InfiniBand Essentials Every HPC Expert Must Know

Oded Paz

April 2014
HPC

- IB Principles
  - Targets
  - Fabric Components
  - Fabric architecture

- Fabric Discovery Stages
  - Topology discovery
  - Information Gathering
  - Forwarding Tables
  - Fabric SDN
  - Fabric Activation

- Protocol Layers Principle
  - Supported Upper Layer protocols
  - Transport layer
  - Link Layer
  - Physical Layer

- Mellanox Products
  - InfiniBand Switches
  - Channel Adapters
  - Cabling
  - Fabric Management
Leading Supplier of End-to-End Interconnect Solutions

Server / Compute

Virtual Protocol Interconnect
56G IB & FCoIB

10/40/56GbE & FCoE

Switch / Gateway

Virtual Protocol Interconnect
56G InfiniBand

10/40/56GbE

Fibre Channel

Storage Front / Back-End

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs

Adapter Cards

Switches/Gateways

Host/Fabric Software

Cables
Mellanox Common Target Implementations

**HPC**
- Up to 10X Performance and Simulation Runtime
- 33% Higher GPU Performance
- Unlimited Scalability

**Web 2.0**
- 2X Hadoop Performance
- 13X Memcached Performance
- 4X Price/Performance

**DB/Enterprise**
- 10X Database Query Performance
- 4X Faster VM Migration
- More VMs per Server and More Bandwidth per VM

**Cloud**
- 12X More Throughput
- Support More Users at Higher Bandwidth
- Improve and Guarantee SLAs

**Financial Services**
- Lowest Latency
- 62% Better Execution Time
- 42% Faster Messages Per Second

**Storage**
Mellanox storage acceleration software provides >80% more IOPS (I/O operations per second)
Mellanox VPI Interconnect Solutions

**ConnectX-3 VPI Adapter**
- Networking
- Storage
- Clustering
- Management

**ConnectX-3**
- Ethernet: 10/40 Gb/s
- InfiniBand: 10/20/40/56 Gb/s

**SwitchX-2 VPI Switch**
- Unified Fabric Manager
- Switch OS Layer
- 64 ports 10GbE
- 36 ports 40GbE
- 48 10GbE + 12 40GbE
- 36 ports IB up to 56Gb/s
- 8 VPI subnets
Bandwidth

- SDR: 10Gb/s
- DDR: 20Gb/s
- QDR: 40Gb/s
- FDR & FDR 10: 56Gb/s
- EDR: 100Gb/s
- NDR: 160/200Gb/s

Same Software Interface

- 2001: 5usec
- 2002: 2.5usec
- 2003: 1.3usec
- 2004: 0.7usec
- 2005: 0.5usec
- 2006: <0.5usec

Latency
InfiniBand Trade Association (IBTA)

- Founded in 1999
- Actively markets and promotes InfiniBand from an industry perspective through public relations engagements, developer conferences and workshops
- InfiniBand software is developed under OpenFabrics Open Source Alliance
  http://www.openfabrics.org/index.html
- InfiniBand standard is developed by the InfiniBand Trade Association (IBTA)
  http://www.infinibandta.org/home

Steering Committee Members:
InfiniBand is a Switch Fabric Architecture

- Interconnect technology connecting CPUs and I/O

- Super high performance
  - High bandwidth (starting at 10Gb/s and up to 100Gb/s)
  - Low latency—fast application response across the cluster < 1µs end to end
    (Mellanox switches 170 nanosec per HOP)
  - Low CPU utilization with RDMA (Remote Direct Memory Access) –
    Unlike Ethernet, TRAFFIC communication bypasses the OS and the CPU’s.

First industry standard high speed interconnect!
InfiniBand is a Switch Fabric Architecture

- InfiniBand was originally designed for large-scale grids and clusters
- Increased application performance
- Single port solution for all LAN, SAN, and application communication
- High reliability CLUSTER management (Redundant Subnet Manager)
- Automatic Cluster switches and ports configuration performed by the Subnet Manager SW

First industry-standard high speed interconnect!
RDMA – How Does it Work

**RDMA over InfiniBand**

**USER**

1. Application
2. Buffer 1

**KERNEL**

1. OS
2. Buffer 1

**HARDWARE**

1. NIC
2. Buffer 1

**RACK 1**

**RACK 2**

TCP/IP
The InfiniBand Architecture

- Industry-standard defined by the InfiniBand Trade Association

- Defines System Area Network architecture
  - Comprehensive specification: from physical to applications

- Architecture supports
  - Host Channel Adapters (HCA)
  - Switches
  - Routers
InfiniBand Components Overview

- **Host Channel Adapter (HCA)**
  - Device that terminates an IB link and executes transport-level functions and support the verbs interface

- **Switch**
  - A device that moves packets from one link to another of the same IB Subnet

- **Router**
  - A device that transports packets between different IBA subnets

- **Bridge/Gateway**
  - InfiniBand to Ethernet
Host Channel Adapters (HCA)

- Equivalent to a NIC (Ethernet)
  - GUID Global Unique ID
- Converts PCI to InfiniBand
- CPU offload of transport operations
- End-to-end QoS and congestion control

HCA bandwidth options:

- **Single Data Rate** 2.5GB/S * 4 = 10
- **Double Data Rate** 5 GB/S * 4 = 20
- **Quadruple Data Rate** 10GB/S * 4 = 40
- **Fourteen Data Rate** 14 Gb/s * 4 = 56
- **Enhanced Data Rate** 25 Gb/s * 4 = 100
Any InfiniBand node requires GUID&LID addresses

**GUID (Global Unique Identifier)**- 64 bits address, “Like a Ethernet MAC address”
- Assigned by IB vendor
- Persistent through reboots

**IB Switch “Multiple” Address GUIDS**
- **Node** = Is meant to identify the HCA as a entity
- **Port** = Identifies the port as a port
- **System** = Allows to combine multiple GUIDS creating one entity
- A single 36 ports IB switch chip, is the Basic Block for every IB switch module

- We create a multiple ports switching module using multiple chips

- In this example we create 72 ports switch, using 6 identical chips:
  - 4 chips will function as lines
  - 2 chips will function as core
Local Identifier- 16 bit L2 Address
• Assigned by the Subnet Manager when port becomes active
• Not persistent through reboots

LID Address Ranges
• 0x 0000 = Reserved
• 0x0001 = 0xBFFF = Unicast
• 0xc001 = 0xFFFE = Multicast
• 0xFFFF = Reserved for special use
- Define different partitions for different customers
- Define different partitions for different applications
- Allows fabric partitioning for security purposes
- Allows fabric partitioning for Quality of Service (QoS)
- Each partition has an Identifier named PKEY
### Usage
- A 128 bit field in the Global Routing Header (GRH) used to route packets between different IB subnets
- Multicast groups port identifier IB & IPOIB

### Structure
- GUID- 64 bit identifier provided by the manufacturer
- IPv6 type header
- Subnet Prefix: A 0 to 64-bit:
  - Identifier used to uniquely identify a set of end-ports which are managed by a common Subnet Manager

```
port GUID: 0x0002c90300455fd1
default gid: fe80:0000:0000:0x0002c90300455fd1
```
IB Basic Management Concepts

- Node: any managed entity—End Node, Switch, Router
- Manager: active entity; sources commands and queries
  - The subnet manager (SM)
- Agent: passive (mostly) entity that will reside on every node, responds to Subnet Managers queries
- Management Datagram (MAD):
  - Standard message format for manager–agent communication
  - Carried in an unreliable datagram (UD)
Objectives of Subnet Management

- Initialization and configuration of the subnet elements
- Establishing best traffic paths between source to destination through the subnet
- Fault isolation
- Continue these activities during topology changes
- Prevent unauthorized Subnet Managers
IB Port Basic Identifiers

- Host Channel Adapter—HCA (IB “NIC”)
- Port number
- Global Universal ID—GUID 64 bit (like mac) ex. 00:02:C9:02:00:41:38:30
  - Each 36 ports “basic” switch has its own switch & system GUID
  - All ports belong to the same “basic” switch will share the switch GUID
- Local Identifier - LID

### LID

- Local Identifier that is assigned to any IB device by the SM and used for packets switching within an IB fabric.
- All ports of the same ASIC unit are using the same LID
Node & Switch Main identifiers

- **Virtual Lane**
  - Each Virtual Lane uses different buffers to send its packet towards the other side
  - VL 15 is used for management only SM traffic
  - VL 0-7 are used for traffic
  - Used to separate different bandwidth & QoS using same physical port
Node & Switch Main identifiers

Traffic Packets VL 0-7

Mux

De-Mux

Traffic Packets VL 0-7

Link Control
VL-15

Packets Transmitted

Traffic Packets VL 0-7

Host

InfiniBand Link

VL-15, VL 0-7

LID 14
GUID-00:02:C9:02:00:41:38:35
Port 17

LID 37
GUID-00:02:C9:02:00:41:38:35
Port 1

Master SM

LID 8

LID 12
Subnet Manager Cluster Discovery
Subnet Manager & Fabric configuration Process

1. Physical Subnet Establishment
2. Subnet Discovery
3. Information Gathering
4. LID Assignment
5. Path Establishment
6. Port Configuration
7. Switch Configuration
8. Subnet Activation
Subnet Manager (SM) Rules & Roles

- Every subnet must have at least one
  - Manages all elements in the IB fabric
  - Discover subnet topology
  - Assign LIDs to devices
  - Calculate and program switch chip forwarding tables (LFT pathing)
  - Monitor changes in subnet

- Implemented anywhere in the fabric
  - Node, Switch, Specialized device

- No more than one **active** SM allowed
  - 1 Active (Master) and remaining are Standby (HA)

```
[root@l-supp-18 ~]# sminfo
sminfo: sm lid 44 sm guid 0x2c9030010392b, activity count 1372449 priority 14 state 3 SMINFO_MASTER
[root@l-supp-18 ~]# saquery -s
IsSM ports
PortInfoRecord dump:
```
1. The **SM wakes up** and starts the Fabric Discovery process

2. The SM starts “**conversation**” with every node, over the InfiniBand link it is connected to. In this stage the **discovery stage**, the SM collects:
   - Switch Information followed by port information
   - Host information

3. Any switch which is already discovered, will be used as a gate for the SM, for further discovery of all **this switch links** and the switches it is connected to known also as its neighbors.
4. The SM gathers information by sending and receiving SMPs (Subnet Management Packets)
   a. These special management packets are sent on Virtual Lane 15 (VL15)
      • VL15 is a special NON flow controlled VL
   b. Two primary “types” of SMPs creating Cluster routing table:
      • Directed routing (DR) table based on Nodes GUIDS & port number
      • This is the type primarily used by OpenSM
   c. LID routing (LR)
      • Topology and than packets routing table, Based on the LIDS which have been assigned to each node by the SM
Fabric Information Gathering During Discovery

- **Node Info Gathered**
  - Node type
  - Num of ports
  - GUID
  - Partition table size

- **Port Info Gathered**
  - Forwarding Database size
  - MTU
  - Width
  - VLS
Fabric Direct Route Information Gathering

- **Building** the direct routing table from & to each one of the fabric elements
- Each node in a path is identified by its port number & GUID
- The table content is saved in the SM LMX table

<table>
<thead>
<tr>
<th>Switch h-1</th>
<th>Switch-2</th>
<th>Switch-5</th>
<th>Switch-3</th>
<th>Switch-6</th>
<th>Switch-4</th>
<th>H-11</th>
<th>H-16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch h-1</td>
<td>Port2</td>
<td>Port2</td>
<td>Port3</td>
<td>Port3</td>
<td>Port8</td>
<td>Port 8 Switch 4_Port30</td>
<td>Switch 3_Port 5 Switch 6_Port29</td>
</tr>
<tr>
<td>H11</td>
<td>Port1 Switch 6_Port2 Switch 3_Port4 Switch 4_Port30</td>
<td>Port 1 Switch 6_Port2 Switch 3_Port1 Switch1_P ort2 Switch2_P ort2</td>
<td>Port 1 Switch 6_Port2 Switch 3_Port4</td>
<td>Port 1 Switch 6_Port2 Switch 3_Port4</td>
<td>Port 1 Switch 6_Port2 Switch 3_Port4 Switch 4_Port30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Diagram of network topology with ports and switches labeled]
LID Assignment

- After the SM finished gathering any needed subnet information, it assigns a base LID and LMC to each one of the attached end ports
  - The LID is assigned to at the port rather than device level
  - Switch external ports do not get/need LIDs

- The DLID is used as the main address for InfiniBand packet switching
- Each **Switch port** can be identified by the **combination** of LID & port number

```bash
[root@v-sup25 ~]# ibswitches
Switch : 0x0008f105060000de ports 36 "Mellanox sLB-4018 #4700-89B8" enhanced port 0 lid 13 lmc 0
Switch : 0x0008f1050650c4a ports 36 "Mellanox sLB-4018 #4700-89B8" enhanced port 0 lid 9 lmc 0
```

```
[root@v-sup25 ~]# saquery LFTR 9
LFT Record dump:

<table>
<thead>
<tr>
<th>LID</th>
<th>Port Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>255</td>
</tr>
<tr>
<td>1</td>
<td>255</td>
</tr>
</tbody>
</table>
```
Linear Forwarding Table Establishment (Path Establishment)

- After the SM finished gathering all Fabric information, including direct route tables, it assigns a LID to each one of the NODES.
- At this stage the LMX table will be populated with the relevant route options to each one of the nodes.
- The output of the LMX will provide the Best Route to Reach a DLID as well as the other Routes.
- The Best Path Result will be based on Shortest Path First (SPF) algorithm.

<table>
<thead>
<tr>
<th>D-LID</th>
<th>PORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>1 2 3 1</td>
</tr>
<tr>
<td>22</td>
<td>2 1 2 1</td>
</tr>
<tr>
<td>23</td>
<td>3 2 1 1</td>
</tr>
<tr>
<td>75</td>
<td>3 2 3 2</td>
</tr>
<tr>
<td>81</td>
<td>4 3 4 3</td>
</tr>
<tr>
<td>82</td>
<td>4 3 2 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Dest. LID</th>
<th>Best Route/exit port</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>75</td>
<td>3</td>
</tr>
<tr>
<td>81</td>
<td>3</td>
</tr>
<tr>
<td>82</td>
<td>8</td>
</tr>
</tbody>
</table>
LID Routed (LR) Forwarding

- Uses the LFT tables
- Based on the data gathered on the LMX – Direct Routing
- It is the standard routing of packets used by switches
- Uses regular link-level headers to define destination and other information, such as:
  - DLID = LID of the final destination
  - SL = Service Level of the path
  - Each switch uses the forwarding table and SL to VL table to decide on the packet’s output port/VL

```
[root@v-sup25 ~]# saquery LFTR 9
LFT Record dump:
    LID.................9
    Block................0
    LFT:
    LID    Port Number
    0      255
    1      255
```

<table>
<thead>
<tr>
<th>The Destination LID</th>
<th>Best Route/exit port</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>75</td>
<td>3</td>
</tr>
<tr>
<td>81</td>
<td>3</td>
</tr>
<tr>
<td>82</td>
<td>8</td>
</tr>
</tbody>
</table>
**LID Routed (LR) Forwarding**

- **LRH: Local Routing Header:**
  - Source & Destination LID
  - Service Level-SL
  - Virtual Lane-VL
  - Packet Length

![InfiniBand Data Packet Diagram]

<table>
<thead>
<tr>
<th>The Dest. LID</th>
<th>Best Route/exit port</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>75</td>
<td>3</td>
</tr>
<tr>
<td>81</td>
<td>3</td>
</tr>
<tr>
<td>82</td>
<td>8</td>
</tr>
</tbody>
</table>
- **Light sweep**:  
  - Routine sweep of the Subnet Manager  
  - By default runs every 30 seconds  
  - Requires all switches to switch and port info

- **Light Sweep traces**:  
  - Ports status change  
  - New SM speaks on the subnet  
  - Subnet Manager changes priority
Any change traced by the light sweep will cause **Heavy Sweep**

**IB TRAP**

- Changes of status of a switch will cause an on line IB TRAP that will be sent to the Subnet Manager and cause **Heavy Sweep**

**Heavy Sweep**

- Will cause all SM fabric discovery to be performed from scratch
InfiniBand Fabric Topologies
InfiniBand Fabric Commonly Used Topologies

Modular switches are based on Fat Tree architecture:

- Back to Back
- 3D Torus
- 2 Tier Fat Tree
- Dual Star
- Hybrid
The IB Fabric Basic Building Block

- A single 36 ports IB switch chip, is the Basic block for every IB switch module

- We create a multiple ports switching Module using multiple chips

- In this example we create 72 ports switch, using 6 identical chips
  - 4 chips will function as lines
  - 2 chips will function as core
CLOS Topology

- **Pyramid Shape Topology**
- The switches at the top of the pyramid are called Spines/Core
  - The Core/Spine switches are interconnected to the other switch environments
- The switches at the bottom of the Pyramid are called Leafs/Lines/Edges
  - The Leaf/Lines/Edge are connected to the fabric nodes/hosts
- In a non blocking CLOS fabric there are equal number of external and internal connections
- External connections:
  - The connections between the hosts and the Line switches

- Internal Connections
  - The connections between the core and the Line switches

- In a non-blocking fabric, there is always a balanced cross-bisectional bandwidth

- In case the number of external connections is higher than internal connections, we have a blocking configuration
The topology detailed here is called CLOS 3

The maximum traffic path between source to destination includes 3 HOPS (3 switches)

Example a session between A to B

- One Hop from A to switch L1-1
- Next Hop from switch L1-1 to switch L2-1
- Last Hop from L2-1 to L1-4
In this example we can see 108 non blocked fabric
• 108 hosts are connected to the line switches
• 108 links connect between the line switches to the core switches to enable non blocking interconnection of the line switches

$18 \times 6 = 108$
IB Fabric Protocol Layers
IB Switch - L2

Upper Level Protocols

Transport Layer

Network Layer

Link Layer

Physical Layer

Client

Transactions

IBA Operations

SAR

Messages

Que Pairs

Inter Subnet Routing

Network

L2 Switching LID Based

Link Encoding
Media Access Control

Packet Relay

MAC

MAC

Link Encoding
Media Access Control

End Node

Switch L2

End Node
**IB Architecture Layers**

- **Software Transport Verbs and Upper Layer Protocols:**
  - Interface between application programs and hardware.
  - Allows support of legacy protocols such as TCP/IP
  - Defines methodology for management functions

- **Transport:**
  - Delivers packets to the appropriate Queue Pair;
  - Message Assembly/De-assembly, access rights, etc.

- **Network:**
  - How packets are routed between different partitions/subnets

- **Data Link (symbols and framing):**
  - From source to destination on the same partition subnet
  - Flow control (credit-based); How packets are routed

- **Physical:**
  - Signal levels and frequency, media, connectors

---

**Diagram:**

- Client
- Transactions
- Client
- IBA Operations
- Messages Que Pairs
- Network
- Inter Subnet Routing
- Network
- Link Encoding
- L2 Switching LID Based
- Link Encoding
- End Node
- Switch L2
- End Node
InfiniBand Header Structure

Upper Layer Protocol
Original Message

Transactions
Messages

Upper Layer Protocol
Original Message

Transport

Packets
Que Pairs

Inter Subnet Routing
Subnet Prefix +GUID

Network

L2 Switching LID Based

Link Layer

Network

Link Layer

Transport

Physical Layer

Physical Layer

End Node

Switch L2

End Node
Transport Layer – Responsibilities

- The **Network** and **Link** protocols deliver a packet to the desired destination.
- The **Transport Layer**
  - **Segmenting Assembly & Reassembly of**:
    - Messages data payload coming from the Upper Layer, into multiple packets that will suit valid **MTU** size
  - Delivers the packet to the proper Queue Pair (assigned to a specific session)
  - Instructs the **QP** how to process the packet’s data (**Work Request Element**)
  - Reassembles the packets arriving from the other side into messages

![Diagram showing Transport Layer responsibilities with WQE, Transmit, Receive, Local QP, Transmit, Receive, Remote QP]
Transport Layer – Responsibilities

Upper Layer Protocol
Original Message

Transactions
Messages

Upper Layer Protocol
Original Message

Transport

SAR SAR SAR

Packets
Que Pairs

Inter Subnet Routing
Subnet Prefix +GUID

Network

Transport

SAR SAR SAR

Link Layer

L2 Switching LID Based

Network

Link Layer

Physical Layer

Physical Layer

End Node

Switch L2

End Node

Physical Layer

Upper Layer Protocol
Original Message

Transactions
Messages

Upper Layer Protocol
Original Message

Transport

SAR SAR SAR

Packets
Que Pairs

Inter Subnet Routing
Subnet Prefix +GUID

Network

Transport

SAR SAR SAR

Link Layer

L2 Switching LID Based

Network

Link Layer

Physical Layer

Physical Layer

End Node

Switch L2

End Node

Physical Layer

Transport Layer – Responsibilities

Upper Layer Protocol
Original Message

Transactions
Messages

Upper Layer Protocol
Original Message

Transport

SAR SAR SAR

Packets
Que Pairs

Inter Subnet Routing
Subnet Prefix +GUID

Network

Transport

SAR SAR SAR

Link Layer

L2 Switching LID Based

Network

Link Layer

Physical Layer

Physical Layer

End Node

Switch L2

End Node

Physical Layer
Layer 2 Forwarding

- Switches use FDB (Forwarding Database)
  - Based on DLID and SL, a packet is sent to the correct output port + specific VL
**Link Layer – Flow Control**

- Credit-based link-level flow control
  - Link Flow Control assures NO packet loss within fabric even in the presence of congestion
  - Link **Receivers** grant packet receive buffer space credits per Virtual Lane
  - Flow Control credits are issued in 64 byte units

- Separate flow control per Virtual Lanes provides:
  - Alleviation of head-of-line blocking

- Virtual Fabrics
  Congestion and latency on one VL, does not impact traffic with guaranteed QoS on another VL, even though they share the same Physical link
Physical Layer- Responsibilities

- InfiniBand is a **Lossless** fabric.

- Maximum Bit Error Rate (BER) allowed by the IB spec is $10^{-12}$. Statistically Mellanox fabrics provides around $10^{-15}$

- The Physical layer should guaranty affective signaling to meet this BER requirement
Physical Layer Cont

- **Industry standard Media types**
  - Copper: 7 Meter QDR, 3 METER FDR
  - Fiber: 100/300m QDR & FDR

- **64/66 encoding on FDR links**
  - Encoding makes it possible to send digital high speed signals to a longer distance enhances performance & bandwidth effectiveness
  - X actual data bits are sent on the line by Y signal bits
  - $64/66 \times 56 = 54.6\text{Gbps}$

- **8/10 bit encoding (DDR and QDR)**
  - X/Y line efficiency (example $80\% \times 40 = 32\text{Gbps}$)
Mellanox Cables – Perceptions Vs. Facts

- Mellanox cables are rebranded from a cable vendor
  - Mellanox cables are manufactured by Mellanox

- Our vendor can sell the same cables
  - No other vendor is allowed to sell Mellanox cables

- Mellanox cables use a different assembly procedure

- Mellanox cables are tested with unique test suite

- Vendors’ “Finished Goods” fail Mellanox dedicated testing

- Mellanox allows the customers to use any IBTA IB approved cables
Mellanox Passive Copper Cables

- Superior design and qualification process
- Committed to Bit Error Rate (BER), better than $10^{-15}$
- Longest reach with Mellanox end-to-end solution

<table>
<thead>
<tr>
<th>Data Rate</th>
<th>PCC Max Reach</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR</td>
<td>3 meter</td>
</tr>
<tr>
<td>FDR10</td>
<td>5 meter</td>
</tr>
<tr>
<td>QDR</td>
<td>7 meter</td>
</tr>
<tr>
<td>40GbE</td>
<td>7 meter</td>
</tr>
<tr>
<td>10GbE</td>
<td>7 meter</td>
</tr>
</tbody>
</table>
Mellanox Active Fiber Cables

- Superior design and qualification process
- Committed to Bit Error Rate (BER), better than $10^{-15}$
- Longest reach with Mellanox end-to-end solution
- Optical Performance Optimization (patent pending)

<table>
<thead>
<tr>
<th>Data Rate</th>
<th>Max Reach</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR</td>
<td>300 meter</td>
</tr>
<tr>
<td>FDR10</td>
<td>100 meter</td>
</tr>
<tr>
<td>QDR</td>
<td>300 meter</td>
</tr>
<tr>
<td>40GbE</td>
<td>100 meter</td>
</tr>
</tbody>
</table>
Mellanox Family Products
Leading Supplier of End-to-End Interconnect Solutions

Server / Compute
- Virtual Protocol Interconnect
- 56G IB & FCoIB
- 10/40/56GbE & FCoE

Switch / Gateway
- Virtual Protocol Interconnect
- 56G InfiniBand
- 10/40/56GbE
- Fibre Channel

Storage Front / Back-End

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

<table>
<thead>
<tr>
<th>ICs</th>
<th>Adapter Cards</th>
<th>Switches/Gateways</th>
<th>Host/Fabric Software</th>
<th>Cables</th>
</tr>
</thead>
</table>

Mellanox Training Center
Training Material
InfiniBand Switches & Gateways
FDR InfiniBand Switch Portfolio

Modular Switches
- 648 port
- 324 port
- 216 port
- 108 port

Edge Switches
- SX6025 – 36 ports externally managed
- SX6015 – 18 ports externally managed
- SX6005 – 12 ports externally managed
- SX6036 – 36 ports managed
- SX6018 – 18 ports managed
- SX6012 – 12 ports managed

Long Distance
- NEW

Bridge – VPI
- NEW

Management
- NEW

**NEW**
SwitchX® VPI Technology Highlights

Virtual Protocol Interconnect® (VPI)
One Switch – Multiple Technologies

1. VPI on Box
   Same box runs InfiniBand OR Ethernet

2. VPI per Port
   Same box runs InfiniBand AND Ethernet

3. VPI Bridging
   Same box bridges InfiniBand AND Ethernet

InfiniBand
Ethernet
MetroX™ - Mellanox Long-Haul Solutions

- Provides InfiniBand and Ethernet Long-Haul Solutions of up to 80km for campus and metro applications.
- Connecting between data centers deployed across multiple geographically distributed sites.
- Extending InfiniBand RDMA and Ethernet RoCE beyond local data centers and storage clusters.
- Perfect cost-effective, low power, easily managed and scalable solution.
- Managed as a single unified network infrastructure.
# MetroDX and MetroX Features

<table>
<thead>
<tr>
<th></th>
<th>TX6100</th>
<th>TX6240</th>
<th>TX6280</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distance</strong></td>
<td>1KM</td>
<td>10KM</td>
<td>40KM</td>
</tr>
<tr>
<td><strong>Throughput</strong></td>
<td>640Gb/s</td>
<td>240Gb/s</td>
<td>80Gb/s</td>
</tr>
<tr>
<td><strong>Port Density</strong></td>
<td>16p X FDR10 long haul</td>
<td>6p X 40Gb/s long haul</td>
<td>2p X 10/40Gb/s long haul</td>
</tr>
<tr>
<td></td>
<td>16p X FDR downlink</td>
<td>6p X 56Gb/s downlink</td>
<td>2p X 56Gb/s downlink</td>
</tr>
<tr>
<td><strong>Latency</strong></td>
<td>200ns + 5us/km over fiber</td>
<td>200ns + 5us/km over fiber</td>
<td>700ns + 5us/km over fiber</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>~200W</td>
<td>~200W</td>
<td>~280W</td>
</tr>
<tr>
<td><strong>QoS</strong></td>
<td>One data VL + VL15</td>
<td>One data VL + VL15</td>
<td>One data VL + VL15</td>
</tr>
<tr>
<td><strong>Space</strong></td>
<td>1RU</td>
<td>1RU</td>
<td>2RU</td>
</tr>
</tbody>
</table>
Mellanox Host Channel Adapters (HCA)

Reference to the following Document:
ConnectX®-3 VPI Single and Dual QSFP Port Adapter Card User Manual
HCA ConnectX-3 InfiniBand Main Features

- Up to 56Gb/s InfiniBand or 40 Gigabit Ethernet per port
- PCI Express 3.0 (up to 8GT/s)
- CPU offload of transport operations
- Application offload
- GPU communication acceleration
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- Dynamic power management
- Fiber Channel encapsulation (FCoIB or FCoE)
- Ethernet encapsulation (EoIB)
Adapters offering

**ConnectX-3 Pro**
- NVGRE and VxLAN HW off-load
- RoCE V2(UDP)
- ECN\QCN

**ConnectX-3**
- VPI
- Up to 56Gb IB
- Up to 56 GbE
- RDMA
- CPU off-load
- SR-IOV

**Connect-IB**
- Up to 56Gb IB
- Greater than 100Gb bi-directional DC
- T10\DIF
- PCIE x16
- More than 130M message/sec
Fabric Management
Goal of Fabric Utilities in HPC Context

- **Enable fast cluster bring-up**
  - Point out issues with devices, systems, cables
  - Provide inventory including cables, devices, FW, SW
  - Perform device specific (proprietary) checks
  - Eye-Opening and BER checks
  - Catch cabling mistakes

- **Validate Subnet Manager work**
  - Verify connectivity at the lowest level possible
  - Report subnet configuration
  - SM agnostic
Goal of Fabric Utilities in HPC Context

- Diagnose L2 communication failures
  - At the entire subnet level
  - On a point to point path

- Monitor the Network Health
  - Continuous and with low overhead
Fabric Management Solution Overview

- **ibutils: ibdiagnet/ibdiagpath**
  - An automated L2 health analysis procedure
  - Text interface
  - No dedicated “monitoring” mode
  - Significant development past year on features and runtime performance at scale

- **UFM**
  - Highend monitoring and Provisioning capabilities
  - GUI based with CLI options
  - Includes ibutils capabilities with additional features
  - **Central device management**
    - Fabric dashboard
    - Congestion analysis
  - **System Integration Capabilities**
    - SMP Traps and Alarms
UFM in the Fabric

- Software or Appliance form factor
- 2 or more High Availability
- Switch and HCA management
- Full Mgmt or Monitoring Only modes

Synchronization, Heartbeat
Multi-Site Management

Site1

Site2

... 

Site n

UFM Multi-Site Portal

Site Name | UFM IP | UFM Vendor | Status | LSs | Networks | Switches | Has | HA Status | Alarms | GUI
---|---|---|---|---|---|---|---|---|---|---
Site 1 | 192.168.10.20 | 3.5 build 7 | INFO | 2 | 2 | 5 | Active | 4 | Launch
Site 2 | 192.168.10.10 | 3.5 build 7 | INFO | 2 | 2 | 5 | Active | 4 | Launch
Site n | | | | | | | | | |
Integration with 3rd Party Systems

- Extensible architecture
  - Based on Web-services

- Open API for users or 3rd-party extensions
  - Allows simple reporting, provisioning, and monitoring
  - Task automation
  - Software Development Kit

- Extensible object model
  - User-defined fields
  - User-defined menus
UFM – Comprehensive Robust Management

Automatic Discovery
Central Device Management
Fabric Dashboard
Congestion Analysis

Fabric Health Reports
Service Oriented Provisioning
Health & Performance Monitoring
UFM Main Features

- Automatic Discovery
- Central Device Mgmt
- Fabric Dashboard
- Congestion Analysis
- Health & Perf Monitoring
- Advanced Alerting
- Fabric Health Reports
- Service Oriented Provisioning
**Advanced Monitoring and Analysis**

- **Monitor & analyze fabric performance**
  - Bandwidth utilization
  - Unique congestion monitoring
  - Dashboard for aggregated fabric view

- **Real-time fabric-wide health monitoring**
  - Monitor events and errors through-out the fabric
  - Threshold based alarms
  - Granular monitoring of host and switch parameters

- **Innovative congestion mapping**
  - One view for fabric-wide congestion and traffic patterns
  - Enables root cause analysis for routing, job placement or resource allocation inefficiencies

- **All is managed at the job/aggregation level**