Introduction to Infiniband

Hussein N. Harake, Performance U! Winter School
Agenda

• Definition of Infiniband
• Features
• Hardware
• Facts
• Layers
• OFED Stack
• OpenSM
• Tools and Utilities
• Topologies
• Infiniband Roadmap
Definition of Infiniband

- Type of communication link that allows data to flow
- Communication Fabric used in the HPC domain
- Open standard interconnect

Developed by Infiniband Trade Association
http://www.infinibandta.org/
Development

- Software Stack is developed by the Openfabric community
- http://www.openfabrics.org/

Open source software stack for HPC network that required bandwidth, scalability and low latency.

OpenFabrics Enterprise Distribution (OFED™)
Infiniband Features

- Scalable HPC Network
- Low Latency ~ 1 microsecond
- Low CPU Overhead
- QoS
- Failover
- Congestion control
Infiniband Hardware

- **Switch**
  - 36 FDR (56Gb/s) ports in a 1U switch
  - IBTA 1.3 and 1.21 compliant
  - SDR/DDR/QDR/FDR10/FDR link speed
  - Congestion control*
  - Adaptive routing*
  - Forward Error Correction (FEC)
  - Port mirroring*
  - FDR10 supports 20% more bandwidth over QDR using the same cables/connections**
  - Up to 8 multiple switch partitions*
  - IPoIB Bridging/Routing *
Infiniband Hardware

- Virtual Protocol Interconnect
- 1µs MPI ping latency
- Up to 56Gb/s InfiniBand or 40 Gigabit Ethernet per port
- Single- and Dual-Port options available
- PCI Express 3.0 (up to 8GT/s)
- CPU offload of transport operations
- Application offload
- GPU communication acceleration
- Precision Clock Synchronization
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- Fibre Channel encapsulation (FCoIB or FCoE)
- Ethernet encapsulation (EoIB)
- RoHS-R6

GT/s gigatransfers per second
Infiniband Hardware

- Greater than 100Gb/s over InfiniBand
- Greater than 130M messages/sec
- 1us MPI ping latency
- PCI Express 3.0 x16
- CPU offload of transport operations
- Application offload
- GPU communication acceleration
- End-to-end internal data protection
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- RoHS-R6
# Infiniband Hardware

<table>
<thead>
<tr>
<th>Ordering Part No.</th>
<th>Speed</th>
<th>Ports</th>
<th>Connectors</th>
<th>ASIC &amp; PCI Dev ID</th>
<th>PCI</th>
<th>Lanes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mellanox ConnectX®-2 VPI Adapter Cards</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHQH198-XTR</td>
<td>QDR and 10GbE</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX®-2 26428</td>
<td>PCIe 2.0</td>
<td>x8</td>
</tr>
<tr>
<td>MHQH29C-XTR</td>
<td>QDR and 10GbE</td>
<td>2</td>
<td>QSFP</td>
<td>ConnectX®-2 26428</td>
<td>PCIe 2.0</td>
<td>x8</td>
</tr>
<tr>
<td>MHRH198-XTR</td>
<td>DDR and 10GbE</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX®-2 25418</td>
<td>PCIe 2.0</td>
<td>x8</td>
</tr>
<tr>
<td>MHRH29C-XTR</td>
<td>DDR and 10GbE</td>
<td>2</td>
<td>QSFP</td>
<td>ConnectX®-2 26418</td>
<td>PCIe 2.0</td>
<td>x8</td>
</tr>
<tr>
<td>MHZ298-XTR</td>
<td>QDR and 10GbE</td>
<td>1 and 1</td>
<td>QSFP and SFP+</td>
<td>ConnectX®-2 26428</td>
<td>PCIe 2.0</td>
<td>x8</td>
</tr>
<tr>
<td><strong>Mellanox ConnectX®-3 VPI Adapter Cards</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCX353A-QCBT</td>
<td>QDR and 10GbE</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX®-3 4099</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td>MCX354A-QCBT</td>
<td>QDR and 10GbE</td>
<td>2</td>
<td>QSFP</td>
<td>ConnectX®-3 4099</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td>MCX353A-TCBT</td>
<td>FDR10 and 10GbE</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX®-3 4099</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td>MCX354A-TCBT</td>
<td>FDR10 and 10GbE</td>
<td>2</td>
<td>QSFP</td>
<td>ConnectX®-3 4099</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td>MCX353A-FCBT</td>
<td>FDR and 40GbE</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX®-3 4099</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td>MCX354A-FCBT</td>
<td>FDR and 40GbE</td>
<td>2</td>
<td>QSFP</td>
<td>ConnectX®-3 4099</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td><strong>Mellanox ConnectX-IB™</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCB193A-FBAT</td>
<td>FDR IB (56Gb/s)</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX-IB</td>
<td>PCIe 2.0</td>
<td>x16</td>
</tr>
<tr>
<td>MCB191A-FCAT</td>
<td>FDR IB (56Gb/s)</td>
<td>1</td>
<td>QSFP</td>
<td>ConnectX-IB</td>
<td>PCIe 3.0</td>
<td>x8</td>
</tr>
<tr>
<td>MCB194A-FCAT</td>
<td>FDR IB (56Gb/s)</td>
<td>2</td>
<td>QSFP</td>
<td>ConnectX-IB</td>
<td>PCIe 3.0</td>
<td>x16</td>
</tr>
</tbody>
</table>
Infiniband Bandwidth Facts

- QDR links use 8b/10b encoding
- Every 10 bits sent, links carry 8 bits of data
- 10Gb link carries 8Gb data
- Applies to SDR, DDR and QDR.

QDR bandwidth is 32Gb/s.

- FDR links use 64b/66b encoding
- Every 66 bits sent carry 64 bits of data.
- From the 14Gb/s links hold 13.64 Gbit/s

FDR bandwidth is 54.54 Gbit/s
Infiniband Layers
Infiniband Layers

Physical Layer:

- How bits are placed on the wire
- Signaling protocol
- Link Rates, link speed and link width
- Cables, copper and optical cables
- Connectors etc..
Infiniband Layers

Link Layer:

Packets, Switching, QoS, Flowcontrol and Data integrity

- Packets operation, format and protocols
- Up to 4K packet size
- Maintain link configuration and route between S. and D. in a subnet.
- Packets forwarded using a 16 bit id (LID) assigned by the subnet manager
- VL (Virtual Lanes) separate logical links which share physical link.
- Up to 15 VLs per link and 1 management lane
- FlowControl manage data flow between links (point to point)
- Data integrity two CRCs per packet, Variant and Invariant CRC to ensure data integrity.
Infiniband Layers

Network Layer:

- Handles routing between different subnets
- Not required within a subnet
- Global Routing Header to route packets by IB subnets
- GUID (Globally Unique ID) is a must per node
- 128bit IPv6 header is used for source and destination packets
Infiniband Layers

Transport Layer:
- QP (Queue Pairs) are used to transport data from one point to another
- Async interface
- Send, receive and complete queue pairs
- IPoIB runs on two different transport mode (UD, CM ..)

Based on the MTU the transport layer:
- Divide the data into packets
- Reassemble based on BTH (Base Transport Header)

Complete Offload and kernel bypass (All mentioned run on the HCA and enables low latency)
Infiniband Layers

Upper Level Protocols:

Categorize ULPs into different categories:

Network
- IPoIB (IP over IB)
- SDP (Socket Direct Protocol)
- WSD (Winsock Direct for windows)

Storage
- NFS over RDMA
- SCSI RDMA Protocol (SRP)
- iSCSI Extensions for RDMA (iSER)

Computing – Clustering
- MPI (Message Passing Interface)
OFED Stack

OFED Stack (OPEN Fabrics Enterprise Distribution)

- A complete Software Stack includes drivers, core, protocols, agents, services, management tools, libraries etc..

- You could obtain the OFED stack from http://www.openfabrics.org/

- Mellanox provides a precompiled version which includes new features, some fixes and firmware update for Mellanox products.

- Intel (Qlogic) download OFED stack from openfabrics or Intel precompiled version.
OpenSM

OpenSM and Subnet Agents are part of the OFED stack

Subnet Management (SM)
• manage the fabric
• discover and configure devices
• Assign local ID (LID) to every port
• provide routing tables

Subnet Agents
• Performance Management
• Connection Manager
• Communication Management
• Device Management
• etc..
Tools and Utilities

Tools and utilities are available to manage, diagnostic and measure performance.

Node or port:
- ibstat
- ibstatus
- ibv_devinfo

Measure Performance
- Ib_write_bw
- Ib_read_bw
- Ib_send_lat
- Ib_read_lat

Diagnostic and query tools on the network
- Ibdiagnet
- Ibhosts
- Iblinkinfo.pl
- Ibswitches
## Infiniband Hardware

### Connect-IB PCI-E 3.0 16X

<table>
<thead>
<tr>
<th>#bytes</th>
<th>#iterations</th>
<th>BW peak[MB/sec]</th>
<th>BW average[MB/sec]</th>
<th>MsgRate[Mpps]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5000</td>
<td>17.52</td>
<td>17.19</td>
<td>9.011615</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>35.53</td>
<td>35.35</td>
<td>9.267570</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>70.33</td>
<td>69.75</td>
<td>9.142398</td>
</tr>
<tr>
<td>16</td>
<td>5000</td>
<td>141.63</td>
<td>140.66</td>
<td>9.218248</td>
</tr>
<tr>
<td>32</td>
<td>5000</td>
<td>283.26</td>
<td>281.03</td>
<td>9.208800</td>
</tr>
<tr>
<td>64</td>
<td>5000</td>
<td>566.47</td>
<td>566.77</td>
<td>9.285990</td>
</tr>
<tr>
<td>128</td>
<td>5000</td>
<td>1133.03</td>
<td>1123.82</td>
<td>9.206333</td>
</tr>
<tr>
<td>256</td>
<td>5000</td>
<td>2242.94</td>
<td>2222.42</td>
<td>9.103021</td>
</tr>
<tr>
<td>512</td>
<td>5000</td>
<td>4455.57</td>
<td>4354.05</td>
<td>8.917087</td>
</tr>
<tr>
<td>1024</td>
<td>5000</td>
<td>7803.83</td>
<td>7653.54</td>
<td>7.837225</td>
</tr>
<tr>
<td>2048</td>
<td>5000</td>
<td>11829.23</td>
<td>11475.45</td>
<td>5.875430</td>
</tr>
<tr>
<td>4096</td>
<td>5000</td>
<td>11921.79</td>
<td>11804.99</td>
<td>3.022076</td>
</tr>
<tr>
<td>8192</td>
<td>5000</td>
<td>11969.12</td>
<td>11937.09</td>
<td>1.527948</td>
</tr>
<tr>
<td>16384</td>
<td>5000</td>
<td>11992.93</td>
<td>11991.61</td>
<td>0.767463</td>
</tr>
<tr>
<td>32768</td>
<td>5000</td>
<td>12030.54</td>
<td>12029.63</td>
<td>0.384948</td>
</tr>
<tr>
<td>65536</td>
<td>5000</td>
<td>12049.43</td>
<td>12048.97</td>
<td>0.192783</td>
</tr>
<tr>
<td>131072</td>
<td>5000</td>
<td>12066.22</td>
<td>12065.84</td>
<td>0.096527</td>
</tr>
<tr>
<td>262144</td>
<td>5000</td>
<td>12066.65</td>
<td>12062.75</td>
<td>0.048251</td>
</tr>
<tr>
<td>524288</td>
<td>5000</td>
<td>12068.60</td>
<td>12068.50</td>
<td>0.024137</td>
</tr>
<tr>
<td>1048576</td>
<td>5000</td>
<td>12069.51</td>
<td>12068.69</td>
<td>0.012069</td>
</tr>
<tr>
<td>2097152</td>
<td>5000</td>
<td>12067.38</td>
<td>12043.58</td>
<td>0.006022</td>
</tr>
<tr>
<td>4194304</td>
<td>5000</td>
<td>11911.23</td>
<td>11735.99</td>
<td>0.002934</td>
</tr>
</tbody>
</table>
### Manage Infiniband

```bash
[root@greina0 ~]# iblinkinfo.pl
Switch 0x0002c903006a5380 MFP:ib36-f4:5x50XX/01:
  1[ ] 14.0625 Gbps Active/ LinkUp)==> 21[ ] "greina03 HCA-1" ( )
  2[ ] 14.0625 Gbps Active/ LinkUp)==> 17[ ] "greina04 HCA-1" ( )
  3[ ] 14.0625 Gbps Active/ LinkUp)==> 15[ ] "greina05 HCA-1" ( )
  4[ ] 14.0625 Gbps Active/ LinkUp)==> 20[ ] "greina02 HCA-1" ( )
  5[ ] 14.0625 Gbps Active/ LinkUp)==> 19[ ] "greina01 HCA-1" ( )
  6[ ] 14.0625 Gbps Active/ LinkUp)==> 22[ ] "greina06 HCA-1" ( )
  7[ ] 14.0625 Gbps Active/ LinkUp)==> 25[ ] "greina08 HCA-1" ( )
  8[ ] 14.0625 Gbps Active/ LinkUp)==> 26[ ] "greina07 HCA-1" ( )
  9[ ] 14.0625 Gbps Active/ LinkUp)==> [ ] "greina09 HCA-1" ( )
 10[ ] 14.0625 Gbps Active/ LinkUp)==> 16[ ] "greina09 HCA-1" ( )
 11[ ] 14.0625 Gbps Active/ LinkUp)==> 24[ ] "greina11 HCA-1" ( )
 12[ ] 14.0625 Gbps Active/ LinkUp)==> 23[ ] "greina10 HCA-1" ( )
 13[ ] 14.0625 Gbps Active/ LinkUp)==> 3[ ] "greina06 HCA-1" ( )
 14[ ] 14.0625 Gbps Active/ LinkUp)==> 34[ ] "greina14 HCA-1" ( )
 15[ ] 14.0625 Gbps Active/ LinkUp)==> 18[ ] "greina12 HCA-1" ( )
 16[ ]=(Down/ Polling)==>[ ] "greina13 HCA-1" ( )
 17[ ]=(Down/ Polling)==>[ ] "greinaagfa01 HCA-1" ( )
 18[ ]=(Down/ Polling)==>[ ] "greina19 HCA-1" ( )
 19[ ]=(Down/ Polling)==>[ ] "router3 HCA-1" ( )
 20[ ]=(Down/ Polling)==>[ ] "router1 HCA-1" ( )
 21[ ]=(Down/ Polling)==>[ ] "router2 HCA-1" ( )
 22[ ]=(Down/ Polling)==>[ ] "greina17 HCA-1" ( )
 23[ ]=(Down/ Polling)==>[ ] "MT2540B ConnectX Mellanox Technologies" ( )
 24[ ]=(Down/ Polling)==>[ ] "greina21 HCA-1" ( )
 25[ ]=(Down/ Polling)==>[ ] "greina20 HCA-1" ( )
 26[ ]=(Down/ Polling)==>[ ] "greina18 HCA-1" ( )
 27[ ]=(Down/ Polling)==>[ ] "greina22 HCA-1" ( )
```
Topologies

Topologies available over IB

- Fat-tree
- Mash
- 2D Torus
- 3D Torus
- DragonFly
- Etc..

Fat-tree Topology
Topologies available over IB

- Fat-tree
- Mash
- 2D Torus
- 3D Torus
- DragonFly
- Etc..

UpDown Routing (Broken Fat-tree)
Topologies
Topologies

- Mesh Network
Topologies

• 2D Torus

32 Compute nodes
16 QDR 8 ports each switch
2D mesh topology
2 IO servers
2 Login nodes
Routing Algorithm

Routing

- Min Hop (shortest length not deadlock free)
- UPDN (Not pure Fat Tree deadlock due to a loop in the subnet)
- Fat-tree (used if the topology is pure Tat-tree)
- DOR (Min Hop deadlock free for Mesh and hypercube topologies)
- Lash (Deadlock free and use the shortest path)
- Torus-2Qos (DOR based deadlock free, used for 2D and 3D Torus)
Infiniband Roadmap

http://www.infinibandta.org/content/pages.php?pg=technology_overview
The HPC Advisory Council and the Swiss Supercomputing Centre will host the HPC Advisory Council Switzerland Conference 2013 in the Lugano Convention Centre, Lugano, Switzerland, from March 13-15, 2013.

The conference will focus on High-Performance Computing essentials, new developments and emerging technologies, best practices and hands-on training. The conference will focus on the following topics:

- Progress of Exascale in the European Union
- High Performance Interconnects, Accelerators and Parallel I/O
- Communication libraries: MPI, SHMEM, PGAS
- GPU computing, CUDA, OpenCL
- Big Data
- Advanced topics / Technologies / development including server and storage systems
- Hands-on: clustering, network, troubleshooting, tuning, optimizations

The conference will bring together system managers, researchers, developers, computational scientists, students and industry affiliates for cross-training and to discuss recent HPC developments and future advancements. Please make sure to register early to ensure your seat at the conference.
Thank you for your attention.