Infiniband fundamentals

José M. Cámara
(checam@ubu.es)
v. 1.0
Introduction I

- SAN (System Area Network) Interconnect.
- Specification produced by a vendors consortium.
- Intended to create low latency high bandwidth connections.
- Message passing system.
Introduction II

• IFB (Infiniband) components.
  – Computational nodes
  – Peripherals
  – I/O nodes
  – Switches
  – Routers

• It is a multi stage switched network. Elements are connected in a mesh like topology.
Low latency is achieved by means of solutions like:

- RDMA: remote direct memory access. Devices are granted access to other one’s local memory.
- Most communication management is hardware implemented (protocol stack).

High bandwidth is achieved through high rate links (up to QDR) and many data lines per link (up to 12). $12 \times \text{QDR}(4 \times 2.5 \text{Gbps}) = 120 \text{ Gbps}$.
Concepts

- **Processor node**: a group of one or more processors and their associated memory.
- **Port**: a bi-directional interface that connects a device to a link.
- **Link**: bi-directional point to point high speed connection between two ports on two different devices.
- **Channel adapter**: hardware system implementing one or more ports.
- **Subnet**: a group of devices sharing a common subnet address managed by a common subnet manager.
- **Subnet manager**: seeks and sets up all devices within the subnet at start up. At runtime it watches any possible modification on the subnet.
  - It is a software entity.
  - It typically resides on a processor node.
  - It communicates with the devices by means of subnet management packets (SMPs) sent through the subnet management interface (SMI) available at each port in a channel adapter and in every router port.
  - Switches implement the SMI at port 0.
Infiniband network layout
Local addressing (Subnet) I

• Channel adapters as well as routers may implement up to 255 ports numbered from 1 to 255.

• Each port’s address is unique within the subnet. It is assigned by the subnet manager at start up. It can be assigned through a range of addresses.

• Switches implement from 3 to 255 ports numbered starting at 0. Port 0 is the mandatory port for switch management purposes. May no have external connection.
  – Only port 0 is given a local address.

• Local addresses (LID) are 16 bits long.
  – Address 0000h is reserved.
  – Addresses 0001h – BFFFh are used as unicast destinations.
  – Addresses C000h – FFFEh are used as multicast destinations.
  – Address FFFFh is a special purpose one.
Local addressing process:
- At the source node the channel adapter builds up the information packets with a local routing header (LRH) containing:
  - LHR:DLID: destination port’s LID.
  - LHR:SLID: source port’s LID.
- The channel adapter injects the packet into the network.
- The packet gets to a port on the first switch on its way.
- The switch’s link layer examines LHR:DLID to see whether the address is unicast or multicast.
  - If unicast, a table generated by the subnet manager selects an output port.
  - If multicast, there is another table to determine all suitable output ports.
- The same procedures applies to the following switches on the road.
- When the destination’s channel adapter is reached, the packet is delivered to the network layer to be consumed.
Global Addressing I

• To address ports placed on different networks (at least one router in between).
• Apart from the LRH, the packet must include a Global Routing Header (GRH) containing:
  – DGID: destination’s port ID and its subnet.
  – SGID: destination’s port ID and its subnet.
Global Addressing II

• Global addressing process:
  – Source node channel adapter inserts the following addressing data into the packet:
    • LRH:SLID
    • LRH:DLID
    • SGID: upper 64 bits for source subnet ID + lower 64 bits for source port’s global address.
    • DGID: upper 64 bits for destination subnet ID + lower 64 bits for the unique destination port’s global address.
  – DLID identifies destination port within the source subnet (always a router port).
  – From GRH:DGID the router determines whether the destination is placed on one of the subnets connected to it or not.
    • In case it is, the Router determines the local address of the destination port.
    • Otherwise, the Router determines which output port leads to the next router.
References

- [http://www.youtube.com/watch?v=pmBpWPqIlvs&feature=related](http://www.youtube.com/watch?v=pmBpWPqIlvs&feature=related)
- “Infiniband Network Architecture”, Tom Shanley, MindShare Inc.