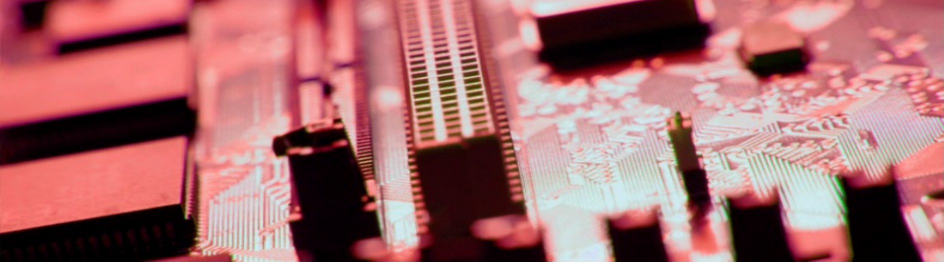

Linux Clusters Institute: Intermediate Networking Practicum

Vernard Martin

HPC Systems Administrator

Oak Ridge National Laboratory

Hand-On with InfiniBand Setup

- Sorta
  - InfiniBand hardware is expensive and rarely just lying around
  - There are few simulators
    - Ibmgtsim: simulates the Fabric as defined by a given topology.
    - Ibsim: Voltaire IB fabric simulator (uses output of ibnetdiscover)
  - Psuedo-Interfaces
    - Psuedo InfiniBand HCA driver: provides IB functions without real IB HCA & fabric. Trys to simulate IB behavior without the physical attributes (i.e. speed, latency, etc)
      - Linux kernel module
      - Userspace plug-in module for libibverbs
      - IB switch emulator for multi-host mode

Quick and Dirty IB Install

- Always upgrade your HBA and switch firmware
  - VERIFY that they are compatible as well!
  - This means scheduled downtime btw.
- Physically connect the network
  - Nodes to switches
  - Switches to switches in Fat Trees
- Choose a subnet manager (embedded switch vs host based)
- Load the kernel modules

What you need (normally)

- Hardware
  - Host Channel Adapter (card)
  - Switch/Router
- Software
  - Subnet Manager (sometimes provided in the switch/router)
  - libibverbs: core user space library that implements the hardware abstracted verbs protocol
  - rdma: package containing kernel modules plus administartion scripts
  - Ibutils, infiniband-diags: various utils for accessing the health of your IB fabric and testing end to end connectivity
  - Perftest,qperf: performance testing tools spefic to RDMA fabrics

PIB Setup and Use

- RedHat/CentOS 6 supported
- RPMs you will need

- Software
  - rdma (service)
  - Kernel modules
  - Various IB diagnostic utils

Testing your Infiniband Network

Check the network health

- From a host
  - Ibstat
  - Ibhosts
  - Ibswitches
  - Ibping
  - Ib_rdma_lat: test RDMA latency
  - Ib_rdma_bw: test RDMA bandwidth
- From the switches
  - Usually a GUI interface
  - If subnet manager available, will show you that as well.

Troubleshooting IB connection issues

OFED tools are very useful and can be used to debug most issues

Knowing where the data is stored is equally important.

“ls /sys/class/infiniband”: S

Shows what IB hardware modules are in use

“cat /sys/class/infiniband/mlx4_0/ports/1/state”:

ACTIVE: hardware is initialized and found by a subnet manager

INIT: the hardware is initialized but no subnet manager has added the port to the fabric yet

(might need to start a subnet manager such as opensm)

lspci/ibv_devices

Utility for displaying information about all PCI buses in the system and all devices connected to them

“lscpi | grep –I Infiniband” to find all cards on the host.

Lists RDMA capable devices installed in the system

“ibv_devices”

Is it active?

Check the state everewhere: “pdsh –a cat /sys/class/infiniband/mlx4_0/ports/1/state”

Once all of the compute nodes report that port 1 is “ACTIVE”, verify the speed on each port

This is a good first check for a bad cable or connection. Each port should report the same speed. For example, the output for double data rate (DDR) InfiniBand cards will be similar to “20 Gb/sec (4X DDR)”

Ibdiagnet

The main OFED tool for troubleshooting performance and connection problems is ibdiagnet. This tool runs multiple tests, as specified on the command line during the run, to detect errors related to the subnet, bad packets, and bad states. These errors are some of the more common seen during initial setup of Infiniband fabrics.

```
Ibdiagnet –pc –c 1000
```

Example output:

```
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
-W- A few ports of local device are up.
Since port-num was not specified (-p option), port 1 of device 1 will be
used as the local port.
-I- Discovering ... 17 nodes (1 Switches & 16 CA-s) discovered.
```

Ibv_devinfo/ibnetdiscover/ibstat

Ibv_devinfo: Query RDMA devices on the system

ibnetdiscover: Discover the infiniband topology

Ibstat: queries basic status of Infiniband devices installed on system

Ibstatus: Default info for each port on each infiniband HCA

Verifying ports are active

View the port status of your hardware with “ibv_devinfo”

Sample output:

```
hca_id: ehca0
        node_guid: 0002:5500:000c:d300
        sys_image_guid: 0000:0000:0000:0000
        vendor_id: 0x5076
        vendor_part_id: 0
        hw_ver: 0x1000003
        phys_port_cnt: 2
        port: 1
                state: PORT_ACTIVE (4)
                max_mtu: 2048 (4)
                active_mtu: 2048 (4)
                sm_lid: 2
                port_lid: 16
                port_lmc: 0x00
        ……..
```

Verify that RDMA is working

- System A:
  - Note the IP address of the host
  - Run the following command “ib_write_bw”
- System B:
  - Run the following command “ib_write_bw <dotted quad>

ibswitches

Show infiniband switch nodes in topology

Ibhosts

To find other hosts with the IB HCAS, run the following command:

```
ibhosts
```

Look for output similar to the following:

```
Ca: 0x0002c90300010458 ports 2 "elm3b199 HCA-1"
Ca : 0x0002c903000128ec ports 2 "elm3b198 HCA-1"
```

To find the switches in the topology, run the command:

```
Ibswitches
```

Look for output similar to the following:

```
Switch : 0x00066a00d9000625 ports 24 "SilverStorm 9024 DDR \
GUID=0x00066a00d9000625" enhanced port 0 lid 1 lmc 0
```

Ibdatacounts

To view the RDMA packet counts, use the **ibdatacounts** command.

To show port counters on LID 4 Port 2 use: “Ibdatacounts 4 2”

Similar output to this:

```
o XmtData:.........................1039851
o RcvData:..........................1040646
o XmtPkts:..........................14447
o RcvPkts:..........................14461
```

Iblinkinfo.pl/perfquery

Report link info for all links in the fabric

Gives all ports on all switches

Usage: iblinkinfo.pl –R

Performance counters per port for each HCA

Usage: perfquery

Problem Determination

- Subnet manager logs
- System logs
- Switch management / switch logs

Problem Determination

- Switch management / switch logs

Problem Determination

- What is a reasonable error rate?

Problem Determination

- How do you know when something isn’t working?
  - Insertion of a node may show errors: this is reasonable!
- Know what to expect in a good configuration (snapshot)
  - OFED tools to look at results

Problem Determination

- How do you clear counters
  - “ibcheckerrors” generates a TON of errors from a port being down to transmissions error. After troubleshooting, you should clear the error counters to make it easier to debug things next time.

    The output of ‘ibcheckerrors’ can be confusing when you’re trying to determine which physical ports the errors are happening on. A good way to see which ‘lid’ an error happens on is by using the tool ‘ibnetdiscover’ to print out your entire, active IB network layout.

  - “ibclearerrors” does what you expect.

Counters

Counter	Explanation
SymbolErrorCounter	Total number of minor link errors detected on one or more physical lanes.
LinkErrorRecovery-Counter	Total number of times the Port Training state machine has successfully completed the link error recovery process.
LinkDownedCounter	Total number of times the Port Training state machine has failed the link error recovery process and downed the link.
PortRcvErrors	Total number of packets containing an error that were received on the port.
PortRcvRemotePhysicalErrors	Total number of packets marked with the EBP delimiter received on the port.
PortRcvSwitchRelayErrors	Total number of packets received on the port that were discarded because they could not be forwarded by the switch relay.
PortXmitDiscards	Total number of outbound packets discarded by the port because the port is down or congested.

Counters (cont.)

PortXmitConstraintErrors	Total number of packets not transmitted from the port for the following reasons: - FilterRawOutbound is true and packet is raw - PartitionEnforcementOutbound is true and packet fails partition key check or IP version check
PortRcvConstraintErrors	Total number of packets received on the port that are discarded for the following reasons: - FilterRawInbound is true and packet is raw - PartitionEnforcementInbound is true and packet fails partition key check or IP version check
LocalLinkIntegrityErrors	The number of times that the count of local physical errors exceeded the threshold specified by LocalPhyErrors
ExcessiveBufferOverrunErrors	The number of times that OverrunErrors consecutive flow control update periods occurred, each having at least one overrun error
VL15Dropped	Number of incoming VL15 packets dropped due to resource limitations (e.g., lack of buffers) in the port.

Counters (cont.)

PortXmitData	Total number of data octets, divided by 4, transmitted on all VLs from the port. This includes all octets between (and not including) the start of packet delimiter and the VCRC, and may include packets containing errors It excludes all link packets.
PortRcvData	Total number of data octets, divided by 4, received on all VLs at the port. This includes all octets between (and not including) the start of packet delimiter and the VCRC, and may include packets containing errors It excludes all link packets.
PortXmitPkts	Total number of packets transmitted on all VLs from the port. This may include packets with errors and excludes link packets.
PortRcvPkts	Total number of packets, including packets containing errors and excluding link packets, received from all VLs on the port.

Problem Determination

- Understanding expected behavior vs abnormal behavior
  - What is baseline? How do you get a baseline performance?
  - IPoIB tests are common
  - Run throughput tests
    - Start server side IB UD throughput test
      - ib_send_bw -a -c UD -d mlx4_0 -i 2
    - Start client side IB UD throughput test
      - ib_send_bw -a -c UD -d mlx4_0 -i 2 10.10.11.8
  - Run latency test
    - Start server side IB UD latency test
      - ib_send_lat -a -c UD -d mlx4_0 -i 2
    - Start client side IB UD latency test
      - ib_send_lat -a -c UD -d mlx4_0 -i 2

Problem Determination

- Tools to monitor
  - Ohio Supercomputer Center tools
    - IB ping, bandwidth, latency tests
  - Applications may show problems that tools may not

Problem Determination

- Cooperative venture with application designers/3rd party/storage/systems and networking
  - Nothing operates in a vacuum