

Infiniband

Christian Külker

2022-06-08

Contents

1	Install The Software	2
2	Test Software Installation	2
3	Find the GUID	4
3.1	ts2	4
4	Configure Opensm	4
4.1	Start Opensm	5
4.2	Verify That Opensm Is Started	5
5	Collect More Host Adapter Information	5
6	Check Extended Hosts On The Network	6
7	Check Switches	6
8	Setting Up IPoverIB	6
9	TCP Performance Tuning	6
10	Test The Connection With Ibping	7
11	Test A Port	7
12	Create A Topology Map File	7
13	Measure Bandwidth	7
14	Other Useful Commands	8
15	Troubleshooting Infiniband	8

16 Find Originally Programmed MAC Address	9
17 Command In Mellanox OFED Environmanets	9
18 History	9
19 Disclaimer of Warranty	10
20 Limitation of Liability	10

The computer networking standard InfiniBand (IB) is used in high-performance computing. It features a very high throughput and very low latency compared to Ethernet. It is used for data and communication interconnect between nodes (computers). InfiniBand can be used as a switched interconnect between nodes and storage or storage and storage.

As of 2014, it was the most commonly used interconnect in supercomputers were general solutions applied. Manly two companies Mellanox and Intel manufacture InfiniBand host bus adapters and network switches. In 2016 it was reported that also Oracle created its own version of InfiniBand switch units and server adapter chips.

Mellanox IB host adapters work with major Linux distributions: RHEL, SLES and Debian, but some might have better support than others for proprietary add-ons.

Infiniband, promoted by the InfiniBand Trade Association, is competing with other network interconnects like Fibre Channel, Intel Omni-Path as well as Ethernet.

The following tests have been performed on Debian and/or CentOS with Mellanox host adapters.

1 Install The Software

```
aptitude install opensm infiniband-diags perftest ibutils
```

2 Test Software Installation

Test if the host adapter is present

```
lspci -v | grep Mellanox
06:00.0 InfiniBand: Mellanox Technologies MT25208 InfiniHost III Ex (Tavor
compatibility mode) (rev 20)
```

In case it is not present or in doubt, sue `dmesg|grep ib`.

Check if kernel modules are loaded:

```
lsmod|grep mlx
mlx4_core                67736  0
```

Load Mellanox module:

```
modprobe mlx4_ib
lsmod|grep mlx4_ib
mlx4_ib                  33590  0
ib_mad                   30017  1 mlx4_ib
ib_core                  40999  2 mlx4_ib,ib_mad
mlx4_core                67736  1 mlx4_ib
lsmod|grep ib
mlx4_ib                  33590  0
ib_mad                   30017  1 mlx4_ib
ib_core                  40999  2 mlx4_ib,ib_mad
libata                   133808  2 ata_generic,ata_piix
mlx4_core                67736  1 mlx4_ib
scsi_mod                 126565  2 sd_mod,libata
```

Load Other IB Modules

```
modprobe ib_sdp
FATAL: Module ib_sdp not found
```

```
root@ts2:~# modprobe ib_srp
root@ts2:~# lsmod |grep ib_srp
ib_srp                   20480  0
scsi_transport_srp       3686  1 ib_srp
ib_cm                    26272  1 ib_srp
ib_sa                    16103  2 ib_srp,ib_cm
ib_core                  40999  5 ib_srp,ib_cm,ib_sa,mlx4_ib,ib_mad
scsi_mod                 126565  5 ib_srp,scsi_transport_srp,scsi_tgt,sd_mod,libata
```

```
root@ts2:~# modprobe ib_ipoib
root@ts2:~# lsmod |grep ib_ipoib
ib_ipoib                 59777  0
inet_lro                 4630  1 ib_ipoib
ib_cm                    26272  2 ib_ipoib,ib_srp
ib_sa                    16103  3 ib_ipoib,ib_srp,ib_cm
ib_core                  40999  6 ib_ipoib,ib_srp,ib_cm,ib_sa,mlx4_ib,ib_mad
```

```
root@ts2:~# modprobe ib_uverbs
root@ts2:~# lsmod |grep ib_uverbs
ib_uverbs          25658  0
ib_core            40999  7
↪ ib_uverbs,ib_ipoib,ib_srp,ib_cm,ib_sa,mlx4_ib,
                               ib_mad
```

```
root@ts2:~# modprobe ib_umad
root@ts2:~# lsmod |grep ib_umad
ib_umad            9879  0
ib_mad             30017  4 ib_umad,ib_cm,ib_sa,mlx4_ib
ib_core            40999  8
↪ ib_umad,ib_uverbs,ib_ipoib,ib_srp,ib_cm,ib_sa,
                               mlx4_ib,ib_mad
```

```
root@ts2:~# modprobe rdma_ucm
root@ts2:~# lsmod |grep rdma_ucm
rdma_ucm           9205  0
rdma_cm            20678  1 rdma_ucm
ib_uverbs          25658  1 rdma_ucm
ib_core            40999  11 rdma_ucm,rdma_cm,iw_cm,ib_umad,ib_uverbs,
                               ib_ipoib,ib_srp,ib_cm,ib_sa,mlx4_ib,ib_mad
```

3 Find the GUID

3.1 ts2

```
ibstat -p
0x002590ffff2e4f6d
```

4 Configure Opensm

SM stands for subnet manager. There different implementations and locations where subnet managers can be installed.

Per default it is started on all ports, open `/etc/default/opensm`

```
vim /etc/default/opensm
```

4.1 Start Opensm

```
/etc/init.d/opensm start
Starting opensm on 0x002590ffff2e4f6d:
```

4.2 Verify That Opensm Is Started

```
tail -f /var/log/syslog
Sep 20 18:10:40 ts2 OpenSM[3527]: /var/log/opensm.0x002590ffff2e4f6d.log
↪ log
file opened
Sep 20 18:10:40 ts2 OpenSM[3527]: OpenSM 3.2.6_20090317#012
Sep 20 18:10:40 ts2 OpenSM[3527]: Entering DISCOVERING state#012
Sep 20 18:10:40 ts2 OpenSM[3527]: SM port is down#012
```

If the above steps are done also on an other node the following message can be seen:

```
Sep 20 18:38:50 ts2 OpenSM[3527]: Entering MASTER state#012
Sep 20 18:38:50 ts2 OpenSM[3527]: SUBNET UP#012
```

5 Collect More Host Adapter Information

```
ibstat
CA 'mlx4_0'
  CA type: MT26428
  Number of ports: 1
  Firmware version: 2.7.200
  Hardware version: b0
  Node GUID: 0x002590ffff2e4f70
  System image GUID: 0x002590ffff2e4f73
  Port 1:
    State: Active
    Physical state: LinkUp
    Rate: 40
    Base lid: 2
    LMC: 0
    SM lid: 1
    Capability mask: 0x0251086a
    Port GUID: 0x002590ffff2e4f71
```

6 Check Extended Hosts On The Network

```
ibhosts
Ca      : 0x002590ffff2e4f6c ports 1 "MT25408 ConnectX Mellanox
↳ Technologies"
Ca      : 0x002590ffff2e4f70 ports 1 "MT25408 ConnectX Mellanox
↳ Technologies"
```

7 Check Switches

As the time of writing I had not a switch attached. Usually there is a long output.

```
ibswitches
```

```
iblinkinfo
```

Other low level information can be obtained by the `sys` file system

```
1 /sys/class/infiniband/DEVICE_NAME
```

8 Setting Up IPoverIB

Infiniband can be used without IP, but for many application it is easier to use IPoverIB. iSCSIoverIB is not covered here.

Check that the module is loaded:

```
modprobe ib_ipoib |grep ib_ipoib
```

This shows nothing but the following should be showing something

```
ifconfig -a |grep ib
ib0    Link encap:UNSPEC  HWaddr
↳ 80-00-00-48-FE-80-00-00-00-00-00-00-00-00-00-00
```

9 TCP Performance Tuning

In order to obtain maximum IPoIB throughput you may need to tweak the MTU and various kernel TCP buffer and window settings. (Jumbo frames) See the details in the `ipoib_release_notes.txt` document in the `ofed-docs` package.

10 Test The Connection With Ibping

Start on one node

```
ibping -S
```

Start on the other node

```
ibping -G 0x002590ffff2e4f6d
```

See:

```
1 Pong from ts2.(none) (Lid 1): time 0.302 ms
```

11 Test A Port

```
smpquery portinfo 24 24
```

12 Create A Topology Map File

```
osmtest -f c -i inventory.txt
cat inventory.txt |grep -e '^lid' -e 'port_guid' -e 'desc'|sed
↪ 's/^lid/\nlid/'\
> mapping.txt
```

13 Measure Bandwidth

One node (nodeABC) start

```
ib_write_bw
```

On a different node start

```
ib_write_bw nodeABC
```

```
-----
                        RDMA_Write BW Test
Number of qps      : 1
Connection type    : RC
TX depth           : 300
```

```
CQ Moderation    : 50
Mtu               : 2048B
Link type        : IB
Max inline data   : 0B
rdma_cm QPs      : OFF
Data ex. method  : Ethernet

-----
local address: LID 0x01 QPN 0x0053 PSN 0xc1675c RKey 0x002d00 VAddr
0x002aaaaaae1000
remote address: LID 0x15 QPN 0x0052 PSN 0xd79849 RKey 0x002000 VAddr
0x002aaaaaae1000
-----

#bytes    #iterations    BW peak[MB/sec]    BW average[MB/sec]
65536     5000             939.37             939.37
-----
```

An other method is: (UNTESTED)

On host A:

```
rdma_bw -b
```

On host B:

```
rdma_bw -b nodeABC
```

14 Other Useful Commands

```
ibnetdiscover
```

```
opensm
```

```
/usr/sbin/ibstatus
```

15 Troubleshooting Infiniband

There many ways to troubleshoot Infiniband and the topic it self can fill a book, a quick start point is to use `libdiagnet` .

```
mkdir libdiagnet
cd libdiagnet
ibdiagnet -ls 10 -lw 4x -vlr > ibdiagnet.out
```

Maybe you have to do this over again. In this case (1) reset the error counters.

```
ibdiagnet -pc
```

And (2) stress the network with benchmark like Intel **IMB-MPI1** benchmark over `mvapich` MPI or other network heavy load. Then (3) read out error counters again. After finish to run the test on all network nodes, run `ibdgnet` again.

```
ibdiagnet -P all=1
```

16 Find Originally Programmed MAC Address

```
ip addr|grep 'link/infiniband'|sed -s 's%.*link/infiniband \
80:00:00:48:fe:80:00:00:00:00:00:00:\(.*\)00:01 brd.%.%11%
```

17 Command In Mellanox OFED Environmanets

Mellanox OFED for Linux is provided as ISO images, one per supported Linux distribution and CPU architecture, that includes source code and binary RPMs, firm-ware, utilities, and documentation. This image also contains firmware.

```
1  ibv_devinfo
2  mlxburn
3  flint
4  spark
```

18 History

Version	Date	Notes
0.1.2	2022-06-08	shell->bash, +history
0.1.1	2020-09-05	
0.1.0	2020-05-18	Initial release

19 Disclaimer of Warranty

THERE IS NO WARRANTY FOR THIS INFORMATION, DOCUMENTS AND PROGRAMS, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE INFORMATION, DOCUMENT OR THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE INFORMATION, DOCUMENTS AND PROGRAMS IS WITH YOU. SHOULD THE INFORMATION, DOCUMENTS OR PROGRAMS PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

20 Limitation of Liability

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE INFORMATION, DOCUMENTS OR PROGRAMS AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE INFORMATION, DOCUMENTS OR PROGRAMS (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE INFORMATION, DOCUMENTS OR PROGRAMS TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.