromised (or by another vendor re-branding Mellanox devices), the information returned by *vstat* may be incomplete or unusable. You may have to trust your instincts and go with what you've seen on the card.

A variety of tools are available to burn a new firmware. If you have those that probably came with the card, use them. I was given the bare card, so I used *flint*, distributed by Mellanox in its WinIB package:

```
flint -d <device reported by mst status> -i <firmware image.bin> burn
```

```
>flint -d mt25204_pci_cr0 -i fw-25204-1_2_000-MHES18-
XTC_A2-A3.bin burn
  Current FW version on flash:  N/A
  New FW version:                N/A
```

Burn image with the following GUIDs:

```
Node:      0005ad000008bb84
Port1:     0005ad000008bb85
Sys.Image: 0005ad000008bb87
```

```
Read and verify Invariant Sector          - OK
Read and verify PPS/SPS in flash          - OK
Burning second FW image without signatures - OK
Restoring second signature                 - OK
```

You may have to specify additional options, like *-nofs* (no fail-safe) *-guid <16-digit hexadecimal number>* when you want to burn without checks and assign a new guid to the device. This may be necessary when the firmware is blank or seriously compromised. Note that the guid must be unique for the device (and the subnet manager) to work properly. One is normally already burned on the device and may be written on the board. If you found it in step (1), keep using it. If not, *vstat* will return the guid of the device (see

node_guid above) if it finds it on the firmware. You must be sure of what you are doing, as you risk leaving the device in an unusable state.

Once the new firmware has been burnt, reboot the machine.

Open a command prompt and type *vstat* again to check whether the firmware has been burnt correctly and the device port is now active.

3. There is no subnet manager running on the infiniband network.

Type *vstat* again and check for the *port_state*. If it is listed as “initializing”, “init” or similar, you may have no subnet manager running, or problems with the subnet manager on your infiniband switch. In the second case, refer back to your switch instructions. In the first case, you can always run *opensm*, which is provided with WinIB. To do a quick test, open a command prompt and type *opensm*, then open another command prompt re-type *vstat*. If *opensm* reports being the subnet master and the port status has changed to active, you had no subnet manager on the network.

In order to run *opensm* as a Windows service in the background, open the Windows 2008 server manager. Go to configuration -> services. Find *opensm* in the list. Change its startup type to “automatic” and start the service.

Note that some manufacturers do not support OpenSM (Cisco does not, for instance) and supply their own.

4. Install the HPC Pack on the head node

You can now install the HPC pack on the head node as per instructions provided with the product. In the network configuration wizard, you should be able to see a “IPoIB” adapter with Network Direct enabled. Choose that for MPI traffic. Note that in HPC Server 2008 you have an option to provide a DHCP service on the private Ethernet and on the infiniband network from the head node. I found this very useful and simple to implement, so I recommend it.

5. Automated deployment to compute nodes

There are several methods to deploy an operating system image to compute nodes and then have them join the cluster. Please refer to HPC Server 2008 documentation. In this document, I would like to focus on the Infiniband stack deployment only. Suffice to say that you can create a cluster just with public and private Ethernet connections, and then add Infiniband support at a second stage.

Once you have established the proper procedure to install the WinIB stack on 1 node, you have several options to deploy it automatically to the compute nodes. The simplest way, in my opinion, is to copy the downloaded WinIB msi package on the target nodes and then run the following command line on them:

```
Msiexec /qn /i <winib package name>.msi ADDLOCAL=ALL
```

If you have a cluster already set up, you can run the copy and the command line above on all the target nodes by preceding them with `clusrun /<node list>`.

Note that the msi package deploys unsigned drivers for now (they're a beta version, after all) and by default Server 2008 will ask you to confirm that you want to do so. If you want to avoid being prompted, you'll have to change the default driver signing policy.

This is not the only way to perform the installation, of course. You can add the command line above to your post-deployment tasks and have it run just after the operating system has been provisioned on the node.

You can also deploy the msi package as part of a node template within the HPC Server 2008 management console or with other Windows tools (group policy, system center etc...).

All these methods will not flash a firmware on the nodes, though. After the msi package installation, you can use `flint` to perform that task. Copy the firmware file on the target nodes, then run:

```
clusrun /<list of nodes> flint -d <device name> -i <firmware binary> -y -s burn
```

Flint options and comments apply: be sure of what you are doing, as you can make all the IB cards in your cluster unusable very quickly!

After installation, reboot the nodes. The following line will reboot the nodes immediately.

```
clusrun /<node list> shutdown /r /t 0
```

6. Conclusion

You should now have a nice Windows HPC Server 2008 cluster with Infiniband. If you like, run the [mpingpong test](#) to verify connectivity, latency and bandwidth. A version of it will be provided with Windows HPC Server 2008 from CTP onwards. A version of it exists for server 2003 as well, in the compute cluster toolpack. Check out <http://www.windowshpc.net> for more details.

Remember that this is not official guidance, just my notes. The software is also still in beta at the time of writing, so things may change and / or differ for you.

7. References

Mellanox Web Site: <http://www.mellanox.com>

Microsoft HPC Web Page: <http://www.microsoft.com/hpc>

HPC on Windows Community: <http://windowshpc.net>

Open Fabrics: <http://www.openfabrics.org>

