Installing and Testing OFED in LINUX

Check the Installation

- ibstat

```
CA 'mlx4_0'
        CA type: MT26438
        Number of ports: 1
        Firmware version: 2.7.9294
        Hardware version: b0
        Node GUID: 0x78e7d1030003ccb0
        System image GUID: 0x78e7d1030003ccb3
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 40
                Base lid: 49
                LMC: 0
                SM lid: 1
                Capability mask: 0x02510868
                Port GUID: 0x78e7d1030003ccb1
                Link layer: IB
```

Check the Installation

```
# cat /sys/class/infiniband/mlx4_0/ports/1/rate
DDR : 20 Gb/sec (4X DDR)

QDR : 40 Gb/sec (4X QDR)


# cat /sys/class/infiniband/mlx4_0/ports/1/state
4: ACTIVE

# cat /sys/class/infiniband/mlx4_0/ports/1/phys_state
5:Link Up

# cat /sys/class/infiniband/mlx4_0/board_id
HP_0190000003

#cat /sys/class/infiniband/mlx4_0/fw_ver
2.7.9294


Counters located at :
                    /sys/class/infiniband/mthca0/ports/1/counters
```

Check the Installation : ibstatus

```
# ibstatus
[root@o187i164 ~]# ibstatus
Infiniband device 'mlx4_0' port 1 status:
        default gid:     fe80:0000:0000:0000:78e7:d103:0003:cb41
        base lid:        0x1d
        sm lid:          0x1
        state:           4: ACTIVE
        phys state:      5: LinkUp
        rate:            40 Gb/sec (4X QDR)
        link_layer:      IB

[root@o187i164 ~]#
```

Check the Installation: ibv_devinfo

```
[root@o187i164 ~]# ibv_devinfo
hca_id: mlx4_0
        transport:                      InfiniBand (0)
        fw_ver:                         2.7.9294
        node_guid:                      78e7:d103:0003:cb40
        sys_image_guid:                 78e7:d103:0003:cb43
        vendor_id:                      0x02c9
        vendor_part_id:                 26438
        hw_ver:                         0xB0
        board_id:                       HP_0190000003
        phys_port_cnt:                  1
                port:   1
                        state:                  PORT_ACTIVE (4)
                        max_mtu:                2048 (4)
                        active_mtu:             2048 (4)
                        sm_lid:                 1
                        port_lid:               29
                        port_lmc:               0x00
                        link_layer:             IB

[root@o187i164 ~]#
```

Check the Installation: ibv_devices

```
#ibv_devices

    device                 node GUID
    ------              ----------------
    mlx4_0              00237dffff94fa2c
```

Check the installation : **ifconfig**

```
# ifconfig ib0
ib0       Link encap:InfiniBand  HWaddr 80:00:00:48:FE:80:00:00:00:00:00:00:00
          inet addr:172.23.0.234  Bcast:172.23.255.255  Mask:255.255.0.0
          inet6 addr: fe80::223:7dff:ff94:fa2d/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
          RX packets:54451217 errors:0 dropped:0 overruns:0 frame:0
          TX packets:57972322 errors:0 dropped:10 overruns:0 carrier:0
          collisions:0 txqueuelen:256
          RX bytes:13419956118 (12.4 GiB)  TX bytes:14717506554 (13.7 GiB)
```

Check the network : ibnodes

- ibnodes – shows all the CA in the system

```
# [root@o187i164 ~]# ibnodes
Ca      : 0x78e7d1030003a2e0 ports 1 "o187i178 HCA-1"
Ca      : 0x78e7d10300035e20 ports 1 "o187i177 HCA-1"
Ca      : 0x78e7d10300031c00 ports 1 "o187i176 HCA-1"
Ca      : 0x78e7d103000339e0 ports 1 "o187i175 HCA-1"
Ca      : 0x78e7d10300039b00 ports 1 "o187i174 HCA-1"
Ca      : 0x78e7d1030003cc20 ports 1 "o187i173 HCA-1"
Ca      : 0x78e7d1030003cca0 ports 1 "o187i172 HCA-1"
Ca      : 0x78e7d10300037f80 ports 1 "o187i171 HCA-1"
Ca      : 0x78e7d10300032e40 ports 1 "o187i170 HCA-1"
Ca      : 0x78e7d10300032840 ports 1 "o187i169 HCA-1"
Ca      : 0x78e7d1030003ed00 ports 1 "o187i168 HCA-1"
Ca      : 0x78e7d10300030dc0 ports 1 "o187i167 HCA-1"
Ca      : 0x78e7d1030003cb60 ports 1 "o187i166 HCA-1"
Ca      : 0x78e7d10300039b20 ports 1 "o187i165 HCA-1"
Ca      : 0x78e7d1030003fe80 ports 1 "o187i163 HCA-1"
Ca      : 0x78e7d1030003cb40 ports 1 "o187i164 HCA-1"
Switch  : 0x0002c902004260c8 ports 32 "Infiniscale-IV Mellanox Technologies" base port 0
lid 2 lmc 0
Switch  : 0x0008f105001088f2 ports 36 "Voltaire 4036 # 4036-88F2" enhanced port 0 lid 1
lmc 0
Switch  : 0x0002c9020042b118 ports 32 "Infiniscale-IV Mellanox Technologies" base port 0
lid 3 lmc 0
[root@o187i164 ~]#
```

Check the network : ibswitches

- ibswitches – shows all the Switches in the system

```
[root@o187i164 ~]# ibswitches
Switch  : 0x0002c902004260c8 ports 32 "Infiniscale-IV Mellanox Technologies" base port 0
lid 2 lmc 0
Switch  : 0x0008f105001088f2 ports 36 "Voltaire 4036 # 4036-88F2" enhanced port 0 lid 1
lmc 0
Switch  : 0x0002c9020042b118 ports 32 "Infiniscale-IV Mellanox Technologies" base port 0
lid 3 lmc 0
[root@o187i164 ~]#
```

Check the network : ibhosts

- ibhosts – shows all the hosts (no SWTICHES) in the system

```
[root@o187i164 ~]# ibhosts
Ca      : 0x78e7d1030003a2e0 ports 1 "o187i178 HCA-1"
Ca      : 0x78e7d10300035e20 ports 1 "o187i177 HCA-1"
Ca      : 0x78e7d10300031c00 ports 1 "o187i176 HCA-1"
Ca      : 0x78e7d103000339e0 ports 1 "o187i175 HCA-1"
Ca      : 0x78e7d10300039b00 ports 1 "o187i174 HCA-1"
Ca      : 0x78e7d1030003cc20 ports 1 "o187i173 HCA-1"
Ca      : 0x78e7d1030003cca0 ports 1 "o187i172 HCA-1"
Ca      : 0x78e7d10300037f80 ports 1 "o187i171 HCA-1"
Ca      : 0x78e7d10300032e40 ports 1 "o187i170 HCA-1"
Ca      : 0x78e7d10300032840 ports 1 "o187i169 HCA-1"
Ca      : 0x78e7d1030003ed00 ports 1 "o187i168 HCA-1"
Ca      : 0x78e7d10300030dc0 ports 1 "o187i167 HCA-1"
Ca      : 0x78e7d1030003cb60 ports 1 "o187i166 HCA-1"
Ca      : 0x78e7d10300039b20 ports 1 "o187i165 HCA-1"
Ca      : 0x78e7d1030003fe80 ports 1 "o187i163 HCA-1"
Ca      : 0x78e7d1030003cb40 ports 1 "o187i164 HCA-1"
[root@o187i164 ~]#
```

Check the network : IBPING using LIDS

- Attention you will need to open 2 connections to 2 compute nodes.

- Step 1
  - Execute ibstat on **NODE 1**
- [root@cne01 ~]# ibstat
- CA 'mlx4_0'
- CA type: MT25418
- Number of ports: 2
- Firmware version: 2.2.0
- Hardware version: a0
- Node GUID: 0x001b78ffff34f8e4
- System image GUID: 0x001b78ffff34f8e7
- Port 1:
- State: Active
- Physical state: LinkUp
- Rate: 10
- **Base lid: 5**
- LMC: 0
- SM lid: 5
- Capability mask: 0x0251086a
- Port GUID: 0x001b78ffff34f8e5

- Step 2
  - Execute ibping on **NODE 1** using –S option

- [root@cne01 ~]# ibping –S

- Step 3
  - Execute ibping **on NODE 2** with the LID of the first node

  - [root@cne02 ~]# ibping 5
  - Pong from hpc4.(none) (Lid 5): time 0.083 ms
  - Pong from hpc4.(none) (Lid 5): time 0.080 ms
  - Pong from hpc4.(none) (Lid 5): time 0.067 ms

Check the nodes : perfquery

- perfquery – checks the local HCA

```
[root@o187i164 ~]# perfquery
# Port counters: Lid 29 port 1
PortSelect:......................1
CounterSelect:...................0x0400
SymbolErrors:....................0
LinkRecovers:....................0
LinkDowned:......................0
RcvErrors:.......................0
RcvRemotePhysErrors:.............0
RcvSwRelayErrors:................0
XmtDiscards:.....................0
XmtConstraintErrors:.............0
RcvConstraintErrors:.............0
CounterSelect2:..................0x00
LinkIntegrityErrors:.............0
ExcBufOverrunErrors:.............0
VL15Dropped:.....................0
XmtData:.........................150783
RcvData:.........................150104
XmtPkts:.........................6032
RcvPkts:.........................6042
[root@o187i164 ~]#
```

Check the state: Ibcheckstate

- ibcheckstate - check the state CA and Switches

```
[root@o187i164 ~]# ibcheckstate

## Summary: 19 nodes checked, 0 bad nodes found
##          48 ports checked, 0 ports with bad state found
[root@o187i164 ~]#
```

Commands – diagnosis

- ibdiagnet – Check all the net for errors

```
# ibdiagnet

Loading IBDIAGNET from: /usr/lib64/ibdiagnet1.2
-W- Topology file is not specified.
    Reports regarding cluster links will use
direct routes.
Loading IBDM from: /usr/lib64/ibdm1.2
-I- Using port 1 as the local port.
-I- Discovering ... 25 nodes (3 Switches & 22
CA-s) discovered.
```

Commands – diagnosis

- ibdiagnet – Easier version

```
[root@o187i188 ~]# mkdir /tmp/test

[root@o187i188 ~]# ibdiag
ibdiagnet   ibdiagpath  ibdiagui
[root@o187i188 ~]# ibdiagnet -o /tmp/test/
Loading IBDIAGNET from: /usr/lib64/ibdiagnet1.5.6
-W- Topology file is not specified.
    Reports regarding cluster links will use direct routes.

[root@o187i188 ~]# cd /tmp/test/
[root@o187i188 test]# ls

ibdiagnet.db  ibdiagnet.fdbs  ibdiagnet_ibis.log
ibdiagnet.log  ibdiagnet.lst  ibdiagnet.mcfdbs  ibdiagnet.pkey
ibdiagnet.sm

[root@o187i188 test]# cat ibdiagnet.sm
ibdiagnet fabric SM report

  SM - master
    Port=2 lid=0x0001 guid=0x0008f105001088f2 dev=23130
priority:4
```

Commands – diagnosis

- ibdiagnet – Check the Speed and the Width

```
[root@o187i169 ~]# ibdiagnet -ls 5 -lw 4x
```

```
What do you find in the section
I---------------------------------------------------
-I- Links With links speed != 5 (as set by -ls option)
-I---------------------------------------------------
```

```
WHY YOU HAVE SO MANY LINKS?
```

Commands – check the net

- Ibcheckwidth – check if all the nodes have 4X

```
[root@o187i164 ~]# ibcheckwidth

## Summary: 19 nodes checked, 0 bad nodes found
##          48 ports checked, 0 ports with 1x width in error found
[root@o187i164 ~]#
```

Check bandwidth btw nodes

Again you will need 2 connections to 2 nodes

- ib_read_bw

```
[root@o187i169 ~]# ib_read_bw
---------------------------------------------------------------------
                    RDMA_Read BW Test
 Number of qp's running 1
 Connection type : RC
 Each Qp will post up to 100 messages each time
 Inline data is used up to 0 bytes message
 Link type is IB
 Mtu : 2048
---------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]
---------------------------------------------------------------------
```

```
[root@o187i187 ~]# ib_read_bw 10.0.0.169
---------------------------------------------------------------------
                    RDMA_Read BW Test
 Number of qp's running 1
 Connection type : RC
 Each Qp will post up to 100 messages each time
 Inline data is used up to 0 bytes message
 Link type is IB
 Mtu : 2048
---------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]
65536      5000           3250.91            3258.14
---------------------------------------------------------------------
```

Check latency btw nodes

Again you will need 2 connections to 2 nodes

- ib_read_lat

```
[root@o187i169 ~]# ib_read_lat
------------------------------------------------------------------
------------------------------------------------------------------
                    RDMA_Read Latency Test
 Connection type : RC
 Link type is IB
 Mtu : 2048
 Number of outstanding reads is 16
 local address: LID 0x17 QPN 0x440049 PSN 0xfda565 OUT 0x10 RKey 0x18001b03 VAddr 0x0000000148c000
 remote address: LID 0x2a QPN 0x180049 PSN 0x2bc0d0 OUT 0x10 RKey 0x20001b00 VAddr 0x000000022ff000
------------------------------------------------------------------
```

```
[root@o187i187 ~]# ib_read_lat 10.0.0.169
------------------------------------------------------------------
                    RDMA_Read Latency Test
 Connection type : RC
 Link type is IB
 Mtu : 2048
 Number of outstanding reads is 16
 local address: LID 0x2a QPN 0x180049 PSN 0x2bc0d0 OUT 0x10 RKey 0x20001b00 VAddr 0x000000022ff000
 remote address: LID 0x17 QPN 0x440049 PSN 0xfda565 OUT 0x10 RKey 0x18001b03 VAddr 0x0000000148c000
------------------------------------------------------------------
 #bytes #iterations    t_min[usec]    t_max[usec]  t_typical[usec]
Warning: measured timestamp frequency 2666.29 differs from nominal 1600 MHz
 2       1000          2.81           35.85        2.84
------------------------------------------------------------------
```

Ping Pong Test

Again you will need 2 connections to 2 nodes

The ping-pong example tests provide basic connectivity tests. Each test has a help message (-h).

- ibv_ud_pingpong
- ibv_rc_pingpong
- ibv_srq_pingpong
- ibv_uc_pingpong

Example: ibv_ud_pingpong -h

Usage:

ibv_ud_pingpong NODE 1 start a server and wait for connection

ibv_ud_pingpong <host> NODE 2 connect to server at <host>

Run them on your cluster and RECORD THE OBTAINED VALUES

RUNNING PING_PONG_RING

Latency: 0.12 usec

```
Host 0 -- ip 16.16.187.163 -- ranks 0 - 4

 host | 0
======|======
    0 : SHM

 Prot -  All Intra-node communication is: SHM
```

BW: 2.8 GB/s

Latency: 19 usec

```
Host 0 -- ip 16.16.187.164 -- ranks 0
Host 1 -- ip 16.16.187.165 -- ranks 1
Host 2 -- ip 16.16.187.166 -- ranks 2
Host 3 -- ip 16.16.187.167 -- ranks 3
Host 4 -- ip 16.16.187.168 -- ranks 4

 host | 0    1    2    3    4
======|=========================
    0 : SHM  TCP  TCP  TCP  TCP
    1 : TCP  SHM  TCP  TCP  TCP
    2 : TCP  TCP  SHM  TCP  TCP
    3 : TCP  TCP  TCP  SHM  TCP
    4 : TCP  TCP  TCP  TCP  SHM

 Prot -  All Intra-node communication is: SHM
 Prot -  All Inter-node communication is: TCP
```

BW: 100 MB/s

Latency: 1.60 usec

```
Host 0 -- ip 16.16.187.164 -- ranks 0
Host 1 -- ip 16.16.187.165 -- ranks 1
Host 2 -- ip 16.16.187.166 -- ranks 2
Host 3 -- ip 16.16.187.167 -- ranks 3
Host 4 -- ip 16.16.187.168 -- ranks 4

 host | 0    1    2    3    4
======|=========================
    0 : SHM  IBV  IBV  IBV  IBV
    1 : IBV  SHM  IBV  IBV  IBV
    2 : IBV  IBV  SHM  IBV  IBV
    3 : IBV  IBV  IBV  SHM  IBV
    4 : IBV  IBV  IBV  IBV  SHM

 Prot -  All Intra-node communication is: SHM
 Prot -  All Inter-node communication is: IBV
```

BW: 3.1 GB/s

Installation of the opensm

- On at least two nodes make sure to install the following packages
  - *opensm*
  - *opensm-libs*
  - *opensm-static*
  - *opensm-devel*
  - *Dependencies :* *glibc-devel-2.5-34*
    glibc-2.5-34

- Check the settings via the /etc/opensm.conf
- On at least two nodes start the SM by typing
  - opensm

Check the logs /var/log/opensm.log to see which one is the Master and the Stand-by

Starting opensm

Open 2 connections to your headnode

- how to start:
- [root@o187i187 ~]# touch/var/log/opensm.log (WINDOW 1)
- [root@o187i187 ~]# tail –f /var/log/opensm.log (WINDOW 1)
- [root@o187i187 ~]# opensm -pXXX (WINDOW 2)

- START ONE AT THE TIME – WAIT FOR MY SIGNAL ☺

- TEAM A : Priority 5 → XXX=5
- TEAM B : Priority 8 → XXX=8
- TEAM C: Priority 10 → XXX=10
- TEAM D: Priority 13 → XXX=13
- TEAM E: Priority 15 → XXX=15

What HAPPENS ?

Starting opensm

Open 2 connections to your headnode

- how to start:
- [root@o187i187 ~]# touch/var/log/opensm.log (WINDOW 1)
- [root@o187i187 ~]# tail –f /var/log/opensm.log (WINDOW 1)
- [root@o187i187 ~]# opensm -pXXX (WINDOW 2)

- KILL THEM ONE AT THE TIME – WAIT FOR MY SIGNAL ☺

- TEAM E : Press “CTRL-C”
- TEAM D : Press “CTRL-C”
- TEAM C: Press “CTRL-C”
- TEAM B: Press “CTRL-C”
- TEAM A: Press “CTRL-C”

What HAPPENS ?