

**IBSwitches.com**

The Infiniband Price / Performance Leader

Download :**Installation Guide
User Guide
OFED Linux Stack
Switch User Guide****Cluster Kits
available on-line :****www.IBSwitches.com**

Building an Infiniband Cluster

Required Material before getting started

- Download Mellanox OFED for Linux Installation Guide
 - http://www.mellanox.com/pdf/products/software/Mellanox_OFED_Installation_Guide_1_20.pdf
- Download Mellanox OFED Stack Linux User's Guide
 - http://www.mellanox.com/pdf/products/software/Mellanox_OFED_Linux_User_Manual_1_20.pdf
- Download Appropriate Linux OFED Stack
 - <http://www.mellanox.com/products/ofed.php>
 - Download links for Red Hat and SLES are at the bottom of the linked page above
- Download Flextronics Switch User's Manual
 - <http://www.ibswitches.com/support>

Step 1 (Order necessary equipment)

In order to build an Infiniband cluster it is most helpful to know what components are required , as well as the corresponding quantities to build a cluster of specific size. To simply things, one can visit the website www.ibswitches.com and order cluster kits which come in sizes of 8 and 24 nodes. The 8 node cluster consists of one Flextronics 8-Port SDR switch, 8 Mellanox Single Port SDR HCA's, and 8 Infiniband Cables. The 24 Node Cluster kit consists of (1) Flextronics 24-Port DDR switch, (24) Mellanox Single Port DDR HCA's, and 24 Infiniband cables.

Step 2 (Mount equipment in Rack)

Infiniband clusters are typically built using 1U or 2U servers mounted on 19" racks. The typical rack size is 42U which we will assume throughout this tutorial. One common sense practice to consider when assembling your servers and associated switches in to the rack, if possible mount switches in the middle and an equal number of servers above and below, this will allow for the shortest cables.

Step 3 (Connect Servers and Switch with IB Cables)

Install Infiniband cables between every server and a corresponding port on the switch.

For details on loading the OFED Stack follow instructions in OFED Linux Installation Guide

Components of the OFED Linux Stack for IB

OpenSM is the open source Subnet Manager included in the OFED Stack

Step 4 (Loading OFED Stack)

- Load OFED Driver on each and every Linux server
- Download appropriate OFED Stack from Mellanox website
 - Mount ISO on shared Drive / Directory
 - Run install script
 - Syntax : `mlxofedinstall`

Important Note : Refer to Mellanox OFED for Linux Installation Guide for detail step by step instructions in loading OFED stack

MLNX OFED Stack contains the following software components:

- HCA drivers
 - mthca, mlx4
- Mid-layer core
 - Verbs, MADs, SA, CM, CMA, uVerbs, uMADs
- Upper Layer Protocols (ULPs)
 - IPoIB, RDS, SDP, SRP Initiator
- MPI
 - Open MPI stack supporting the InfiniBand interface
 - OSU MVAPICH stack supporting the InfiniBand interface
 - MPI benchmark tests (OSU BW/LAT, Intel MPI Benchmark, Presta)
- OpenSM : Infiniband Subnet Manager
- Utilities
 - Diagnostic tools
 - Performance tests
- Firmware tools
- (MFT)Source code for all the OFED software modules (for use under the conditions mentioned in the modules' LICENSE files)
- Documentation

Step 5 (Starting OpenSM)

- Every Infiniband cluster must run at least one subnet management agent somewhere on the fabric
- Majority of clusters use un-managed switches where the subnet Manager runs on one of the servers
- In this case it is necessary to designate one of the servers as the head node and start up the OpenSM subnet manager provided in the OFED ISO.

Default values for OpenSM typically sufficient for small clusters

For detailed description of OpenSM options and features, see Section 9 of OFED Linux Users Manual

Verify all Server and Switch Ports are up and running at the correct speed



Step 5 (Continued)

- For most sub 100 node clusters the defaults set in the OpenSM program are sufficient. In this case simple execute
 - Syntax
 - **Opensm**
- A handy way to ensure the head node runs OpenSM automatically after a re-boot is to edit the OpenSM config file on the head node as follows :
 - Edit file —> **/etc/ofa/opensm.conf**
 - Change to **onboot = yes**
- OpenSM has developed in to a very sophisticated and powerful Subnet Management program and as such has many different options that can be experimented with in order to achieve maximum performance.
 - Section 9 of the Mellanox OFED Linux Users Manual has detailed information on each option available
- After having installed OFED on all the servers, and initiating OpenSM on at least one node, it is now time to bring up the cluster

Step 6 (Initial Bring Up)

- Reboot all Servers after loading OFED stack on each machine
- Each HCA port and Switch port will have two status LED's
 - One LED provides status on the Physical Link
 - The other LED provides status on the Logical Link, does the Subnet Manager acknowledge the port as initialized.
- All LED's designating the Physical ports should be illuminated GREEN on both the HCA and Switch port
 - If a particular LED designating a Physical port does not illuminate GREEN, remove cable from both server side and switch side, note this is a suspect cable.
 - Unplug a known good cable from another server/switch pair.
 - Take the known good cable and plug it in to the suspect server and switch port
 - If both Server and Switch Port Lights now illuminate GREEN
 - Cable could be bad, check on other ports to verify
 - Mark this cable as suspect even if it checks out good (easiest way to ID flakey cables)
- Continue above operation until all Physical port LED's on applicable HCA's and Switch Ports illuminate green



Verify all ports are connected and running at the appropriate throughput

Diagnostic tools are preloaded with OFED Stack Install

For detailed description of diagnostic tools options and features, see Section 10 of OFED Linux Users Manual

In this example, Port 1 is up and running with 4X SDR link. Port 2 is not connected

Step 6 (Continued)

- All LED's designating the Logical ports should be illuminated YELLOW on both the HCA and Switch port
 - If a particular LED designating a Logical port does not illuminate YELLOW, it is possible that the OFED stack did not install properly
 - For additional troubleshooting refer to diagnostic section below.
- The Logical port LED's illuminate solid YELLOW when no traffic is running, they rapidly blink YELLOW when traffic is flowing on the port
- When all Physical Ports are illuminated GREEN and all Logical Ports are illuminated YELLOW, it is time to run further tests verifying configuration, stability, and performance of the cluster

Step 7 (Running Diagnostics)

- Once OFED has successfully installed on each server, a head node has been designated and is running OpenSM, we can run diagnostics to verify all connections are functional and operating as expected
- All the following Diagnostic Utilities are loaded with OFED Stack

1. Run **ibstat** on each server with a HCA installed

SCREEN DUMP from **ibstat**

```
[root@mtlab32 ofa]# ibstat
CA 'mlx4_0'
CA type: MT25408
Number of ports: 2
Firmware version: 2.5.0
Hardware version: a0
Node GUID: 0x00000002c9002138
System image GUID: 0x00000002c900213b
Port 1:
  State: Active
  Physical state: LinkUp
  Rate: 10
  Base lid: 1
  LMC: 0
  SM lid: 4
  Capability mask: 0x02510868
  Port GUID: 0x00000002c9002139
Port 2:
  State: Down
  Physical state: Polling
  Rate: 10
  Base lid: 0
  LMC: 0
  SM lid: 0
  Capability mask: 0x02510868
  Port GUID: 0x00000002c900213a
```

Note : Firmware loaded on HCA

State of Logical Link
YELLOW LED

State of Physical Link
GREEN LED

Port Configuration
2 = 1X SDR (2.5Gb/s)
5 = 1X DDR (5 Gb/s)
10 = 4X SDR (10Gb/s)
20 = 4X DDR (20Gb/sec)
40 = 4X QDR (40Gb/s)

Step 7 (Continued)

2. Run `ibdiagnet -ls XX -lw XX`

Tells diag tool we are expecting SDR links (5 for DDR, 10 for QDR)

SCREEN DUMP from `ibdiagnet -ls 2.5 -lw 4x`

```
Root> ibdiagnet -ls 2.5 -lw 4x
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 ... 3 nodes (1 Switches & 2 CA-s) discovered.
-I- Discovering ... 6 nodes (1 Switches & 5 CA-s) discovered.
```

Tells diag tool we are expecting 4x links (other options : 1x and 12x)

```
-I-----
-I- Bad GUIDs/LIDs Info
-I-----
-W- Found Device with SystemGUID=0x0000000000000000:
   a Switch PortGUID=0x000b8cfff0020c1 at direct path="1"
```

Warnings can often be ignored

```
-I-----
-I- Links With Logical State = INIT
-I-----
-I- No bad Links (with logical state = INIT) were found
```

```
-I-----
-I- Bad Fabric SM Info
-I-----
-E- Missing master SM in the discover fabric
```

Errors must be addressed, in this case no subnet manager running

```
-I-----
-I- PM Counters Info
-I-----
-I- No illegal PM counters values were found
```

```
-I-----
-I- Links With links width != 4x (as set by -lw option)
-I-----
-I- No unmatched Links (with width != 4x) were found
```

Verified 4x links as expected from command line

```
-I-----
-I- Links With links speed != 2.5 (as set by -ls option)
-I-----
-I- No unmatched Links (with speed != 2.5) were found
```

Verified SDR link speed as expected from command line

```
-I-----
-I- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-I-----
-I- PKey:0x7fff Hosts:5 full:5 partial:0
```

```
-I-----
-I- IPoIB Subnets Check
-I-----
```

```
-I-----
-I- Bad Links Info
-I-----
-I- No bad link were found
-I-
```

SCREEN DUMP Continued

```

-----
-|- Links With Logical State = INIT
-|------
-|- No bad Links (with logical state = INIT) were found

-|------
-|- Bad Fabric SM Info
-|------
-|- E- Missing master SM in the discover fabric

-|------
-|- PM Counters Info
-|------
-|- No illegal PM counters values were found

-|------
-|- Links With links width != 4x (as set by -lw option)
-|------
-|- No unmatched Links (with width != 4x) were found

-|------
-|- Links With links speed != 2.5 (as set by -ls option)
-|------
-|- No unmatched Links (with speed != 2.5) were found

-|------
-|- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-|------
-|- PKey:0x7fff Hosts:5 full:5 partial:0

-|------
-|- IPoB Subnets Check
-|------

-|------
-|- Bad Links Info
-|------
-|- No bad link were found

-|- Stages Status Report:
STAGE      Errors Warnings
Bad GUIDs/LIDs Check      1      0
Link State Active Check   0      0
SM Info Check              1      0
Performance Counters Report 0      0
Specific Link Width Check  0      0
Specific Link Speed Check  0      0
Partitions Check           0      0
IPoB Subnets Check       0      0
    
```

Summary of results :

Says there are two errors while in actuality there was one error and one warning.

This can be verified by looking at details above as noted

In this case I would start OpenSM on the head node and run the ibdiagnet again to verify error condition went away

Please see /tmp/ibdiagnet.log for complete log

Step 7 (Continued)

- Verify expected bandwidth
 - Performance tests need to be run on two different machines so traffic can be passed between them
- 3. Run `ib_write_bw` on machine 1
 Run `ib_write_bw` IP ADDR on machine 2

SCREEN DUMP from `ib_write_bw`

`[root@mtilab32 ~]# ib_write_bw` ← Command initiated on Machine 1

```
-----
RDMA_Write 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 400 bytes message
 local address: LID 0x01, QPN 0x8004a, PSN 0x4f676 RKey 0x90002400 VAddr
0x002aaaab522000
 remote address: LID 0x02, QPN 0x6004a, PSN 0x751bb0, RKey 0x5a002400 VAddr
0x002aaaaaef000
Mtu : 2048
-----
```

```
#bytes #iterations BW peak[MB/sec] BW average[MB/
sec]
```

← No results initially as command on Machine 2 has not been started

`[root@mtilab31 ~]# ib_write_bw 10.2.2.32`

```
-----
RDMA_Write 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 1 bytes message
 local address: LID 0x02, QPN 0x6004a, PSN 0x751bb0 RKey 0x5a002400 VAddr
0x002aaaaaef000
 remote address: LID 0x01, QPN 0x8004a, PSN 0x4f676, RKey 0x90002400 VAddr
0x002aaaab522000
Mtu : 2048
-----
```

← Command initiated on machine 2, note addition of IP address

```
#bytes #iterations BW peak[MB/sec] BW average[MB/sec]
65536 5000 935.95 935.93
```

← Results of 935MB consistent with expectations for 4x SDR

- Bandwidth, will vary depending on chipset, memory, and CPU
- Some basic guidelines are as follows
 - DDR and PCIe Gen 1, expected unidirectional full wire speed bandwidth is 1400 MB/sec
 - DDR and PCIe Gen 2, expected unidirectional full wire speed bandwidth is 1800 MB/sec
 - QDR and PCIe Gen 2, expected unidirectional full wire speed bandwidth is 3000 MB/sec

For additional details and specifics on Infiniband Performance Troubleshooting, refer to Appendix B of the OFED Linux User's Guide

Step 7 (Continued)

- Verify expected latency
 - Performance tests need to be run on two different machines so traffic can be passed between them
- 4. Run `ib_write_lat` on machine 1
Run `ib_write_lat` IP ADDR on machine 2

SCREEN DUMP from `ib_write_lat`

```
[root@mtilab32 ~]# ib_write_bw
-----
RDMA_Write Latency Test
Inline data is used up to 400 bytes message
Connection type : RC
  local address: LID 0x01, QPN 0x8004a, PSN 0x4f676 RKey 0x90002400 VAddr
0x002aaaab522000
  remote address: LID 0x02, QPN 0x6004a, PSN 0x751bb0, RKey 0x5a002400 VAddr
0x002aaaaaef000
Mtu : 2048
-----
#bytes #iterations  t_min[usec]  t_max[usec]  t_typical
[usec]
-----
```

← Command initiated on Machine 1

No results initially as command on Machine 2 has not been started

```
[root@mtilab31 ~]# ib_write_lat 10.2.2.32
-----
RDMA_Write Latency Test
Inline data is used up to 400 bytes message
Connection type : RC
  local address: LID 0x02 QPN 0x8004a PSN 0xab8b63 RKey 0x5c002400 VAddr
0x0000000060b002
  remote address: LID 0x01 QPN 0xa004a PSN 0x74d635 RKey 0x92002400 VAddr
0x000000026e1002
Mtu : 2048
-----
#bytes #iterations  t_min[usec]  t_max[usec]  t_typical[usec]
  2      1000      1.28      20.85      1.31
-----
```

← Command initiated on machine 2, note addition of IP address

← Results consistent with expected latency on Connect-X cards running SDR

Step 7 (Continued)

- Gather detailed information on all HCA's and Switches in the cluster

5. Run **ibnetdiscover** from any machine on the cluster

SCREEN DUMP from **ibnetdiscover**

```
[root@mtilab31 ~]# ibnetdiscover
#
# Topology file: generated on Tue Aug 26 05:29:42 2008
#
# Max of 2 hops discovered
# Initiated from node 0002c9030000057c port 0002c9030000057d

vendid=0x2c9
devid=0xa87c

switchguid=0xb8cfff0020c1(b8cfff0020c1)
Switch 8 "S-000b8cfff0020c1" # "MT43132 Mellanox Technologies" base port 0 lid 9 lmc 0
[8] "H-0002c90200251480"[1](2c90200251481) # "HCA-1" lid 4 4xSDR
[4] "H-0002c902002170d8"[1](2c902002170d9) # "HCA-1" lid 5 4xSDR
[2] "H-0002c9020025148c"[1](2c9020025148d) # "mtilab33 HCA-1" lid 3 4xSDR
[1] "H-00000002c9002138"[1](2c9002139) # "mtilab32 HCA-1" lid 1 4xSDR
[3] "H-0002c9030000057c"[1](2c9030000057d) # "mtilab31 HCA-1" lid 2 1xSDR

vendid=0x2c9
devid=0x6282
sysimgguid=0x2c90200251483
caguid=0x2c90200251480
Ca 2 "H-0002c90200251480" # "HCA-1"
[1](2c90200251481) "S-000b8cfff0020c1"[8] # lid 4 lmc 0 "MT43132 Mellanox Technologies"
lid 9 4xSDR

vendid=0x2c9
devid=0x634a
sysimgguid=0x2c902002170db
caguid=0x2c902002170d8
Ca 2 "H-0002c902002170d8" # "HCA-1"
[1](2c902002170d9) "S-000b8cfff0020c1"[4] # lid 5 lmc 0 "MT43132 Mellanox Technologies"
lid 9 4xSDR

vendid=0x2c9
devid=0x6282
sysimgguid=0x2c9020025148f
caguid=0x2c9020025148c
Ca 2 "H-0002c9020025148c" # "mtilab33 HCA-1"
[1](2c9020025148d) "S-000b8cfff0020c1"[2] # lid 3 lmc 0 "MT43132 Mellanox Technologies"
lid 9 4xSDR

vendid=0x2c9
devid=0x6340
sysimgguid=0x2c900213b
caguid=0x2c9002138
Ca 2 "H-00000002c9002138" # "mtilab32 HCA-1"
[1](2c9002139) "S-000b8cfff0020c1"[1] # lid 1 lmc 0 "MT43132 Mellanox Technologies" lid 9
4xSDR

vendid=0x2c9
devid=0x634a
sysimgguid=0x2c9030000057f
caguid=0x2c9030000057c
Ca 2 "H-0002c9030000057c" # "mtilab31 HCA-1"
[1](2c9030000057d) "S-000b8cfff0020c1"[3] # lid 2 lmc 0 "MT43132 Mellanox Technologies"
lid 9 4xSDR
```

Step 7 (Continued)

- Verify no Symbol errors as this can affect cluster stability

6. Run **perfquery** from any machine on the cluster

SCREEN DUMP from **perfquery**

```
[root@mtilab31 ~]# perfquery
# Port counters: Lid 2 port 1
PortSelect:.....1
CounterSelect:.....0x0000
SymbolErrors:.....0
LinkRecovers:.....0
LinkDowned:.....0
RcvErrors:.....0
RcvRemotePhysErrors:.....0
RcvSwRelayErrors:.....0
XmtDiscards:.....0
XmtConstraintErrors:.....0
RcvConstraintErrors:.....0
LinkIntegrityErrors:.....0
ExcBufOverrunErrors:.....0
VL15Dropped:.....0
XmtData:.....165841472
RcvData:.....111400
XmtPkts:.....322326
RcvPkts:.....12325
```

```
perfquery -h FOR HELP
perfquery -R TO RESET ALL COUNTERS
```

- Additional Diagnostic commands for gathering cluster information include :
 - **ibswitches**
 - Detailed info on all switches in the cluster
 - **ibhosts**
 - Detailed info on all HCA's in the cluster
- For additional information on Diagnostic tools included in the OFED Stack including the specific switches and options available, consult Chapter 10 of the OFED Linux Users Manual