Tutorial:
InfiniBand clusters with Open Fabrics Software Stack

HPC Advisory Council
Stanford Workshop
December 6-7th, 2011

Todd Wilde - Director of Technical Computing and HPC
Basics of the OFED InfiniBand Stack
Open Fabrics Enterprise Distribution (OFED) is a complete SW stack for RDMA capable devices.

- Contains low level drivers, core, Upper Layer Protocols (IPoIB, MPI), Cluster Tools and documentation.

- Available on OpenFabrics.org.

Mellanox version (MLNX_OFED) includes:
- Pre-built RPMS for popular OS’es
- Documentation and complete User Manual
- Includes new features not released into OFED yet.
OFED Installation

- **Pre-built RPM install.**
  1. mount -o rw,loop MLNX_OFED_LINUX-1.4-rhel5.3.iso /mnt
  2. cd /mnt
  3. ./mlnxofedinstall

- **Building RPMs for un-supported kernels.**
  1. mount -o rw,loop MLNX_OFED_LINUX-1.4-rhel5.3.iso /mnt
  2. cd /mnt/src
  3. cp OFED-1.4.tgz /root (this is the original OFED distribution tarball)
  4. tar zxvf OFED-1.4.tgz
  5. cd OFED-1.4
  6. copy ofed.conf to OFED-1.4 directory
  7. ./install.pl -c ofed.conf
Loading and Unloading the IB stack

- /etc/infiniband/openib.conf controls boot time configuration and other options
  
  # Start HCA driver upon boot
  ONBOOT=yes

  # Load IPoIB
  IPOIB_LOAD=yes

- Optionally manually start and stop services
  - /etc/init.d/openibd start|stop|restart|status
IPoIB in a Nut Shell

- Encapsulation of IP packets over IB

- Uses IB as “layer two” for IP
  - Supports both UD service (up to 2KB MTU) and RC service (connected mode, up to 64KB MTU)

- IPv4, IPv6, ARP and DHCP support

- Multicast support

- VLANs support

- Benefits:
  - Transparent to legacy applications
  - Allows leveraging of existing management infrastructure
IPoIB Configuration

- Requires assigning an IP address and a subnet mask to each HCA port (like any other network adapter)

- The first port on the first HCA in the host is called interface ib0, the second port is called ib1, and so on

- Configuration can be based on DHCP or on a static configuration
  - Modify `/etc/sysconfig/network-scripts/ifcfg-ib0`:
    ```
    DEVICE=ib0
    BOOTPROTO=static
    IPADDR=10.10.0.1
    NETMASK=255.255.255.0
    NETWORK=10.10.0.0
    BROADCAST=10.10.0.255
    ONBOOT=yes
    ```
  - `ifconfig ib0 10.10.0.1 up`
OpenSM in a nutshell

- The Subnet Manager (SM) is mandatory for setting up port ID, links and routes
- OpenSM is an Infiniband compliant subnet manager included with OFED
- Ability to run several instance of osm on the cluster in a Master/Slave(s) configuration for redundancy.
- Partitions P-key (similar to VLANs) support
- QoS support
- Enhanced routing algorithms:
  - Min-hop, up-down, fat-tree, LASH, DOR, Torus2QOS
Running opensm

**Command line**
- Default (no parameters)
  - Scans and initializes the IB fabric and will occasionally sweep for changes
- opensm –h for usage flags
  - e.g. to start with up-down routing: opensm --routing_engine updn
- Run is logged to two files
  - /var/log/messages – registers only major events
  - /var/log/opensm.log – detailed report

**Start on boot/daemon**
- /etc/init.d/opensmd start|stop|restart|status
- /etc/opensm/opensm.conf for default parameters
  ```
  # ONBOOT
  # To start OpenSM automatically set ONBOOT=yes
  ONBOOT=yes
  ```

**SM detection**
- /etc/init.d/opensmd status
  - Shows opensm runtime status on a machine
- sminfo
  - Shows master and standby subnets running on the cluster
Running Benchmarks
InfiniBand benchmarks

- Bandwidth and Latency performance tests
  - /usr/bin/ib_write_bw
  - /usr/bin/ib_write_lat
  - /usr/bin/ib_read_bw
  - /usr/bin/ib_read_lat
  - /usr/bin/ib_send_bw
  - /usr/bin/ib_send_lat

- Usage
  - Server: <test name> <options>
  - Client: <test name> <options> <server IP address>

---

Note: Same options must be passed to both server and client. Use –h for all options.
InfiniBand benchmarks

[root@lisbon001 ~]# ib_send_bw

[root@lisbon002 ~]# ib_send_bw lisbon001

Send BW Test
Number of qps : 1
Connection type : RC
RX depth : 600
CQ Moderation : 50
Link type : IB
Mtu : 2048

Inline data is used up to 0 bytes message
local address: LID 0x5c QPN 0x68004a PSN 0x821f36
remote address: LID 0x5d QPN 0x64004a PSN 0xce92a9

#bytes  #iterations  BW peak[MB/sec]  BW average[MB/sec]
65536  1000   2719.89   2719.47

------------------------------------------------------------------
MPI benchmarks

- **Prerequisites for Running MPI:**
  - The `mpirun_rsh` launcher program requires automatic login (i.e., password-less) onto the remote machines.
  - Must also have an `/etc/hosts` file to specify the IP addresses of all machines that MPI jobs will run on.
  - Make sure there is no loopback node specified (i.e. 127.0.0.1) in the `/etc/hosts` file or jobs may not launch properly.
  - Details on this procedure can be found in Mellanox OFED User’s manual

- **Basic format (mvapich):**
  - `mpirun_rsh --np procs node1 node2 node3 BINARY`

- **Other flags:**
  - `-show`: show only
  - `-paramfile`: environment variables
  - `-hostfile`: list of host
  - `-ENV=VAL` (i.e. `VIADEV_RENDEZVOUS_THRESHOLD=8000`)
MPI benchmarks

```bash
[root@lisbon001 ~]# mpirun_rsh -np 2 lisbon001 lisbon002
/usr/mpi/gcc/mvapich-1.2.0/tests/osu_benchmarks-3.1.1/osu_latency
# OSU MPI Latency Test v3.1.1

<table>
<thead>
<tr>
<th>Size</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.37</td>
</tr>
<tr>
<td>1</td>
<td>1.36</td>
</tr>
<tr>
<td>2</td>
<td>1.37</td>
</tr>
<tr>
<td>4</td>
<td>1.37</td>
</tr>
<tr>
<td>8</td>
<td>1.38</td>
</tr>
<tr>
<td>16</td>
<td>1.38</td>
</tr>
<tr>
<td>32</td>
<td>1.45</td>
</tr>
<tr>
<td>64</td>
<td>1.55</td>
</tr>
<tr>
<td>128</td>
<td>2.43</td>
</tr>
<tr>
<td>256</td>
<td>2.56</td>
</tr>
<tr>
<td>512</td>
<td>2.86</td>
</tr>
<tr>
<td>1024</td>
<td>3.47</td>
</tr>
<tr>
<td>2048</td>
<td>4.75</td>
</tr>
<tr>
<td>4096</td>
<td>6.03</td>
</tr>
<tr>
<td>8192</td>
<td>8.76</td>
</tr>
<tr>
<td>16384</td>
<td>13.35</td>
</tr>
<tr>
<td>32768</td>
<td>18.46</td>
</tr>
<tr>
<td>65536</td>
<td>30.28</td>
</tr>
<tr>
<td>131072</td>
<td>52.84</td>
</tr>
<tr>
<td>262144</td>
<td>99.88</td>
</tr>
<tr>
<td>524288</td>
<td>191.46</td>
</tr>
<tr>
<td>1048576</td>
<td>375.02</td>
</tr>
<tr>
<td>2097152</td>
<td>748.70</td>
</tr>
<tr>
<td>4194304</td>
<td>1481.48</td>
</tr>
</tbody>
</table>
```
OFED InfiniBand Diagnostic Tools
The world of *ib-diags*

### ib-diags: Command Line Tools

- `ibstat`
- `ibstatus`
- `ibaddr`
- `ibroute`
- `sminfo`
- `smpdump`
- `smpquery`
- `perfquery`
- `ibcheckport`
- `ibchecknode`
- `ibcheckerrs`
- `ibportstate`
- `ibcheckportwidth`
- `ibcheckportstate`
- `ibsysstat`
- `ibtracert`
- `ibping`
- `ibnetdiscover`
- `ibdiscover.pl`
- `ibchecknet`
- `ibnetdiscover`
- `ibswitches`
- `ibhosts`
- `ibnodes`
- `ibcheckwidth`
- `ibcheckstate`
- `ibcheckerrors`
- `ibclearerrors`
- `ibclearcounters`
- `saquery`
The world of *ib-diags*

*ib-diags*: Command Line Tools

- ibstat
- ibstatus
- ibaddr
- ibroute
- sminfo
- smpdump
- smpquery
- perfquery
- ibcheckport
- ibchecknode
- ibcheckerrs
- **ibportstate**
- ibcheckportwidth
- ibcheckportstate

- ibsysstat
- ibtracert
- ibping

- ibnetdiscover
- ibdiscover.pl
- ibchecknet
- ibnetdiscover
- ibswitches
- ibhosts
- ibnodes
- ibcheckwidth
- ibcheckstate
- ibcheckerrors
- ibclearerrors
- ibclearcounters
- saquery
The world of *ib-diags*

### Single Node Scope

- **ibstat** - show host adapters status
- **ibstatus** - similar to ibstat but implemented as a script
- **ibaddr** - shows the lid range and default GID of the target (default is the local port)
- **ibroute** - display unicast and multicast forwarding tables of switches
- **sminfo** - query the SMInfo attribute on a node
- **smpdump** - simple solicited SMP query tool. Output is hex dump
- **smpquery** - formatted SMP query tool
- **perfquery** - dump (and optionally clear) performance/error counters of the destination port
- **ibcheckport** - perform some basic tests on the specified port
- **ibchecknode** - perform some basic tests on the specified node
- **ibcheckerrs** - check if the error counters of the port/node have passed some predefined thresholds
- **ibportstate** - get the logical and physical port state of an IB port or enable/disable port
- **ibcheckportwidth** - perform 1x port width check on specified port
- **ibcheckportstate** - perform port state (and physical port state) check on specified port

### Node based tools can be run on any machine with OFED stack installed

- man pages available for all utilities
- `-h` option for online help
The world of *ib-diags*

- **Source/Destination Path Scope**
  - *ibsysstat* - obtain basic information for node (hostname, cpus, memory) which may be remote
  - *ibtracert* - display unicast or multicast route from source to destination
  - *ibping* - ping/pong between IB nodes (currently using vendor MADs)

- **Subnet Scope**
  - *saquery* - issue some SA queries
  - *ibnetdiscover* - scan topology
  - *ibchecknet* - perform port/node/errors check on the subnet.
  - *ibnetdiscover* - topology output
  - *ibswitches* - scan the net or use existing net topology file and list all switches
  - *ibhosts* - scan the net or use existing net topology file and list all hosts
  - *ibnodes* - scan the net or use existing net topology file and list all nodes
  - *ibcheckwidth* - perform port width check on the subnet. Used to find ports with 1x link width.
  - *ibclearerrors* - clear all error counters on subnet
  - *ibclearcounters* - clear all port counters on subnet
  - *ibcheckstate* - perform port state (and physical port state) check on the subnet.
  - *ibcheckerrors* - perform error check on subnet. Find ports above the indicated thresholds
Each Subnet must have a Subnet Manager (SM):
- Topology Discovery
- Forwarding Table Initialization
- Fabric Maintenance

Every entity (CA, SW, Router) must support a Subnet Management Agent (SMA)

SM communicates with SA using Subnet Management Packets (SMPs)

Local IDs (LIDS) are used to identify end ports/nodes and route packets
### ibstatus

- Displays basic information obtained from the local IB driver
- Output includes Firmware version, GUIDS, LID, SMLID, port state, link width active, and port physical state

```bash
> ibstatus
Infiniband device 'mlx4_0' port 1 status:
  default gid: fe80:0000:0000:0000:0000:0000:0007:3896
  base lid: 0x3
  sm lid: 0x3
  state: 4: ACTIVE
  phys state: 5: LinkUp
  rate: 20 Gb/sec (4X DDR)

Infiniband device 'mlx4_0' port 2 status:
  default gid: fe80:0000:0000:0000:0000:0000:0007:3897
  base lid: 0x1
  sm lid: 0x3
  state: 4: ACTIVE
  phys state: 5: LinkUp
  rate: 20 Gb/sec (4X DDR)
```

- **Down** = Physical Link is Down
- **Initialize** = SM has not configured yet
- **Active** = Ready to transfer data
**ibportstate** – port status/control

- **ibportstate**
  - Enables querying the logical (link) and physical port states of an InfiniBand port. It also allows adjusting the link speed that is enabled on any InfiniBand port.
  - If the queried port is a switch port, then the command can also be used to:
    - Disable, enable or reset the port
    - Validate the port’s link width and speed against the peer port

---

```plaintext
> ibportstate 56 3
PortInfo:
# Port info: DR path slid 65535; dlid 65535; 0 port 1
LinkState: Initialize
PhysLinkState: LinkUp
LinkWidthSupported: 1X or 4X
LinkWidthEnabled: 1X or 4X
LinkWidthActive: 4X
LinkSpeedSupported: 2.5 Gbps or 5.0 Gbps
LinkSpeedEnabled: 2.5 Gbps or 5.0 Gbps
LinkSpeedActive: 5.0 Gbps
```
**perfquery**

- Queries InfiniBand ports’ performance and error counters
- It can also reset counters

```plaintext
perfquery  
# Port counters: Lid 6 port 1  
PortSelect:......................1  
CounterSelect:.....................0x1000  
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:..........................55178210  
RcvData:..........................55174680  
XmtPkts:.........................766366  
RcvPkts:.........................766315
```
smpquery – attribute details

- **smpquery**
  - Reports relevant node, port, switch and other interesting info

```plaintext
-> smpquery 5 6
Lid:.............................0
SMLid:..........................0
CapMask:........................0x0
DiagCode:.......................0x0000
MkeyLeasePeriod:...............0
LocalPort:......................10
LinkWidthEnabled:.............1X or 4X
LinkWidthSupported:..........1X or 4X
LinkWidthActive:..............4X
LinkSpeedSupported:.........2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkState:......................Active
PhysLinkState:................LinkUp
LinkDownDefState:............Polling
ProtectBits:....................0
LMC:............................0
LinkSpeedActive:.............10.0 Gbps
LinkSpeedEnabled:..........2.5 Gbps or 5.0 Gbps or 10.0 Gbps
NeighborMTU:..................2048
VLCap:.........................VL0-7
......................................continued
```
ibnetdiscover – cluster topology report

- Reports a complete topology of cluster

- Shows all interconnect connections reporting:
  - Port LIDs
  - Port GUIDs
  - Host names
  - Link Speed

- GUID to switch name file can be used for more readable topology
ibnetdiscover – cluster topology report

ibnetdiscover –node-name-map my_guid_map_file

```
vendid=0x2c9
devid=0xb924
sysimgguid=0xb8cfff004207
switchguid=0xb8cfff004207

Switch 24 SWITCH_1  "MT47396 Infiniscale-III Mellanox Technologies" base port 0 lid 9 lmc 0
[5]  "H-0002c90200230e54"[1](2c90200230e55)  # "mtilab55 HCA-1" lid 22 4xDDR
[6]  "H-0002c902002312a8"[1](2c902002312a9)  # "mtilab47 HCA-1" lid 12 4xDDR
[14] "H-0002c90300000268"[2](2c9030000026a)  # "mtilab40 HCA-1" lid 20 4xDDR
[18] "H-0002c9020021ad78"[1](2c9020021ad79)  # "mtilab54 HCA-1" lid 21 4xDDR

devid=0x6282
sysimgguid=0x2c902002312ab
caguid=0x2c902002312a8
Ca 2 "H-0002c902002312a8"  # "mtilab47 HCA-1"
[1](2c902002312a9)  "S-000b8cfff004207"[6]  # lid 12 lmc 0 "MT47396 Infiniscale-III Mellanox Technologies" lid 9 4xDDR

vendid=0x2c9
devid=0x6274
sysimgguid=0x2c90200230e57
caguid=0x2c90200230e54
Ca 1 "H-0002c90200230e54"  # "mtilab55 HCA-1"
[1](2c90200230e55)  "S-000b8cfff004207"[5]  # lid 22 lmc 0 "MT47396 Infiniscale-III Mellanox Technologies" lid 9 4xDDR
```
ibutils (ibdiagnet/path)
Integrated Cluster Utilities
**Ibutils – in a nutshell**

- **ibdiagnet**
  - Examine all the paths in the network
    - Look for cross paths issues
    - Network balancing
    - Covers all L2 issues on links

- **ibdiagpath**
  - Source to Destination path based analysis
    - Cover all L2 issues on the path
    - Include extensive link level analysis
ibdiagnet – functionality

- **Topology**
  - Info: dump topology in “topo”, “lst” and “ibnetdiscover” formats
  - Info: Optionally report on cable information (i.e. vendor, cable length, part number)
  - Error: duplicated GUIDs
  - Error: connectivity mismatch to reference topology
  - Warn: link speed/width change from reference topology
  - Error: optional report on any port below given speed/width

- **SM**
  - Info: all active SMs their status and priority
  - Error: missing or multiple masters
  - Error: Illegal LID: 0, duplicated, not meeting LMC
  - Error: invalid link parameters: OpVLs, MTU
  - Error: link width/speed not matching maximal supported

- **Error Counters**
  - Info: a full dump of all IB port counters of the entire subnet
  - Error: error counters over some limit (user controlled)
  - Error: error counters increasing during the run
Routing
• Info: histogram of hops from CA to CA
• Info: histogram of number of CA to CA paths on every port
• Info: multicast groups and their members (include sender only)
• Error: no unicast route between every CA to every other CA
• Error: (on request) no unicast route between every CA/SW to every other CA/SW
• Error: credit loops found (optionally include multicast)
• Error: multicast routing loops, disconnects, garbage

Partitions
• Info: All partitions ports and membership status
• Error: Mismatching host partitions and attached switches ingress port tables

IPoIB
• Info: available broadcast domains and their parameters and member end-points
• Warn: sub-optimal domain parameters (rate to small, rate not met by some nodes)

Bit Error Check
• Error: given some threshold and time between samples
QoS
- Info: Admissible SL’s on the path (including the details where they block etc)
- Info: PathRecord for every SL (optionally limit by given ServiceID, DSCP and SL)
- Error: no common SL to be used
- Error: no PathRecord for given ServiceID, DSCP and SL

Cable Reports
- Reports vendor information, part number, cable length, etc
HW failures – reporting faults using topology comparison

- **Case 1: remove 2 cables**
  - SL2-1 P10 to SL1-5 P14
  - SL2-6 P19 to SL1-10 P23

- **Case 2: remove hosts**
  - H-49, H-12

- **Case 3: remove a switch, or a FRU within a switch system**

```
Case 1:
ibdiag -t `pwd`/network.top
-I-Topology matching results
-I- Topology matching results
   Missing cable connecting:SL2-1/P10 to:SL1-5/P14
   Missing cable connecting:SL2-6/P19 to:SL1-10/P23

Case 2:
ibdiag -t `pwd`/network.top
-I-Topology matching results
-I- Topology matching results
   Missing System:H-12(MT23108)
       Should be connected by cable from port: P1(H-12/U1/P1)
       to:SL1-1/P12(SL1-1/U1/P12)

   Missing System:H-49(MT23108)
       Should be connected by cable from port: P1(H-49/U1/P1)
       to:SL1-5/P1(SL1-5/U1/P1)

Case 3:
ibdiag -t `pwd`/network.top
-I-Topology matching results
-I- Topology matching results
   Missing System Board:SL1-1/leaf3
```
Helpful Hint – use ibdiaginet to write out a topology file

- Writing out the topology
  - Use –wt network.topo to generate the reference topology
  - Host names are already correct…
  - For switches
    - Do some automatic naming modification (rename all switches to SW<index>):
      - grep S000 network.topo |\
        sed 's/.*/\1\^\*\(S000[^\*]\)*.*/\2/' | sort -u |\
        awk '{printf("s/%s/\SW%d/g\n",$1,i++)}' > name_switches.sed
      - sed -f name_switches.sed network.topo > named.topo
  - Or manually edit for setting some names
Cable/connector faults and test

- **Link Faults**
  - Bad cables need to be found in cluster bring-up
  - Error counters provide on every IB port report these issues

- **Reporting Link Faults across Network**
  - Error: When any port counter change rate > than threshold
  - Report: Entire set of counters for each port on the subnet

```
ibdiagnet -t `pwd`/network.topo
-I----------------------------------------------------------------------
-I- PM Counters Info
-I----------------------------------------------------------------------
-W- "H-37/P1" lid=0x0087 guid=0x0002c9000000ee dev=23108
    Performance Monitor counter : Value
    port_rcv_errors              : 0x307 (Increase by 34 during ibdiagnet scan.)
-W- "SL1-2/P11" lid=0x0008 guid=0x0002c9000000207 dev=47396
    Performance Monitor counter : Value
    port_rcv_errors              : 0xd1 (Increase by 5 during ibdiagnet scan.)
-W- "SL1-2/P16" lid=0x0008 guid=0x0002c9000000207 dev=47396
    Performance Monitor counter : Value
    port_rcv_errors              : 0x6c (Increase by 4 during ibdiagnet scan.)
-W- "SL1-4/P1" lid=0x000c guid=0x0002c900000020b dev=47396
    Performance Monitor counter : Value
    port_xmit_discard            : 0x307 (Increase by 34 during ibdiagnet scan.)
```
Cable/connector faults and test

```
ibdiagpath -t `pwd`/network.topo -n H-3
-I- PM Counters Info
-I- PM Counters Info
-W- "SL1-1/U1/P3" lid=0x0002 guid=0x0002c90000000201 dev=47396
    Performance Monitor counter : Value
    port_rcv_errors : 0xcd4 (Increase by 7 during ibdiagpath scan.)
```

```
ibdiagpath -t `pwd`/network.topo -n H-23
-I- PM Counters Info
-I- PM Counters Info
-W- "SL1-2/U1/P11" lid=0x0008 guid=0x0002c900000000207 dev=47396
    Performance Monitor counter : Value
    port_rcv_errors : 0x603 (Increase by 8 during ibdiagpath scan.)
```
Advanced ibutil Topics
Subnet Manager faults

- The Subnet Manager (SM) is mandatory for setting up port ID, links and routes

Subnet Manager Reporting

- One and only one “master” SM
  - Error: When no master or more then one master
  - Report: All master and standby SM ports

- SM is responsible for configuring links
  - Error: When Neighbor MTU is not correctly set by SM
  - Error: If operational VLs does not match the other sides of the link

- Packet routes are configured by the SM
  - Error: When not all nodes are assigned an unique address (LID)
  - Error: If routes from every nodes to every other node are not set
  - Error: If multicast routes for each member of each group are not proper
  - Error: If “credit loops” are caused by the routing
Multiple SM’s - *ibdiagnet* results

**No SM:**
- I- --------------------------------------------------
- I- Bad Fabric SM Info
- I- --------------------------------------------------
- E- Missing master SM in the discover fabric

**The normal case: one master one standby**
- I- --------------------------------------------------
- I- Summary Fabric SM-state-priority
- I- --------------------------------------------------
  SM - master
  "H-1/P1" lid=0x0001 guid=0x0002c9000000002 dev=23108 priority:0
  SM - standby
  The Local Device: "H-2/P1" lid=0x0062 guid=0x0002c900000000a2 dev=23108 priority:0

**Two masters?? (hard to create ...)**
- I- --------------------------------------------------
- I- Bad Fabric SM Info
- I- --------------------------------------------------
- E- Found more then one master SM in the discover fabric
  The Local Device: H-2/P1 priority:0
  H-1/P1 priority:0
ibdiagnet -t `pwd`/network.topo -r
...
-I----------------------------------------------------
-I- Fabric qualities report
-I----------------------------------------------------
-I- Verifying all CA to CA paths ...
-E- Unassigned LFT for lid:70 Dead end at:H-122/U1
-E- Fail to find a path from:H-1/U1/1 to:H-24/U1/1
...
-E- Found 1380 missing paths out of:13340 paths

# Multicast disconnect and unneeded entries
-I- Scanning all multicast groups for loops and connectivity...
-I- Multicast Group:0xC000 has:6 switches and:8 FullMember CA ports
-W- Switch: S0002c900000000004/U1 has unconnected MFT entries for MLID:0xC000
-W- Switch: S0002c900000000005/U1 has unconnected MFT entries for MLID:0xC000
-E- Found 2 connection groups for MLID:0xC000
    Group:1 has 4 CAs: H-[9..12]/U1
    Group:1 has 1 SWs: S0002c900000000001/U1
    Group:2 has 4 CAs: H-[13..16]/U1
    Group:2 has 1 SWs: S0002c900000000003/U1
Credit Loops? What are these?

- “loss-less fabric” = “link level flow control” = packet not sent if there is no buffer for it
- If traffic to DST-1 waits on traffic for DST-2 which in turn depends on traffic to DST-3 which depends on DST-1 we have a dependency loop and the fabric deadlocks
SM/SA Test Cases

- Credit Loops in real world
Credit Loop issues - *ibdiagnet* results

```
ibdiagnet -t `pwd`/network.topo -r
...
-I----------------------------------------------------------------
-I- Checking credit loops
-I----------------------------------------------------------------
-I-
-I- Analyzing Fabric for Credit Loops 1 SLs, 1 VLs used.
   Found credit loop on: SW_L2_1/P3 VL: 0
   - BT credit loop through: SW_L1_8/P2 VL: 0
   - BT credit loop through: SW_L2_8/P3 VL: 0
   - BT credit loop through: SW_L1_4/P2 VL: 0
   - BT credit loop through: SW_L2_4/P3 VL: 0
   - BT credit loop through: SW_L1_7/P2 VL: 0
   - BT credit loop through: SW_L2_7/P3 VL: 0
   - BT credit loop through: SW_L1_3/P2 VL: 0
   - BT credit loop through: SW_L2_3/P3 VL: 0
   - BT credit loop through: SW_L1_6/P2 VL: 0
   - BT credit loop through: SW_L2_6/P3 VL: 0
   - BT credit loop through: SW_L1_2/P2 VL: 0
   - BT credit loop through: SW_L2_2/P3 VL: 0
   - BT credit loop through: SW_L1_5/P2 VL: 0
   - BT credit loop through: SW_L2_5/P3 VL: 0
   - BT credit loop through: SW_L1_1/P2 VL: 0
   - BT credit loop through: SW_L2_1/P3 VL: 0
   - BT credit loop through: H-1/P1 VL: 0
-E- credit loops in routing
-E- Total Credit Loop Check Errors:1
```
Partition Configuration Issues

- Partitions are similar to VLAN IDs, but enforced on hosts and switch ports

- Network
  - Warning: Partition enforcement by leaf switches mismatch hosts
  - Report: Node groups – which nodes can communicate

- Path
  - Error: No common partition for the path
  - Error: Mismatch between leaf switch and host partitions
  - Report: Which partitions can be used for the path
  - Verbose: On each port (if enforcing) show list of PKeys
Partitions Test Case

- Two partitions with some common nodes.

---

**Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)**

-W- Missing PKey:0x8001 on remote switch of node:"H-95/P1" lid=0x0089 guid=0x0002c900000001ee dev=23108

-I- PKey:0x0001 Hosts:87 full:87 limited:0
-I- PKey:0x0002 Hosts:84 full:84 limited:0
-I- PKey:0x7fff Hosts:128 full:1 limited:127

---
IPoIB Subnets

- Each IPoIB subnet is attached to a partition and a broadcast group
- Network
  - Warn: Not all members of the subnet can join the group
  - Warn: All members support higher than setup rate
- Path
  - Error: When no common IPoIB subnet
  - Report: Common IPoIB subnets and their parameters

---

-I- IPoIB Subnets Check
-I- IPoIB Subnets Check
-I- Subnet: IPv4 PKey:0x0001 QKey:0x00000000b1b MTU:2048Byte rate:20Gbps SL:0x00
-W- Port "H-40/P1" lid=0x0090 guid=0x0002c900000000fe dev=23108 can not join due to rate:5Gbps < group:20Gbps
-I- Subnet: IPv4 PKey:0x0002 QKey:0x00000000b1b MTU:2048Byte rate:10Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0003 QKey:0x00000000b1b MTU:2048Byte rate:10Gbps SL:0x00
-W- Suboptimal rate for group. Lowest member rate:20Gbps > group-rate:10Gbps
...
Questions?
ibdiagnet run – a good case

- Parsing Subnet file:/tmp/ibmgtsim.7021/ibdiagnet.lst
- Defined 145/145 systems/nodes
- Bad Guides/LIDs Info
- skip option set. no report will be issued
- Links With Logical State = INIT
- No bad Links (with logical state = INIT) were found
- General Device Info
- PM Counters Info
- No illegal PM counters values were found
- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
- PKey:0x0001 Hosts:128 full:128 limited:0
- PKey:0x0002 Hosts:128 full:128 limited:0
- PKey:0x0003 Hosts:128 full:128 limited:0
- PKey:0x7fff Hosts:128 full:1 limited:127
- IPoIB Subnets Check
- Subnet: IPv4 PKey:0x0001 QKey:0x00000b1b MTU:2048 Byte rate:20Gbps SL:0x00
- Subnet: IPv4 PKey:0x0002 QKey:0x00000b1b MTU:2048 Byte rate:20Gbps SL:0x00
- Subnet: IPv4 PKey:0x0003 QKey:0x00000b1b MTU:2048 Byte rate:20Gbps SL:0x00
ibdiagnet run – a good case

-I- Bad Links Info
-I- No bad link were found
-I- Summary Fabric SM-state-priority
-I- Summary Fabric SM-state-priority

SM - master
The Local Device: H-1/P1 lid=0x0001 guid=0x0002c900000002 dev=23108
priority:0

-I- Fabric qualities report
-I- Fabric qualities report

-I- Parsing FDBs file:/tmp/ibmgtosim.7021/ibdiagnet.fdb
-I- Defined 2465 fdb entries for:17 switches
-I- Parsing Multicast FDBs file:/tmp/ibmgtosim.7021/ibdiagnet.mcfdbs
-I- Defined 450 Multicast Fdb entries for:17 switches
-I-
-I- Verifying all CA to CA paths ...

---------------------------- CA to CA : LFT ROUTE HOP HISTOGRAM -------------------
The number of CA pairs that are in each number of hops distance.
This data is based on the result of the routing algorithm.

HOPS NUM-CA-CA-PAIRS
2 1364
4 14892
----------------------------
ibdiagnet run – a good case

---------- LFT CA to CA : SWITCH OUT PORT - NUM DLIDS HISTOGRAM ----------
Number of actual Destination LIDs going through each switch out port considering all the CA to CA paths. Ports driving CAs are ignored (as they must have = Nca - 1). If the fabric is routed correctly the histogram should be narrow for all ports on same level of the tree.
A detailed report is provided in /tmp/ibdmchk.sw_out_port_num_dlids.

NUM-DLIDS  NUM-SWITCH-PORTS
1   20
2   84
3   21
4   2
5   1
9   28
10  72
11  28

-I- Scanned:16256 CA to CA paths

-I- Scanning all multicast groups for loops and connectivity...
-I- Multicast Group:0xC000 has:12 switches and:128 HCAs
-I- Multicast Group:0xC001 has:12 switches and:128 HCAs
-I- Multicast Group:0xC002 has:12 switches and:128 HCAs
-I- Checking credit loops
-I- Analyzing Fabric for Credit Loops 1 SLs, 1 VLs used.
-I- no credit loops found
ibdiagnet run – a good case

-I- mgid-mlid-HCAs table

<table>
<thead>
<tr>
<th>mgid</th>
<th>mlid</th>
<th>PKey</th>
<th>QKey</th>
<th>MTU</th>
<th>rate</th>
<th>HCAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xff12401b80010000:0x00000000ffffffff</td>
<td>0xc000</td>
<td>0x8001</td>
<td>0x00000b1b</td>
<td>=2048</td>
<td>=20Gbps</td>
<td>128</td>
</tr>
<tr>
<td>0xff12401b80020000:0x00000000ffffffff</td>
<td>0xc001</td>
<td>0x8002</td>
<td>0x00000b1b</td>
<td>=2048</td>
<td>=20Gbps</td>
<td>128</td>
</tr>
<tr>
<td>0xff12401b80030000:0x00000000ffffffff</td>
<td>0xc002</td>
<td>0x8003</td>
<td>0x00000b1b</td>
<td>=2048</td>
<td>=20Gbps</td>
<td>128</td>
</tr>
</tbody>
</table>

-I- Stages Status Report:

<table>
<thead>
<tr>
<th>STAGE</th>
<th>Errors</th>
<th>Warnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad GUIDs/LIDs Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Link State Active Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>General Devices Info Report</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Performance Counters Report</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Partitions Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IPoIB Subnets Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subnet Manager Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fabric Qualities Report</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Credit Loops Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multicast Groups Report</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Please see /tmp/ibmgtsim.7021/ibdiagnet.log for complete log
ibdiagpath -f network.topo -l 128
-I- Parsing topology definition:/local/ez/OSM_REGRESSION/SRC/ibutils/ibdiag/demo/network.topo
-I- Defined 145/145 systems/nodes
-I- Traversing the path from local to destination
-I- Traversing the path from local to destination
-I- From: "H-1/U1/P1"  lid=0x0001 guid=0x0002c90000000002 dev=23108
-I- To: "SL1-1/U1/P1"  lid=0x0002 guid=0x0002c90000000201 dev=47396

-I- From: "SL1-1/U1/P21"  lid=0x0002 guid=0x0002c90000000201 dev=47396
-I- To: "SL2-5/U1/P1"  lid=0x0020 guid=0x0002c9000000021f dev=47396

-I- From: "SL2-5/U1/P5"  lid=0x0020 guid=0x0002c9000000021f dev=47396
-I- To: "SL1-3/U1/P21"  lid=0x0009 guid=0x0002c90000000209 dev=47396

-I- From: "SL1-3/U1/P9"  lid=0x0009 guid=0x0002c90000000209 dev=47396
-I- To: "H-33/U1/P1"  lid=0x0080 guid=0x0002c900000000de dev=23108
-I- PM Counters Info
-I- No illegal PM counters values were found
-I- Path Partitions Report
-I- Source "H-1/U1/P1"  lid=0x0001 guid=0x0002c90000000002 dev=23108 Port 1
  PKeys:0xffff 0x8001 0x8002 0x8003
-I- Destination "H-33/U1"  lid=0x0080 guid=0x0002c9000000000de dev=23108
  PKeys:0x7fff 0x8001 0x8002 0x8003
-I- Path shared PKeys: 0x8001 0xffff 0x8002 0x8003
Ibdiagpath run – a good case

-I------------------------------------------------
-I- IPoIB Path Check
-I------------------------------------------------
-I- Subnet: IPv4 PKey:0x0001 QKey:0x00000b1b MTU:2048 Byte rate:20Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0002 QKey:0x00000b1b MTU:2048 Byte rate:20Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0003 QKey:0x00000b1b MTU:2048 Byte rate:20Gbps SL:0x00
-I------------------------------------------------
-I- QoS on Path Check
-I------------------------------------------------
-I- The following SLs can be used:0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
-I------------------------------------------------
-I- Stages Status Report:

<table>
<thead>
<tr>
<th>STAGE</th>
<th>Errors</th>
<th>Warnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFT Traversal: local to destination</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Performance Counters Report</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Path Partitions Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Path IPoIB Check</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>QoS on Path Check</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Please see /tmp/ibdiagpath.log for complete log