CHAPTER

InfiniBand, iWARP, and RoCE

11

Manoj Wadekar

Fellow, Chief Technologist QLogic Corporation

11.1 Introduction

InfiniBand (IB) is a point-to-point interconnect. Its features, such as zero-copy and remote direct memory access (RDMA), help reduce processor overhead by directly transferring data from sender memory to receiver memory without involving host processors. This chapter covers the overall IB architecture (IBA) and its various layers. The emphasis of this chapter is on the link layer and network layer. As IB evolves to provide connectivity for low-latency applications over Ethernet, Internet Wide Area RDMA Protocol (iWARP) and RoCE are becoming attractive options for providing RDMA functionality to applications. This chapter covers these two protocols in detail.

11.2 InfiniBand architecture

IBA defines a switched communications fabric allowing many devices to concurrently communicate with high bandwidth and low latency in a protected, remotely managed environment. An end node can communicate over multiple IBA ports and can utilize multiple paths through the IBA fabric.

Figure 11.1 demonstrates various components in the IB [10] system network. This network consists of various processor nodes and I/O units connected through cascaded switches and routers. It allows low-latency interconnect for interprocessor communication, support connectivity for storage devices to storage devices, and also demonstrates that routers can be used to extend the connectivity to wide area networks (WANs), local area networks (LANs, over Ethernet), or storage area networks (SANs). Routers also provide connectivity between multiple IB subnets.

IBA also defines architectural components that allow communication with other Layer 2 technologies like Small Computer System Interface (SCSI), Ethernet, and Fibre Channel (FC).

IBA defines a layered protocol that specifies physical, link, network, transport, and upper layers. It defines communication over various media including printed circuit boards (PCBs), copper, and also fiber cable. IB allows three link speeds:

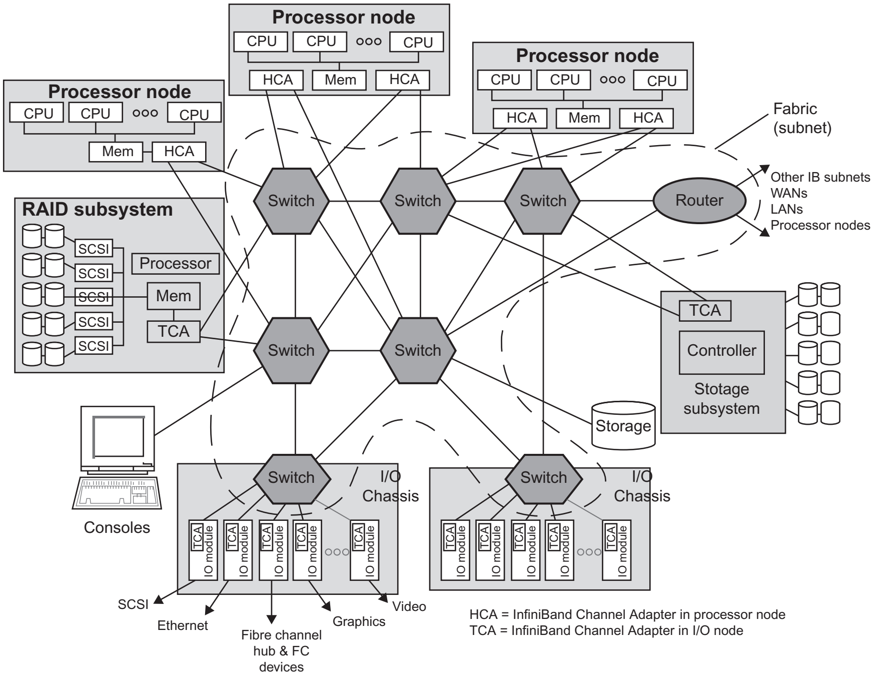


FIGURE 11.1

IB fabric and components.

over 4 wires (1X: single lane), 16 wires (4X: 4 lanes), or 48 wires (12X: 12 lane). So if a single lane (4 wires: differential pairs RX+/RX− and TX+/TX−) is at 2.5 Gbps, then 4X connectivity provides 10 Gbps and 12X connectivity provides communication at 30 Gbps.

IBA supports unicast and multicast traffic between nodes. Such traffic can be carried out in reliable or unreliable mode. It also supports connection or datagram mode for communication. Various QoS (quality of service) mechanisms are provided within the architecture to guarantee lossless and differentiated traffic in the network. The following sections will describe all these aspects of IBA in more detail.

11.3 IB network

11.3.1 Network topology

The IBA network is comprised of multiple subnets that can be connected through routers. Each end node can be a processing node, an I/O unit, or a storage subsystem. IBA allows communication between these participating nodes using the RDMA protocol. This enables very low latency data transfers and also low CPU utilization for Inter Process Communication (IPC) applications. Remote data placement is achieved directly between source nodes and destination nodes—so data copy is avoided and OS involvement is minimized. These factors together reduce overall latency as well as CPU overhead (Figure 11.2).

Any IB device can be connected with one or multiple IB devices or switches. One or multiple links can be used for such connectivity.

11.3.2 Subnet components

11.3.2.1 Channel Adapters

IBA defines two types of adapters that reside in servers or I/O systems. If the adapter resides in the host system (e.g., servers), it is called a Host Channel Adapter (HCA). If the adapter resides in the storage target system—it is called a Target Channel Adapter.

A Channel Adapter provides connectivity for operating systems and applications to physical ports. An HCA provides an interface to the operating system and provides all the verb interfaces defined by IB. Verb is an abstract interface between the application and the functionality provided by a Channel Adapter. Each adapter can have one or multiple ports. Each port provides further differentiation of traffic through a "virtual lane" (VL). Each VL can be flow-controlled independently. As can be seen in Figure 11.3, DMA can be initiated from local as well as remote applications.

A Channel Adapter carries a unique address called a globally unique identifier (GUID). Each port of the adapter is also assigned a unique identifier, a Port GUID. GUIDs are assigned by the adapter vendor. The management entity for a given subnet, called the subnet manager (SM), assigns local identifiers (LIDs) to each port of a Channel Adapter.

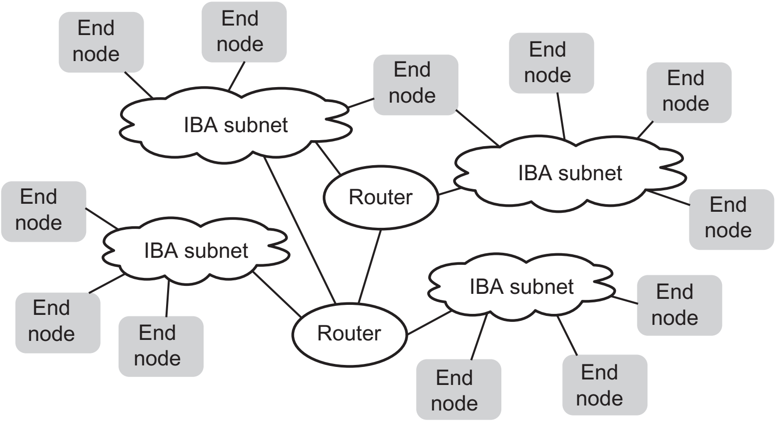


FIGURE 11.2

IBA network components.

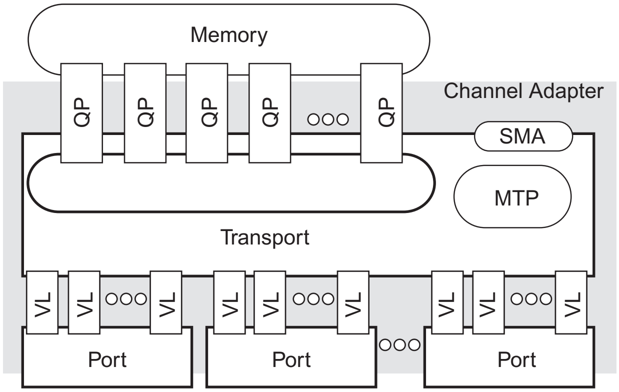


FIGURE 11.3

IB Channel Adapter.

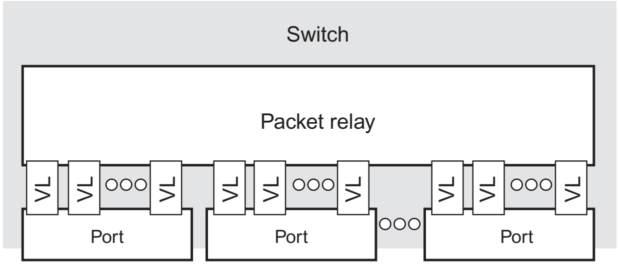


FIGURE 11.4

IB switch.

11.3.2.2 Switches

A switch contains multiple IB ports. It forwards packets between adapter ports based on the destination address in the local routing header (LRH) of each packet. Switches forward a unicast packet to a single destination port and multicast packets to multiple ports as configured in their forwarding database. The forwarding database used by a switch is configured by the SM. Switches just forward the packets—they do not modify packets' headers (Figure 11.4).

11.3.2.3 Routers

Similar to switches, routers forward packets from source to destination. However, routers forward packets based on the global routing header (GRH). They modify the LRH while forwarding from one subnet to another subnet. Each subnet is identified with a subnet prefix. Routers exchange routing information using protocols specified by IPv6 (Figure 11.5).

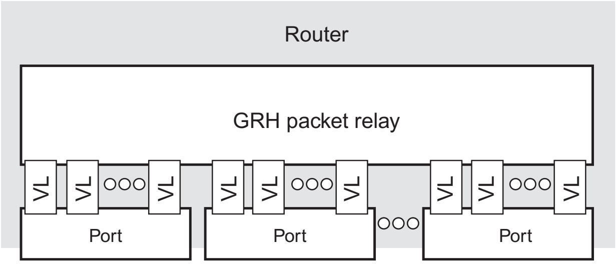


FIGURE 11.5

IB router.

The source node specifies the LID of the router and global identifier (GID) of the destination that packet is being sent to. Each router forwards packets to the next router using subnet information and routing tables. Routing is performed until the packet reaches the destination subnet. The last router forwards packets to the final destination using the local ID associated with the destination GID.

11.3.2.4 Subnet manager

The SM configures local components in subnet. It provides LIDs to all the nodes in the subnet, and it also provides forwarding information to switches in the subnet. SMs communicate to all the nodes within the subnet via subnet management agents (SMAs). Each IB node is required to implement SMA.

There can be multiple SMs in a subnet—but only one can be active at a given time. All the inactive/backup SMs maintain a copy of an active SM's forwarding information and use it to continue to provide management services to the subnet if the active SM goes down.

11.4 Communication mechanisms

11.4.1 Communication services (transport)

IBA supports different types of communication mechanisms between IB nodes based on the needs of the application.

11.4.1.1 Reliable connection and reliable datagram

For reliable communication, data is delivered reliably through a combination of sequence numbers and acknowledgment messages (ACK/NAK). Upon detecting an error or loss of packet, the source can recover by retransmitting the packet without involvement from the user application. This mode guarantees the delivery of a message packet exactly once. When applications need to rely on the underlying transport to guarantee delivery of messages to its destination, this mode is used.

This mode frees up an application to safeguard against the unreliability of the underlying media or delivery mechanisms.

Reliable connection (RC) mode provides reliable data transfer between nodes using a direct dedicated connection between the source and destination end nodes.

Reliable datagram (RD) mode provides reliable packet message delivery to any end node without a dedicated connection between the source and destination end nodes. This is an optional mode.

11.4.1.2 Unreliable datagram and unreliable connection

Unreliable modes are useful for applications that are not sensitive to packet loss or that are capable of handling the packet loss themselves, but desire fast data transmission. In this mode, transmission of data from the source node to the destination end node is not guaranteed.

In unreliable datagram (UD) mode, data can be sent from the source node to the destination end node without any connection establishment. Packet delivery is not guaranteed. In this mode, data loss is not detected.

In unreliable connection (UC) mode, a dedicated connection is established between the source and destination end nodes, and message transfer is carried out without transmission guarantee. Errors (including sequence errors) are detected and logged and are not informed back to the source end node.

11.4.1.3 Raw IPv6 datagram and Raw Ethertype datagram

This is special mode of UD in which only local transport header information is used. This mode is used by non-IBA transport layers to tunnel data across IB networks.

11.4.2 Addressing

IB communication among nodes requires unique identification for each addressable entity (node, card, port, queue pair (QP) within a port, etc.) so that packets can be delivered appropriately. Such packets could be part of communication within a subnet or they could belong to flows that cross subnets through a router. Flows could be unicast where communication is between exactly two addressable entities. Multicast flows are used for communication between multiple entities (Figure 11.6).

LID: A LID is a 16-bit unicast or multicast identifier and is unique within a subnet; it cannot be used to route between subnets. A LID is assigned by an SM. LIDs are contained within the LRH of each packet.

GID: A GID is a 128-bit unicast or multicast address and is unique globally—which allows it to be used for routing packets across subnets. A GID is a valid IPv6 address with additional restrictions defined by the IBA. GID assignment ranges from default assignment (calculated from the manufacturer-assigned identifier) to an address assigned by the SM.

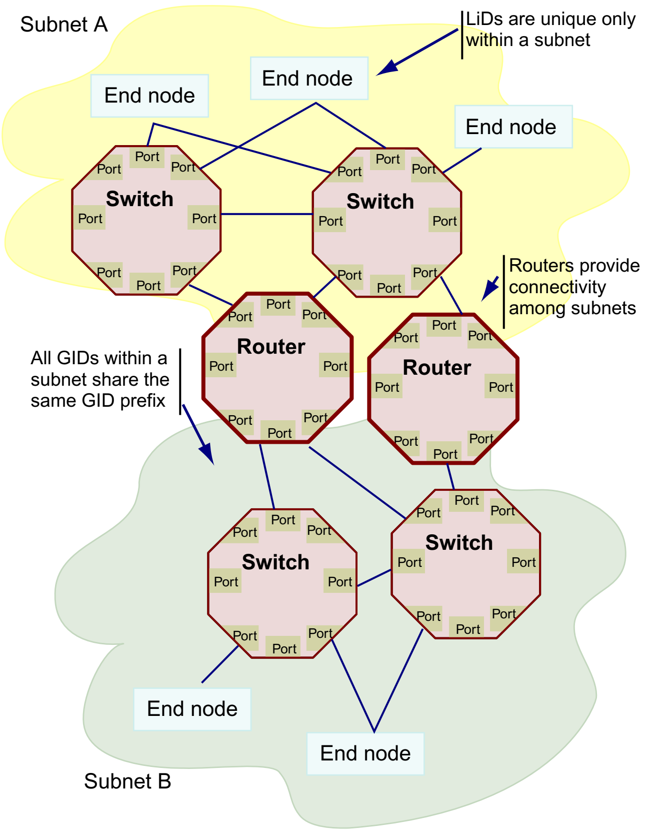

Subnet A
LiDs are unique only within a subnet
End node
End node
End node
Port
Switch
Switch
Routers provide connectivity among subnets
All GIDs within a subnet share the same GID prefix
Router
Router
Switch
Switch
End node
End node
Subnet B

FIGURE 11.6

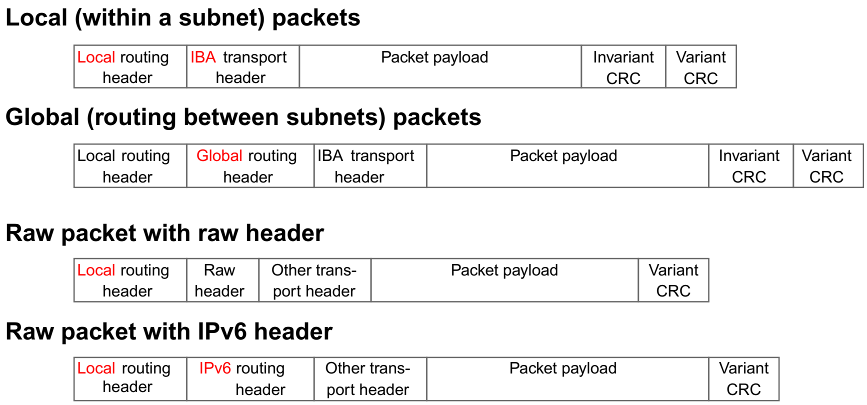
IB addressing scope.

Unicast identifier: A unicast LID or GID identifies a single addressable entity. A packet addressed to the unicast identifier will be delivered to a single end port.
Multicast identifier: A multicast LID or GID identifies a set of addressable end ports. A packet sent to a multicast identifier must be delivered to all the end ports that are part of that identifier.

11.4.3 Packet formats

There are two categories of packets that are defined in IB networks.

IBA packets: IB packets that carry transport headers are routed on IBA fabrics and use native IBA transport facilities.
Raw packets: These packets are typically used for transferring non-IBA packets over an IB network. So these packets do not contain IBA transport headers.



Local (within a subnet) packets

Local routing header	IBA transport header	Packet payload	Invariant CRC	Variant CRC

Global (routing between subnets) packets

Local routing header	Global routing header	IBA transport header	Packet payload	Invariant CRC	Variant CRC

Raw packet with raw header

Local routing header	Raw header	Other transport header	Packet payload	Variant CRC

Raw packet with IPv6 header

Local routing header	IPv6 routing header	Other transport header	Packet payload	Variant CRC

FIGURE 11.7

IB packet formats.

The packet formats defined by IBA are illustrated in Figure 11.7.

All the packets require a local route header (8 bytes). This header is used for forwarding the packets within a local subnet [11].

A GRH (40 bytes) is required for all the packets that need to be routed to a different subnet and on all multicast packets. The link next header field in the LRH indicates the presence of a GRH header in a packet.

Raw packets contain only an LRH and a Raw or IPv6 routing header.

IBA supports two types of CRC (cyclic redundancy check) in the packets. Invariant CRC (4 bytes) covers all the fields that do not change as a packet moves through the fabric from the source to the destination end node. Variant CRC (VCRC) (2 bytes) covers all the fields in the packet. Each IBA packet carries invariant CRC followed by VCRC. Raw packets carry only VCRC.

11.5 Layered architecture

Layered architecture allows solutions to be built with different components that can interoperate with flexibility of implementation, yet with correctness of operation due to well-defined interfaces between layers. It also provides architectural clarity and separation for different functional blocks in systems, which helps applications to communicate with each other over a variety of protocols, networks, and physical connectivity options. This provides a variety of possible application deployments without requiring top-to-bottom change to the implementation. One can run a system over copper or fiber cable with a change in physical media without requiring a change in any of the protocols above the physical layer. Similarly, one could run the application within a subnet or across multiple subnets without the application being aware of network separation between the communicating systems (Figure 11.8).

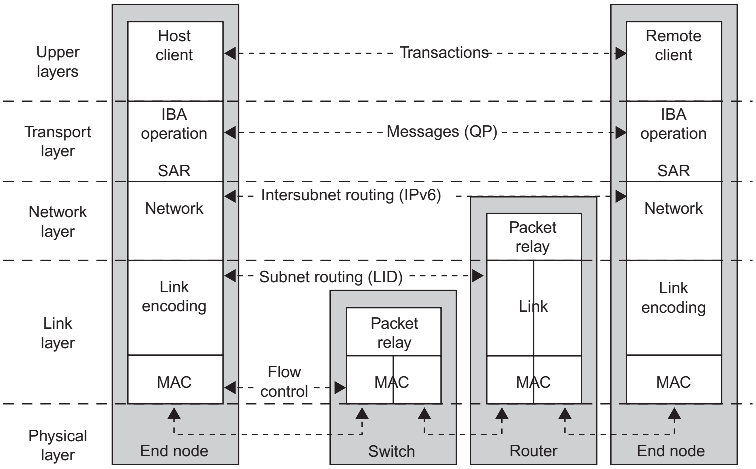


FIGURE 11.8

IB layered architecture.

IBA establishes a layered architecture across five layers. The top layer provides an interface to applications while the bottom layer defines physical connectivity for systems with each other.

11.5.1 Physical layer

The physical layer defines how actual bits flow on physical connectivity between systems. Systems can be connected to each other over backplane or over fiber or copper media. The length of cable can vary, as can the mechanical connector with which they are connected to the systems. The physical layer for IB defines the electrical as well as mechanical aspects of such connectivity.

The physical layer is responsible for receiving (control and data) bytes from the link layer, sending them in electrical or optical form to the link peer, and then delivering the received bytes to the link layer at the receiver. It provides transparency to the link layer about the actual physical media that connects to the link peer.

The physical layer is also responsible for speed and width negotiation for the underlying physical media with the link peer.

IBA defines four types of physical connectivity options to connect IB devices. It provides electrical, optical, and mechanical specifications for all of these. The following lists the physical connectivity options:

1. Backplane port
2. Copper port
3. Fiber port
4. Active cable port

11.5.1.1 Packet formats and link widths

Packets are delimited by special symbols on the wire called SDL (start of data packet delimiter) and EGP (end of good packet delimiter) or EBP (end of bad packet delimiter). A link is formed with multiple "lanes" of connectivity between two nodes. Ports with a single lane are called 1X ports. Similarly, ports with 4 lanes or 12 lanes are called 4X or 12X, respectively.

For ports with multiple lanes, packets are "byte-striped" across all the lanes (Figure 11.9).

11.5.1.2 Speed and width negotiation

IBA operation of the link at different speeds, as new generations have evolved to run links at higher speeds. In SDR (single data rate), signaling is at 2.5 Gbps. In the following generations, speeds have increased to 5 Gbps (DDR), 10 Gbps (QDR), etc. Figure 11.10 shows the succession of speeds for the IB link layer (with future speeds projected as well) [12].

11.5.2 Link layer

The link layer in the IB architecture defines mechanisms for sending and receiving packets across physical connections. These mechanisms include addressing, buffering, flow control, QoS, error detection, and switching. Addressing was discussed in more detail in Section 11.4.2.

11.5.2.1 Packet forwarding

The link layer defines forwarding of packets within an IBA subnet.

Within a subnet, packets are forwarded using LIDs. Figure 11.11 shows the format for an LRH. These identifiers are configured at a device by the SM as described in section 11.3.2.4. Switches use the destination LID to look up the destination port to which a given packet needs to be forwarded.

IBA requires that in-order packet delivery is maintained within unicast packets in a flow (packets between the same source and destination LIDs within a subnet).

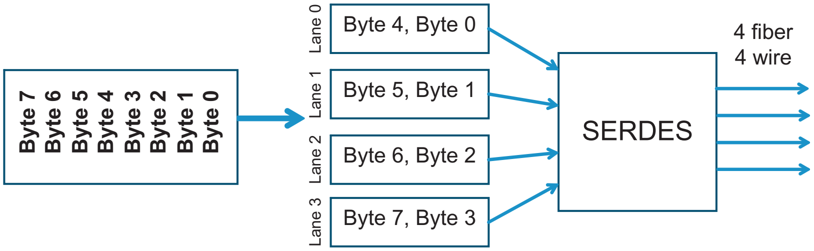
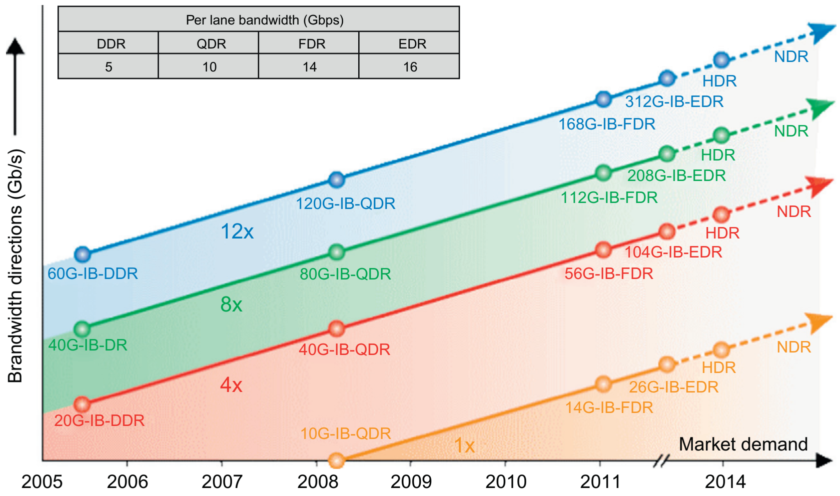


FIGURE 11.9

Byte-striping data across lanes.



SDR – Single data rate
DDR – Double data rate
QDR – Quad data rate
FDR – Fourteen data rate
EDR – Enhanced data rate
HDR – High data rate
NDR – Next data rate

FIGURE 11.10

IBTA signaling rate roadmap.

<table>
<tr><th>Bits
Bytes</th><th colspan="3">31-24</th><th colspan="3">23-16</th><th>15-8</th><th>7-0</th></tr>
<tr><td>0–3</td><td>VL</td><td colspan="2">LVer</td><td>SL</td><td>Rsv2</td><td>LNH</td><td colspan="2">Destination local identifier</td></tr>
<tr><td>4–7</td><td colspan="2">Reserve 5</td><td colspan="4">Packet length (11 bits)</td><td colspan="2">Source local identifier</td></tr>
</table>

FIGURE 11.11

Local routing header (LRH).

The same applies to multicast and broadcast packets. However, the in-order requirement does not apply between unicast and multicast packets received at a node.

11.5.2.2 Data integrity

IBA guarantees the integrity of data as it flows through the network using a CRC field.

A 32-bit invariant CRC covers all the fields in the packet that are invariant from end to end. This field does not require recalculation at each hop. This field is not present in raw packets.

Fields that vary in the network are covered by a 16-bit VCRC. The VCRC is calculated at each hop to catch any data integrity compromise that may have caused modification of headers during routing. This field is present in all packets, including raw packets.

11.5.2.3 Virtual lanes

IB allows the creation of virtual links over the, physical connection of wires with an abstraction called VLs. Each VL can operate independently of each other as far as mechanisms like link-level flow control are concerned. This allows traffic differentiation of different "flows" on a given physical port or a physical link. Flows can be grouped to belong to a particular "VL," and consistent traffic engineering discipline can be applied across multiple devices to provide a certain QoS to given VL.

Each packet carries information about which VL it belongs to in a 4-bit VL field in the link header.

Each VL is provided independent buffering resources guaranteeing that they do not interfere with each other's operation. Link-level flow control for each VL assures that flows in one VL will not be affected by flow control asserted on another VL.

IBA defines 15 VLs on each link. VL15 is considered higher in priority as compared to VL0. Higher-priority traffic can be serviced more promptly as compared to lower-priority flows. Bandwidth/resources can be allocated to flows according to their VL discipline. Management traffic uses the highest-priority VL, VL15. Each device is required to support VL0 and VL15.

11.5.2.4 Service level

In addition to VLs, IBA also defines a mechanism to assign a QoS identifier to flows—service level (SL). This 4-bit field is included in the LRH and it identifies SL for a given flow within a subnet. This field in the packet does not get modified as the packet traverses the network. The actual meaning of SL is left for implementation—however, it is intended to be used by products to provide traffic differentiation for flows as dictated by the SM.

IBA defines mechanisms to map the SL field to a VL for a given port. This SL to VL mapping is achieved through a mapping table, and it allows IB nodes to provide QoS for flows according to the discipline defined through the mapping table and inherent expectation of VL assignment. For example, an SL mapping to a higher-priority VL gets higher priority at a given port.

Since each SL gets different scheduling on a given port, ordering is not maintained between different SLs.

11.5.2.5 Buffering and flow control

IBA provides a mechanism to guarantee lossless transport of packets across a link through a buffer-to-buffer credit-based flow control mechanism. This requires each receiver to provide information about the availability of a buffer to the transmitter so that the transmitter can deliver the packet to the receiver on the wire. Since there is a reserved buffer waiting for the arriving packet, there is no scenario for the receiver to drop a packet due to congestion.

Each VL is required to have separate buffering on a given port. This allows the use of separate flow control on each VL.

IBA defines a mechanism for the receiver to inform the transmitter about the amount of data it is allowed to transmit at a given point in time. IBA also specifies protocols to ensure that the communication protocol to exchange information about flow control itself is error-free and can recover from error conditions, if they arise. Transmitters and receivers resynchronize their information periodically to correct any inconsistency in information about credit availability on a given VL.

In addition to link-level flow control, IBA also specifies congestion control mechanisms that allow a congested port in a network to request the actual source of a flow to slow down (as compared to just transmitting the port of a previous node in link-level flow control). IBA specifies a mechanism in which congestion on a VL can be detected and forward notification of congestion (FECN: forward explicit congestion notification) is marked by a switch for the offending packet. This bit is interpreted by destination and turned around as a special management packet for backward notification (BECN: backward explicit congestion notification) to the source of the offending flow. The source then interprets this packet and reduces the injection rate of data to the given congested destination temporarily (the original injection rate resumes over time).

11.5.3 Network layer

The network layer provides routing across multiple subnets. It specifies forwarding of unicast and multicast packets across IBA subnets. Such routing can be accomplished by routers conforming to IBA as well as non-IBA (e.g., IP) specifications.

The fields provided in the GRH in Figure 11.12 can be used for such routing. Typically, these fields include SGID, DGID, TClass, and flow label (these could easily be mapped into the IPv6 vocabulary, and this is intentional). Source Global Identifier (SGID) and Destination Global Identifier (DGID) are 128-bit fields that can be mapped to a IPv6 addresses. Routing works very similar to IP routing where the Destination Local Identifier (DLID) within a source subnet will be

<table>
<tr><th>Bits
Bytes</th><th colspan="2">31-24</th><th colspan="2">23-16</th><th colspan="2">15-8</th><th colspan="2">7-0</th></tr>
<tr><td>0–3</td><td>IPVer</td><td colspan="2">TClass</td><td colspan="5">Flow label</td></tr>
<tr><td>4–7</td><td colspan="4">PayLen</td><td colspan="2">NxtHdr</td><td colspan="2">HopLmt</td></tr>
<tr><td>8–11</td><td colspan="8">SGID[127–96]</td></tr>
<tr><td>12–15</td><td colspan="8">SGID[95–64]</td></tr>
<tr><td>16–19</td><td colspan="8">SIGID[63–32]</td></tr>
<tr><td>20–23</td><td colspan="8">SIGID[31–0]</td></tr>
<tr><td>24–27</td><td colspan="8">DGID[127–96]</td></tr>
<tr><td>28–31</td><td colspan="8">DGID[95–64]</td></tr>
<tr><td>32–35</td><td colspan="8">DGID[63–32]</td></tr>
<tr><td>36–39</td><td colspan="8">DGID[31–0]</td></tr>
</table>

FIGURE 11.12

Global routing header (GRH).

mapped to a local router address, and the destination router will make sure that the packet is delivered to the destination node by changing the DLID of the packet to the final destination port in that subnet.

11.5.4 Transport layer

The transport layer provides an interface for upper layer protocols (ULPs) (and applications) to communicate within and across subnets over network layer using a QP for send and receive operations. It is responsible for delivering a data payload from the source end node to the destination end node using the delivery characteristics desired by the application (e.g., reliable versus unreliable and connection versus datagram). The transport layer delivers packets to the right QP based on the information in the transport header.

The transport layer is also responsible for providing segmentation and reassembly services to ULPs. It segments consumer data in the transmit path into the right-sized payload based on the maximum transfer unit supported by the underlying network layer. Each segment is encapsulated with headers and CRC during transmission. Upon reception, a QP reassembles all the segments in a specified ULP buffer in memory.

Actual transport of data and its delivery is dependent on the type of service a given QP is configured with. Details about these mechanisms are discussed in Section 11.5.1.

11.6 RDMA over converged Ethernet (RoCE)

11.6.1 Overview (DCB and RoCE)

There has been an increasing desire for converging different types of fabrics, adapters to allow for reduction in overall TCO (total cost of ownership). Instead of running separate networks for LAN, SAN (storage), and IPC (low latency), there are a lot of benefits in running all these protocols on a single physical infrastructure. With this in mind, Ethernet standards have been enhanced to support different types of networks (Figure 11.13).

IEEE defined new enhancements to Ethernet that allow application of “lossless” characteristics to a L2 network—this enhanced Ethernet is called DCB (data center bridging).

With DCB networks, one can get “lossless” characteristics in an Ethernet L2 network that are similar to an IB network. Although the mechanisms used to achieve this “lossless” behavior are different between these two technologies, for all practical purposes, they achieve similar results of delivering a packet across a link in lossless fashion (avoiding a drop in case of congestion).

Since the IB protocol was designed to operate over lossless Layer 2 connectivity, DCB provides the required functionality in Ethernet to carry IB packets.

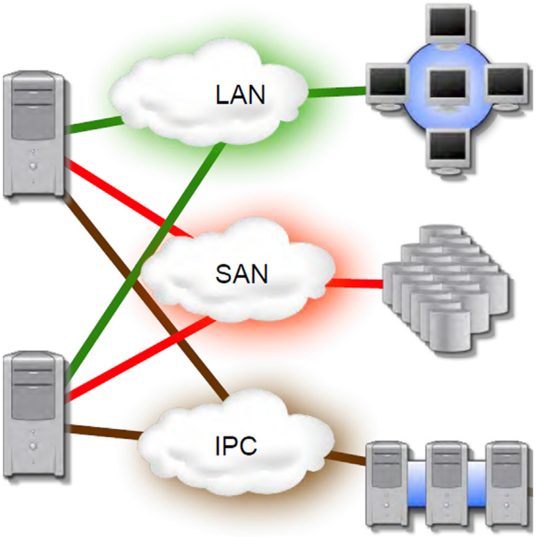


FIGURE 11.13

I/O convergence.

RoCE is a protocol that tunnels IB packets over Ethernet. It maintains most of the layers intact and uses the Ethernet L2 layer as the physical and link layers [9] (Figure 11.14).

11.6.2 Layer architecture

As can be seen from Figure 11.14, RoCE maintains all the layers except the link, MAC, and physical layers.

RoCE achieves the following goals through this modification:

1. Uses DCB (lossless Ethernet) as Layer 2 network to provide physical connectivity.
2. No change to applications that are using RDMA as the ULP interface is maintained unchanged.
3. Maintains existing IB transport constructs and services (RC, UC, RD, etc.) (Figure 11.15).

11.6.3 Packet formats

RoCE tunnels most of the IB packet into an Ethernet packet [13]. The Ethernet header provides similar functionality to the IB LRH. It allows Ethernet nodes to communicate with each other in a given subnet. So the RoCE packet does not include the LRH in the tunneled Ethernet packet. LRH fields are mapped into equivalent Ethernet header fields.

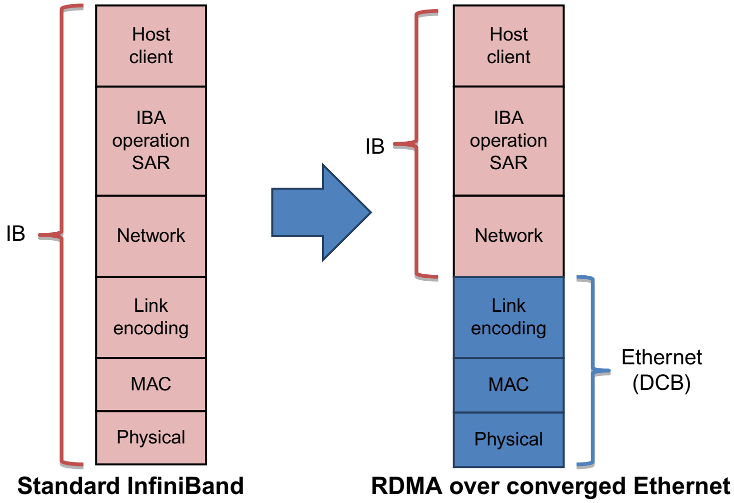


FIGURE 11.14

Layer comparison between IB and RoCE.

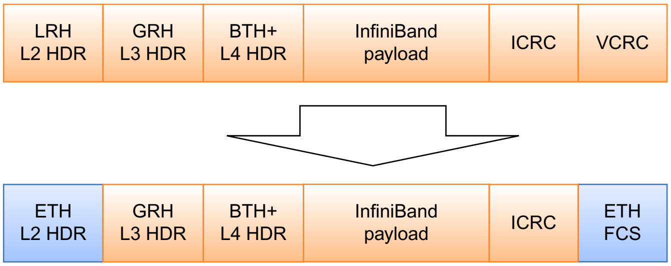


FIGURE 11.15

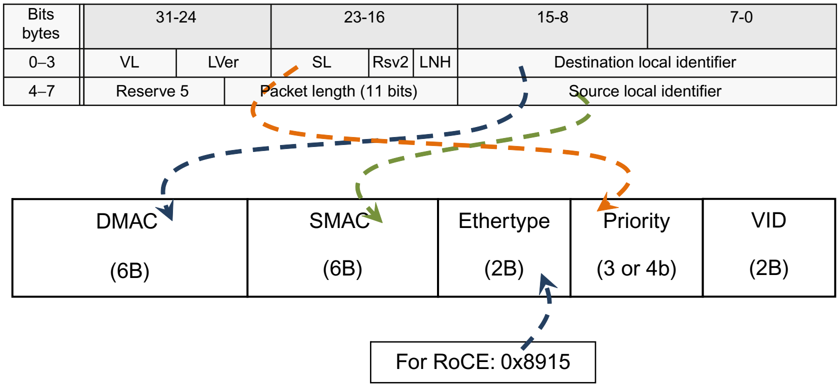
IB and RoCE packet format comparison.

Since the Ethernet packet is covered with Frame Check Sequence (FCS), VCRC from the IB packet is not required in RoCE packets. The remaining fields in an IB packet are carried intact in an RoCE packet.

11.6.4 Header and address mapping

Ethernet has a similar header structure as to the IB LRH (Figure 11.16).

As the IB packet gets mapped into the Ethernet format, LRH fields are replaced with the Ethernet L2 header. DLID and SLID are replaced with 6-byte Ethernet MAC addresses. (IB allows information about subnet LIDs to be



Bits bytes	31-24		23-16			15-8	7-0
0–3	VL	LVer	SL	Rsv2	LNH	Destination local identifier	
4–7	Reserve 5		Packet length (11 bits)			Source local identifier	

FIGURE 11.16

IB header mapping for RoCE.

accessed by ULPs through a verb interface. Since RoCE does not carry LRH, these LIDs are not carried through the interface.)

MAC addresses are generated and used through normal Ethernet methods (each end point has its assigned MAC address). The association of a GID to a MAC address is left to implementation (through well-known mechanisms similar to ARP, neighbor discovery, etc.)

Since there is no LRH in a RoCE packet, Raw services are not supported in RoCE. (As can be seen in Section 11.5.3, Raw services do not carry the GRH and other IBA headers and rely on LRH headers; hence, they cannot be supported in RoCE).

SLs are represented in a priority/drop eligibility field in the VLAN header. Since there are eight priority values for 16 SLs the, 0—7 SL values are mapped directly to the 0—7 priority values. SL the values 8—15 are reserved in RoCE.

The Ethernet header does not have a field like VL that identifies local resources (e.g., queues) at each node. The Ethernet standard allows mapping of priority values to local queues (called TC, the traffic class) through programmatic interface; however, it does not have a mechanism to provide such mapping on each flow/packet. Thus, this mapping needs to be achieved for RoCE through an out-of-band mechanism.

RoCE has assigned Ethertype (0 × 8915), which identifies RoCE packets on an Ethernet link.

11.6.5 Ethernet fabric requirement

The RoCE specification does not expressly require DCB or "lossless" Ethernet—however for comparative performance/features, it is expected that RoCE will be used only with DCB-compliant Ethernet switches.

IEEE 802.1 defined the following standards for providing converged traffic over Ethernet in 2011.

1. IEEE 802.1Qbb [14]: Priority-based Flow Control (PFC)
   a. PFC allows selective flow control of traffic flows identified with particular priority bit in Ethernet header [7]
   b. Provides no-drop behavior required for Fibre Channel over Ethernet (FCoE) and RoCE
2. IEEE 802.1Qaz: Enhanced Transmission Selection (ETS)
   a. ETS provides for bandwidth allocation to traffic classes; an alternative to strict priority
   b. DCBX uses Link Layer Discovery Protocol (LLDP) to coordinate configuration of DCB features across links
3. IEEE 802.1Qau: Congestion notification
   a. Allows a congestion point to notify the traffic source (reaction point) of congestion

Although most of the implementations are expected to move to these standards, current implementations in the market follow prestandard multivendor agreement specifications for #1 and #2 above [1–3].

11.7 8 iWARP

11.7.1 Overview

Internet Wide Area RDMA Protocol enables usage of the RDMA protocol over TCP/IP in an Ethernet environment. Specifications for iWARP are standardized by the IETF (Internet Engineering Task Force [1,4–6]). iWARP provides a similar verb interface to ULPs as IB and RoCE.

11.7.2 Layer architecture

IB and iWARP are defined by different standards bodies, but both of them address similar network needs and provide a similar verb interface to applications. Figure 11.17 shows an approximate comparison of the layered architectures of IB and iWARP.

Figure 11.18 shows the layers for iWARP as defined in IETF specifications.

iWARP uses Ethernet for local routing within a subnet, IP as a networking layer to route traffic across subnets, and TCP as a transport layer to provide reliable and connection-oriented packet delivery across the network. iWARP also supports Stream Control Transmission Protocol (SCTP) as another alternative, as a transport layer for Remote Direct Memory Access Protocol (RDMAP). The primary difference between TCP and SCTP is that TCP is a streaming protocol (converts a message into a stream of bytes) and SCTP is a message-oriented protocol. Both

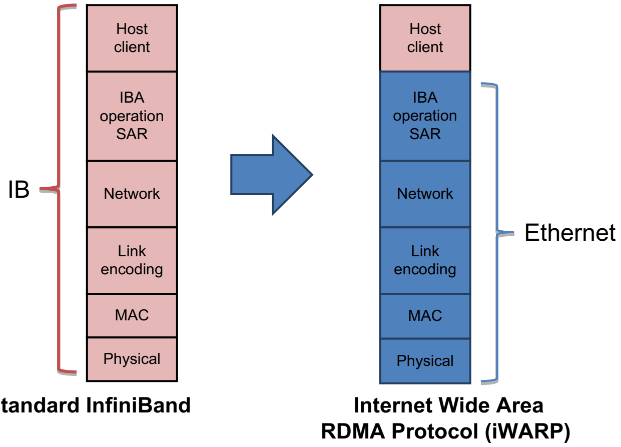


FIGURE 11.17

Comparison between IB and iWARP.

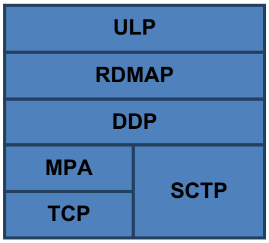


FIGURE 11.18

iWARP layer architecture.

protocols run over IP and are friendly with each other with regard to their congestion management mechanisms. iWARP provides RC for RDMA over both the TCP and SCTP layers. The most dominant usage of iWARP in industry currently uses TCP as the transport layer—so the remainder of the section will focus on TCP.

Since iWARP uses TCP as the transport layer, it does not have dependency on the underlying Ethernet fabric to be lossless. And since IP is routable across the Internet, iWARP is routable in a data center deployment (RoCE, eventhough it uses the Ethernet fabric, is not routable across IP subnets in the Ethernet data centers).

11.7.3 Packet formats

11.7.3.1 RDMAP

RDMAP is an interface for RDMA applications to carry data over an underlying iWARP infrastructure. It uses underlying layer direct data replacement (DDP) to

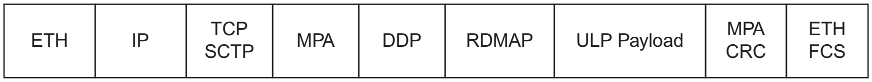


FIGURE 11.19

iWARP packet headers.

enable the applications to read and write into a remote node's memory (RDMA) (Figure 11.19).

11.7.3.2 DDP

DDP can move data directly into a destination node's (data sink) memory without requiring the network interface to copy data into an intermediate buffer. This layer provides the following functionality:

- Tagged buffer model: Ability to name buffers and share that information with peers (this enables placing data directly into the destination node's memory)
- Untagged buffer model: Also allows data transfer to anonymous buffers at the data sink
- Reliable, in-order delivery
- Segmentation and reassembly of ULP messages; can handle out-of-order segments without requiring an additional copy

11.7.3.3 MPA (marker PDU aligned framing for TCP)

TCP transfers a message across the network by creating segments that carry a stream of bytes. In order to identify the boundaries of a message in the given stream of bytes, MPA has been defined. It places boundary identifiers (markers) in the TCP stream to allow a receiver to identify the boundaries of the given message.

MPA is not required when running DDP over SCTP since SCTP is a message-oriented protocol. MPA includes an additional CRC check to increase data integrity when running over TCP.

References

[1] R. Recio, B. Metzler, P. Culley, J. Hilland, D. Garcia. A Remote Direct Memory Access Protocol Specification [Online]. Available from: <http://tools.ietf.org/html/rfc5040>.

[2] CEE Specifications—DCBX, CEE Specifications—DCBX [Online]. Available from: <http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf>.

[3] CEE Specifications—ETS [Online]. Available from: <http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf>.

[4] H. Shah, J. Pinkerton, R. Recio, P. Culley. Direct Data Placement over Reliable Transports [Online]. Available from: <http://tools.ietf.org/html/rfc5041>.

[5] C. Bestler, R. Stewart, Stream Control Transmission Protocol (SCTP) Direct Data Placement (DDP) Adaptation [Online]. Available from: <http://tools.ietf.org/html/rfc5043>.

[6] P. Culley, U. Elzur, R. Recio, S. Bailey, J. Carrier, Marker PDU Aligned Framing for TCP Specification [Online]. Available from: <http://tools.ietf.org/html/rfc5044>.

[7] CEE Specifications—PFC, CEE Specifications—PFC XE "PFC: Priority-based Flow Control" [Online]. Available from: <http://www.ieee802.org/1/files/public/docs2008/bb-pelissier-pfc-proposal-0508.pdf>.

[8] IBM, HPC Clusters Using InfiniBand on IBM Power Systems [Online]. Available: <http://www.redbooks.ibm.com/redbooks/pdfs/sg247767.pdf>.

[9] RDMA over Converged Ethernet: Supplement to IB Architecture Specification Volume 1 Release 1.2.1 [Online]. Available from: <http://www.inifinibandta.org>.

[10] InfiniBand Architecture Specification Volume 1 & 2 [Online]. Available from: <http://www.infinibandta.org>.

[11] G.F. Pfister [Online]. Available from: <http://gridbus.csse.unimelb.edu.au/~raj/superstorage/chap42.pdf>.

[12] C. DeCusatis, Handbook of Fiber Optic Data Communication: A Practical Guide to Optical Networking, third ed., Academic Press, 2008.

[13] A. Ayoub, RDMA over Converged Ethernet (RoCE), September 9, 2011 [Online]. Available from: http://www.itc23.com/fileadmin/ITC23_files/slides/WDC_3_RoCE-DC-CaVeS-9Sep2011-nb.pdf.

[14] IEEE 802.1Q, "Data Center Bridging WG" [Online]. Available from: <http://www.ieee802.org/1/pages/dcbridges.html>.