InfiniBand Trade Association

InfiniBand Interoperability

Method of Implementation

Contents

Introduction

This document describes the InfiniBand systems interoperability (IB Interop) testing procedure. The procedure uses the Intel® MPI Benchmarks (IMB) to test the point-to-point and fabric-wide operations for a variety of message sizes. It utilizes Open MPI for the message passing protocol. The IMB tests include PingPong, Gather, Sendrecv, Scatter, Allreduce, Alltoall, Allgatherv. For a more complete list, please check the Intergrators' List.

System Configuration

This section provides a detailed explanation on how to setup servers, xCAs and switches.

Cluster Setup

Please follow the steps outlined in this section to configure your servers. You must follow these steps carefully to avoid frustration and save yourself time.

xCA Setup

If you have completed the Cluster Setup section above, then you can skip this as the steps outlined here are a subset of the Cluster Setup procedure.

Switch Setup

Appendix B describes the steps required to configure a managed switch, if one needs to be configured. You can set the link speed, MTU size, flow control and other fields.

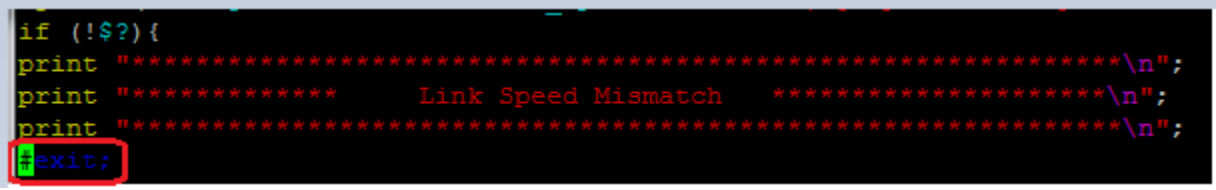
Test Scenarios

Interoperability test scenarios vary at each event depending on the devices registered for testing. Please refer to the Integrators' List to get some idea of the different test scenarios. The list provides scenarios from previous Plugfest events.

Test Procedure

1. Download the latest version of the IBTA Open MPI perl script.
2. Connect your cable following one of the scenarios chosen for testing in the plugfest.
3. Ensure that you have a link to all nodes. If there is no link, try re-inserting the cable. You may need to reboot the systems.
4. Make sure that opensm is running on one of the servers and that you can ping both through eth0/1 and ib0. (***ping sm-node-1*** and ***ping sm-node-1-ib***; or ***ping sm-node-2*** and ***ping sm-node-2-ib*** for example if you are testing between sm-node-1 and sm-node-2).
5. Run the ibta perl script with the command *perl ibta-mpi_script<date>.pl openmpi b 14 "someFolderName"*

   For EDR-2-EDR tests, you need to change the option "b 14" to "b 25". Please note that if the network setup has some links at EDR and others at FDR, the script will return a "**Link Speed Mismatch**" error and the test will exit before it completes. You can comment out the exit at line 217 in the script to allow the test to run to completion so that you are able to verify if there are other issues that you need to address. Please see the code snippet below. The part you need to comment out is highlighted.



6. The test run should take no more than 10 minutes to complete. Most runs will take a lot less than that. If the test run takes a long time, you may stop the script with the command ctrl+C.

   Once the text completes, record all the errors. The first few lines near the beginning of the run and at the end of the output to the screen will show if there are errors. You must check both to ensure that there were no errors at the beginning of the test run and none at the end. If you miss the output to the screen, you can view the errors after the test completes in the log files. The file you would need to check will usually if you missed the test output on the monitor has a name with the string "openmpi-mpi-part-b-test-output<date/time>" in it.

```
Errors for "SM-Node-6 HCA-1"
    GUID 0x11750000791748 port 1: [PortRcvRemotePhysicalErrors == 603]
Errors for 0x66a00e300289e "QLogic 12200 GUID=0x00066a00e300289e"
    GUID 0x66a00e300289e port ALL: [SymbolErrorCounter == 3287] [LinkErrorRecoveryCounter == 14] [PortRcvErrors == 956]
     [PortRcvSwitchRelayErrors == 1] [PortRcvConstraintErrors == 1] [LocalLinkIntegrityErrors == 3]
    GUID 0x66a00e300289e port 20: [SymbolErrorCounter == 3287] [LinkErrorRecoveryCounter == 14] [PortRcvErrors == 956]
     [PortRcvSwitchRelayErrors == 1] [PortRcvConstraintErrors == 1] [LocalLinkIntegrityErrors == 3]
Errors for 0x2c903005cc7b0 "SwitchX -  Mellanox Technologies"
    GUID 0x2c903005cc7b0 port ALL: [LinkErrorRecoveryCounter == 14] [PortXmitDiscards == 26]
    GUID 0x2c903005cc7b0 port 4: [LinkErrorRecoveryCounter == 14] [PortXmitDiscards == 26]
```

When you look at the errors, please make sure that you take note of the node at which the error occurs. If the error is at a connection other than that of the cable under test, you may need to use a different

control cable, and run the test again. You cannot fail the cable under test if the test output shows that the failure is at a control cable.

7. Conditions for passing the Interoperability tests
    a. ***Link width***: Link width check must return the expected value: 4x.
    b. ***Link Speed***: Link speed check must return the expected value: FDR, EDR or HDR.
    c. ***Link Recovery***: There must be no link recovery errors during the MPI run.
    d. ***Port Receive Errors***: There must be no port receive errors during the MPI run.
    e. ***Symbol Errors***: There must be no symbol errors during the MPI run.
    f. ***Port xmit Discard***: There must be no port xmit discard errors during the MPI run.
    g. ***MPI Test***: The MPI test must run to completion without error and without extreme performance degradation.

If the network operates normally and then suddenly you note that either you are having link issues or you see errors when you expect none, a reboot of all the systems might resolve the issue.

Known Issues

1. If there is no link between an HCA and a switch or another HCA, and rebooting or reinserting the cable fails, then you may need to reseat the HCA(s). It is also always a good idea to try the link with known good cables first.
2. The ssh-key exchange process tends to be error prone. Make sure that IPoIB and the host files are set correctly before you attempt to do the ssh-key exchange.
3. You may see the error shown in the picture below if there were changes made to the system.

```
[root@sm-node-1 ~]# ssh root@sm-node-1
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle attack)!
It is also possible that a host key has just been changed.
The fingerprint for the RSA key sent by the remote host is
6e:45:f9:a8:af:38:3d:a1:a5:c7:76:1d:02:f8:77:00.
Please contact your system administrator.
Add correct host key in /home/hostname /.ssh/known_hosts to get rid of this message.
Offending RSA key in /var/lib/sss/pubconf/known_hosts:4
RSA host key for pong has changed and you have requested strict checking.
Host key verification failed.
```

   To resolve this problem open the offending file in a text editor and delete the entry listed after the file name. For this exact error message you would:
    a. vim /var/lib/sss/pubconf/known_hosts
    b. delete entry # 4
4. If you run into an error that a directory such as (more likely) */var/tmp/ibdiagnet2* cannot be found, then please create it. This is the default location where ibdiagnet saves its output.
5. For Open MPI version 3.0.0 and newer, the '*sm*' option is replaced by '*vader*'
    a. So, instead of *$options = "--allow-run-as-root --mca btl openib,self,sm --mca pml ob1";* we now have *$options = "--allow-run-as-root --mca btl openib,self,vader --mca pml ob1";*
6. *mpirun* hangs on FDR systems

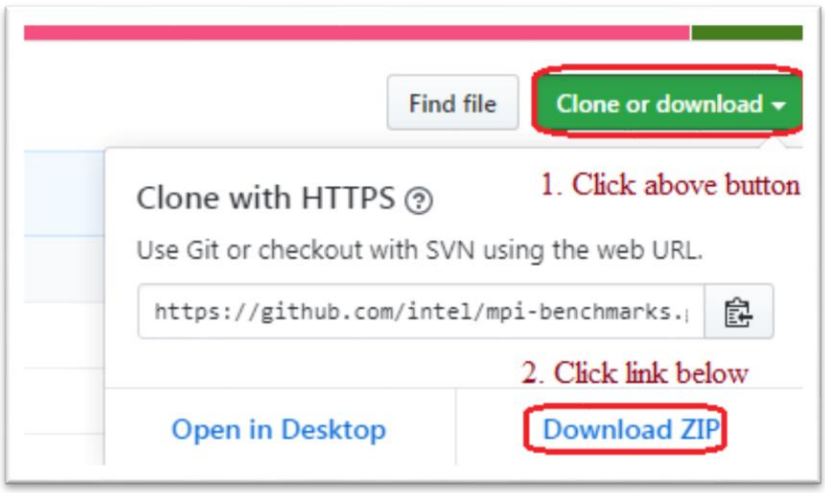
a. 2018-12-11 Yossi Itigin @ Mellanox suggested that we drop openib and add the MTU size option. So, the options line now looks like this: *$options = "--allow-run-as-root --mca btl self,vader --mca pml ucx -x UCX_RC_PATH_MTU=4096";* Note that this path MTU is for hardware fragmentation as opposed to the software fragmentation when you set the MTU size via ifcfg-ib0.

Appendix A

Cluster Setup

System Requirements

1. CentoOS 7.7 or newer. Please only have one IB card in each server. If you have more than one card, the IP address assignments may be handled incorrectly which can be difficult to resolve. Make sure you configure the host name. Change the *localhost.localdomain* under the network set up to something like *sm-node-** where '*' is the node number.
2. OFED 4.17-1 or newer. It is recommended that you have the same version of OFED installed in the same file system location on all systems. Please remember to review the OFED Release Notes. After the OFED installation, please run the command "`dracut -f -v`" to update device drivers.
3. OpenMPI 3.1.4 or newer. You must have the same version of Open MPI in the same file system location on all systems.
4. Intel® IMB_2019 or newer. When you get to the GitHub page, click on the "Clone or download" tab and select "Download ZIP"



Open MPI Installation

1. cd to the location where you unpacked the Open MPI download
2. Invoke the command ***./configure –prefix=/usr/local*** ← You can build without the prefix option */usr/local* because the default directory is */usr/local*. This command takes a while to complete.

3. Invoke the command ***make all install***.

You can verify the version of Open MPI with the command '***mpirun –version***'.

Intel® IMB Installation

1. cd to the '*src/src_c*' location directory of the location where you unpacked the Intel MPI Benchmark tarball.
2. Open the *make_ict* file and change line 3 from ***CC = mpiicc*** to ***CC = mpicc***. "mpiicc" uses the icc compiler while "mpicc" allows you to select the compiler and defaults to gcc.
    a. This step is not required for Intel ® MPI Benchmarks 2019
3. While still in the '*src/src_c*' directory, invoke the command '***make all'.*** If you get the error "mpiCC: error while loading shared libraries: libopen-pal.so.13: cannot open shared object file: No such file or directory"
    a. Run the command "*ldconfig"*
    b. Run *make all* again
4. Copy the IMB-MPI1 file which has just been built to the directory ***/usr/local/bin***.

Updating the IBTA MPI Script

Once all of the above installations are complete, update the mpi_location and imb_location in the script as follows

1. $mpi_location="/usr/local/bin".
2. $imb_location="/usr/local/bin/IMB-MPI1".
3. As of openmpi 1.8.2, you must include the mpirun option "—allow-run-as-root". Without this option, mpirun will abort when you run the script as root.

Lockable Memory Limits

The lockable memory limits in each system must be set to allow unlimited locked memory per process. This means that you must edit the ***/etc/security/limits.conf*** file so the following two lines read thus:

* hard memlock unlimited ← the (*) is not a bullet! It is part of the entry in the file
* soft memlock unlimited ← the (*) is not a bullet! It is part of the entry in the file

SELinux

SELinux is on by default. You must disable it. Edit the file ***/etc/selinux/config*** and set *SELINUX=disabled*

Disable the Firewall Daemon

You must disable the ***firewalld*** service. Issue the command sequence

1. systemctl stop firewalld
2. systemctl mask firewalld
3. systemctl status firewalld

The last command just checks to confirm that the firewalld daemon is really disabled. You should see something similar to the following:

```
● firewalld.service
   Loaded: masked (/dev/null)
   Active: inactive (dead) since Mon 2016-05-09 10:12:03 EDT; 1h 1min ago
 Main PID: 1075 (code=exited, status=0/SUCCESS)

May 09 10:01:56 sm-node-7 systemd[1]: Starting firewalld - dynamic firewall daemon...
May 09 10:01:56 sm-node-7 systemd[1]: Started firewalld - dynamic firewall daemon.
May 09 10:12:02 sm-node-7 systemd[1]: Stopping firewalld - dynamic firewall daemon...
May 09 10:12:03 sm-node-7 systemd[1]: Stopped firewalld - dynamic firewall daemon.
May 09 10:47:37 sm-node-7 systemd[1]: Cannot add dependency job for unit firewalld.service, ignoring: Unit firewalld.service is masked.
May 09 10:49:30 sm-node-7 systemd[1]: Cannot add dependency job for unit firewalld.service, ignoring: Unit firewalld.service is masked.
May 09 10:54:57 sm-node-7 systemd[1]: Cannot add dependency job for unit firewalld.service, ignoring: Unit firewalld.service is masked.
May 09 10:55:01 sm-node-7 systemd[1]: Cannot add dependency job for unit firewalld.service, ignoring: Unit firewalld.service is masked.
May 09 11:13:22 sm-node-7 systemd[1]: Cannot add dependency job for unit firewalld.service, ignoring: Unit firewalld.service is masked.
```

If you do not disable the firewalld service, the Open MPI script will return a “Broken Pipe” error. Since you are stopping the firewalld daemon, it may be best to set up your test system on an isolated subnet, one that is not connected to the internet or any other network.

You can restart the firewalld service by running the commands “*systemctl unmask firewalld*” and then “*systemctl start firewalld*”

Create Host Files

1. Create a data directory. We generally just create ***/data***
2. Create the file *master_node_list* and put this file in the ***/data*** directory. This file contains all the possible hosts you may run against. The file includes only host names, and not domains. So, for instances if you have four servers, you might have the names *sm-node-01 sm-node-02 sm-node-03 sm-node-04* (one per line).
3. Create the file *current-cluster*. This file contains a subset of the host names from the file */data/master_node_list* that you wish to test against. This file just includes the host names and not domain names: *sm-node-01 sm-node-02 sm-node-03 sm-node-04* (one per line). Please place this file in the user’s home directory.
4. Create the file *mpi-hosts* and put this in root home directory. This file contains the number of instances you want the MPI script to run. So, it contains a series of entries like this (not numbered, and one per line)
    a. sm-node-01-ib
    b. sm-node-02-ib
    c. sm-node-03-ib
    d. sm-node-04-ib

xCA Setup

Configure IPoIB

Add ***/etc/sysconfig/network-scripts/ifcfg-ib0*** and configure it. You may need to avoid quotes as some operating systems (e.g. RH 6.4) are not very tolerant of them. You could use something like 192.168.30.x, where ‘x’ is the last entry of the corresponding Ethernet IP address (read more on private subnet in RFC1918).

Modify /etc/hosts

You need to add lines for all the hosts you want to test, assuming you do not have DHCP and DNS configured.

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
10.20.1.101     sm-node-01
10.20.1.102     sm-node-02
10.20.1.103     sm-node-03
10.20.1.104     sm-node-04

192.168.30.101  sm-node-01-ib
192.168.30.102  sm-node-02-ib
192.168.30.103  sm-node-03-ib
192.168.30.104  sm-node-04-ib
```

Note When running on a network on which DHCP and DNS are configured, you should set the eth0 to DHCP and set */etc/sysconfig/network-scripts/ifcfg-ib0* to the IP address assigned by the DHCP and DNS servers. This can be discovered by typing the command ***nslookup sm-node-01-ib*** assuming that sm-node-01-ib is the name you have assigned to the IB card.

SSH Key Exchange

Using the root account for the ssh key exchange and testing is the simplest method but can lead to security concerns. If you cannot use root, then all the systems must be set up with at least one identical user account. The account must be able to ssh to all systems from the system which launches the Open MPI tests. This means that the ssh host keys must already be cached.

1. Connect the IB cards directly or through an IB switch and start opensm
2. Type the following command and accept all the defaults
    a. ssh-keygen
3. Do a copy to the other machine you want to share with. Below we assume you want nodes sm-node-01 and sm-node-02 and their IB nodes to share the keys
    a. ssh-copy-id -i .ssh/id_rsa.pub root@sm-node-01
    b. ssh-copy-id -i .ssh/id_rsa.pub root@sm-node-01-ib
    c. ssh-copy-id -i .ssh/id_rsa.pub root@sm-node-02
    d. ssh-copy-id -i .ssh/id_rsa.pub root@sm-node-02-ib

    If you have more than the two servers as in this example, then you would follow steps (2) and (3) for all the servers you want to use.
4. Test each one of your servers with the command *ssh root@sm-node-** so your keys get registered

Return to Top

Appendix B

Switch Setup

If you are using a managed Mellanox InfiniBand switch, you can change the link speed and other fields via the command line or through the web interface. If you are having problems accessing your Mellanox switch this link may be helpful: https://community.mellanox.com/docs/DOC-2172. We show how to discover the IP

address of a Mellanox switch and change the link speed using both the command line and the web interface here.

Command Line

There are two methods of passing commands to a Mellanox switch using a terminal. Remember the default login credentials of a Mellanox switch are -- **Account**: admin, **Password**: admin.

1. Using a serial connection tool (Tera Term) to connect to the switch via the console port.
    a. This can be used to ascertain the IP address of the management port connected to the network.
        i. Log in to switch
        ii. en[able]
        iii. con[figure] t[erminal]
        iv. show ip interface mgmt0

```
switch-EN [standalone: master] > show ip interface mgmt0
Interface mgmt0 status:
   Comment:
   Admin up:           yes
   Link up:            yes
   DHCP running:       no
   IP address:         10.20.0.190
   Netmask:            255.255.255.0
   IPv6 enabled:       yes
   Autoconf enabled:   no
   Autoconf route:     yes
   Autoconf privacy:   no
   DHCPv6 running:     no
   IPv6 addresses:     1
   IPv6 address:       fe80::7efe:90ff:fe28:cd56/64
   Speed:              1000Mb/s (auto)
   Duplex:             full (auto)
   Interface type:     ethernet
   Interface source:   bridge
   MTU:                1500
   HW address:         7C:FE:90:28:CD:56
```

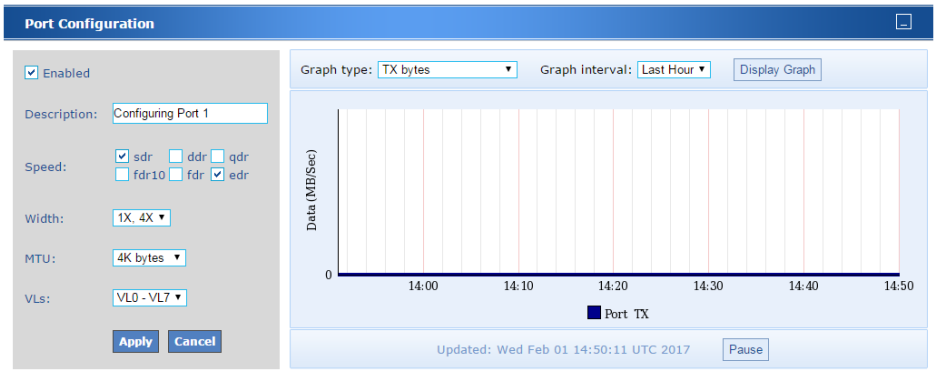
2. Using an SSH client to connect to the switch via the management IP interface. This assumes that you know the mgmt0 or mgmt1 IP address.
    a. ssh admin@*IP address of switch* Once connected, you should be presented with the same interface as shown in the serial port connection case above.

Web Interface

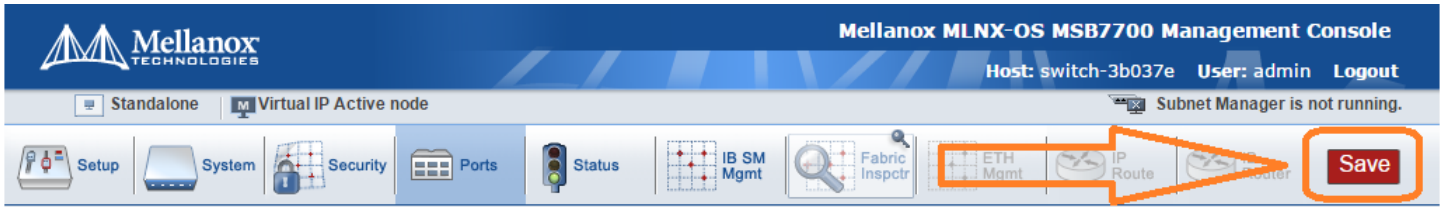
Get the switch's IP address (via Tera Term, for example), and login to the switch's web interface.

1. Click on the Ports tab and then select a port
2. Scroll down to the Port Configuration section of the page and change the link speed via the interface as shown in the screen below. The image below is for an EDR switch. While the interface is slightly

different for FDR switches, the fields you need to change will be obvious in either case. In the image below, we have selected SDR and EDR speeds.



3. After configuring all the fields you need, click Apply.
4. Make sure you **save** the settings you have just applied. If your switch loses power or is restarted it will lose any unsaved configuration.



Performance Testing

Users interested in high performance testing may wish to consider the [Software Forge HPC Performance Monitor]. This tool measures and monitors both RDMA and/or TCP connections.

[Return to Top]

Revision History

Revision	Date	Author	Comments
1.0.00	2017-03-28	Llolsten Kaonga	• Initial version of the IB Interop MOI re-write
1.0.01	2017-07-14	Llolsten Kaonga	• Update MOI for PF32
1.0.02	2018-02-12	Llolsten Kaonga	• Update MOI for PF33
1.0.03	2018-03-12	Llolsten Kaonga	• Added dracut command after OFED installation
1.0.04	2018-07-23	Llolsten Kaonga	• Update MOI for PF34
1.0.05	2019-01-24	Llolsten Kaonga	• Update MOI for PF35
1.0.06	2019-10-09	Llolsten Kaonga	• Update MOI for PF36
			•