Using InfiniBand for high performance computing

technology brief, 2nd edition

Abstract

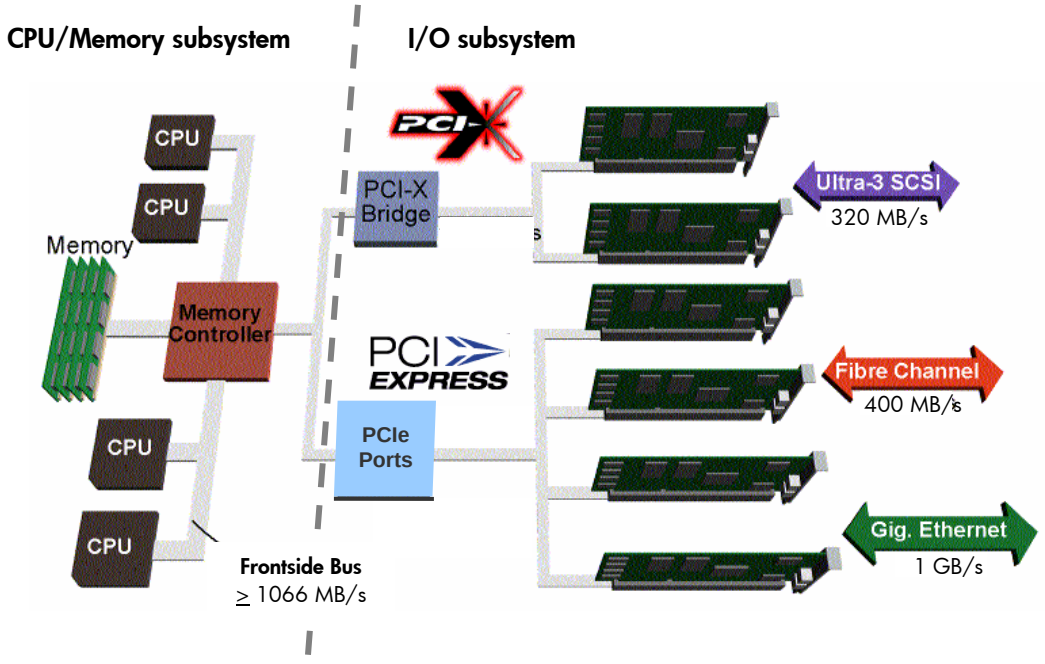
With business models constantly changing to keep pace with today's Internet-based, global economy, IT organizations are continually challenged to provide customers with high performance platforms while controlling their cost. With broader adoption of high performance computing (HPC) in various industry segments, more enterprise businesses are implementing parallel compute cluster architectures to provide a cost-effective approach for scalable, HPC platforms.

InfiniBand is one of the most important technologies that enable the adoption of cluster computing. This technology brief describes InfiniBand as an interconnect technology used in cluster computing, provides basic technical information, and explains the advantages of implementing the InfiniBand architecture.

Introduction

The overall performance of enterprise servers is determined by the synergetic relationship between three main subsystems: processing, memory, and input/output (Figure 1). Using multiple processing cores sharing common memory space, the multiprocessor architecture of a single server provides a high degree of parallel processing capability.

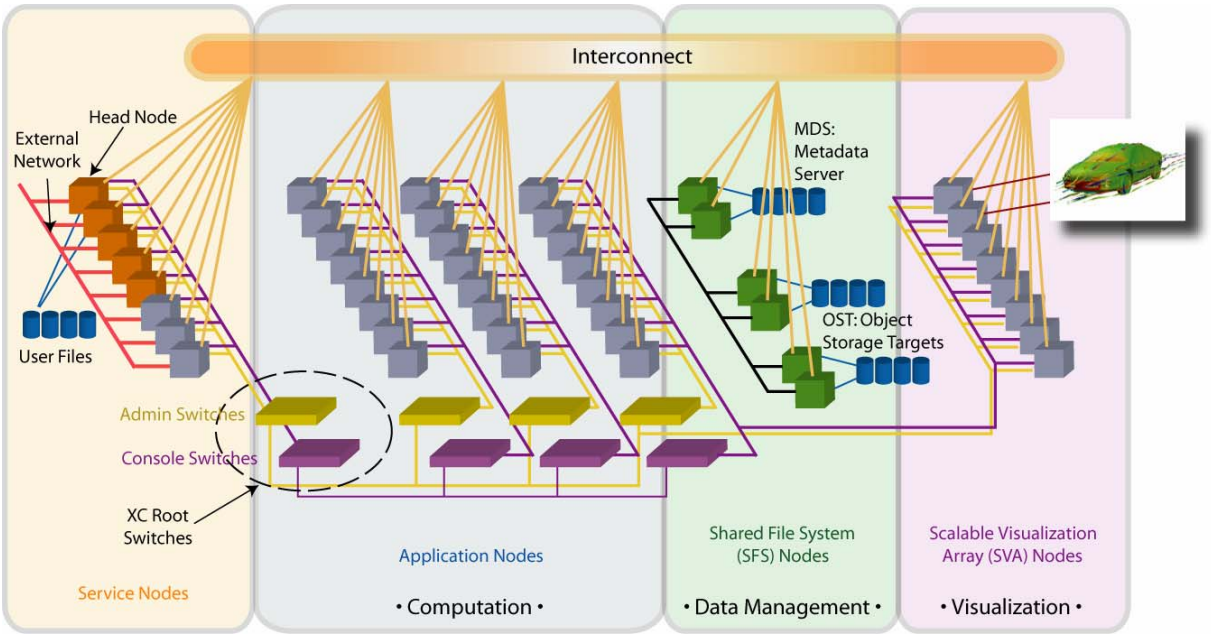
Figure 1. Single server (node) architecture



However, multiprocessor server architecture cannot scale cost effectively to a large number of processing cores. Cluster computing that builds an entire system by connecting stand-alone systems with interconnect technology has become widely implemented at the HPC and enterprise data centers around the world.

Figure 2 shows an example of cluster architecture that integrates computing, storage, and visualization functions into a single system. Applications are usually distributed to compute nodes through job scheduling tools.

Figure 2. Sample clustering architecture



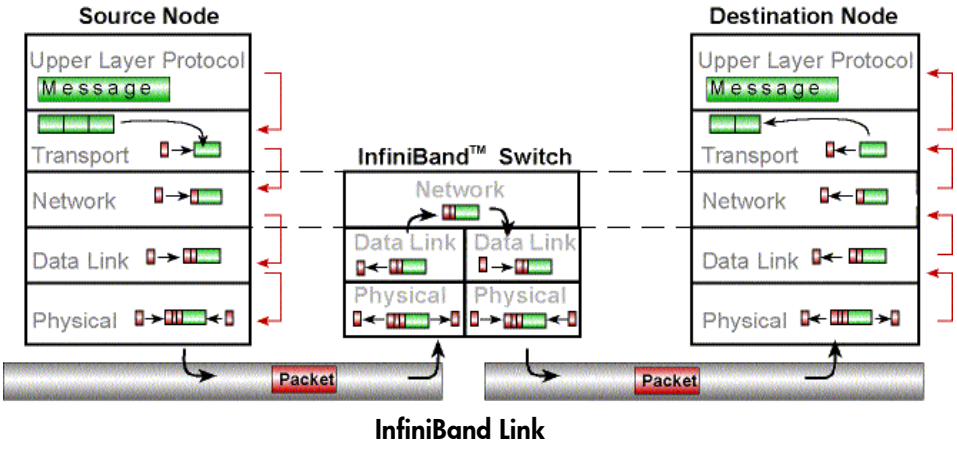
Clustered systems allow infrastructure architects to meet performance and reliability goals, but interconnect performance, scalability, and cost are key areas that must be carefully considered. A cluster infrastructure works best when built with an interconnect technology that scales easily and economically with system expansion.

InfiniBand technology

InfiniBand (IB) is an industry-standard, channel-based architecture that features high-speed, low-latency interconnects for cluster computing infrastructures. InfiniBand uses a multiple-layer architecture to transfer data from one node to another. In the InfiniBand layer model (Figure 3), separate layers perform different tasks in the message passing process.

The upper layer protocol (ULP) layer works closest to the operating system and application; it defines how much software overhead will be required by the data transfer. The InfiniBand transport layer is responsible for communication between applications. Each message can be up to 2 GB in size. The transport layer splits the messages into data payloads and encapsulates each data payload and an identifier of the destination node into one or more packets. Packets can contain data payloads of up to four kilobytes, although one to two kilobytes is typical depending on the IB adapter and type of transport protocol being used. The packets are passed to the network layer, which selects a route to the destination node and attaches the route information to the packets. The data link layer attaches a local identifier (LID) to the packet for communication at the subnet level. The physical layer transforms the packet into an electromagnetic signal based on the type of network media—copper or fiber.

Figure 3. Distributed computing using InfiniBand architecture



NOTE:
While InfiniBand infrastructures usually include the use of switches, direct host-channel-adaptor to host-channel-adapter (HCA-to-HCA) operation is supported in some implementations.

InfiniBand offers key advantages including:

- Increased bandwidth with double data rate (DDR) at 20 Gbps available now and Quad Data Rate (QDR) in the future
- Low latency end-to-end communication
- Hardware-based protocol handling, resulting in faster throughput due to efficient message passing and memory-efficient data transfers such as RDMA

InfiniBand components

InfiniBand architecture involves four key components:

- Host channel adapter
- Subnet manager
- Target channel adapter
- InfiniBand switch

A host node or server requires a host channel adapter (HCA) to connect to an InfiniBand infrastructure. An HCA can be a card installed in an expansion slot or integrated onto the host's system board. An HCA can communicate directly with another HCA, with a target channel adapter, or with an InfiniBand switch.

InfiniBand uses subnet manager (SM) software to manage the InfiniBand fabric as well as to provide monitoring services for interconnect performance and health at the fabric level. A fabric can be as simple as a point-to-point connection or multiple connections through one or more switches. The SM software resides on a node or switch within the fabric, and provides switching and configuration information to all of the switches in the fabric. Additional backup SMs can be located within the fabric for failover should the primary SM fail. All other nodes in the fabric will contain an SM agent that processes management data. Managers and agents communicate using management datagrams (MADs).

A target channel adapter (TCA) is used to connect an external device (storage unit or I/O interface) to an InfiniBand infrastructure. The TCA includes an I/O controller specific to the device's protocol (SCSI, Fibre Channel, Ethernet, etc.) and can communicate with an HCA or an InfiniBand switch.

An InfiniBand switch provides scalability to an InfiniBand infrastructure by allowing a number of HCAs, TCAs, and other IB switches to connect to an InfiniBand infrastructure. The switch handles network traffic by checking the local link header of each data packet received and forwarding the packet to the proper destination.

The most basic InfiniBand infrastructure will consist of host nodes or servers equipped with HCAs and subnet manager software. More expansive networks will include multiple switches.

InfiniBand software architecture

Traditional Ethernet communication uses a layered network protocol stack provided by the operating system. To improve overall system performance, two Ethernet developments have occurred:

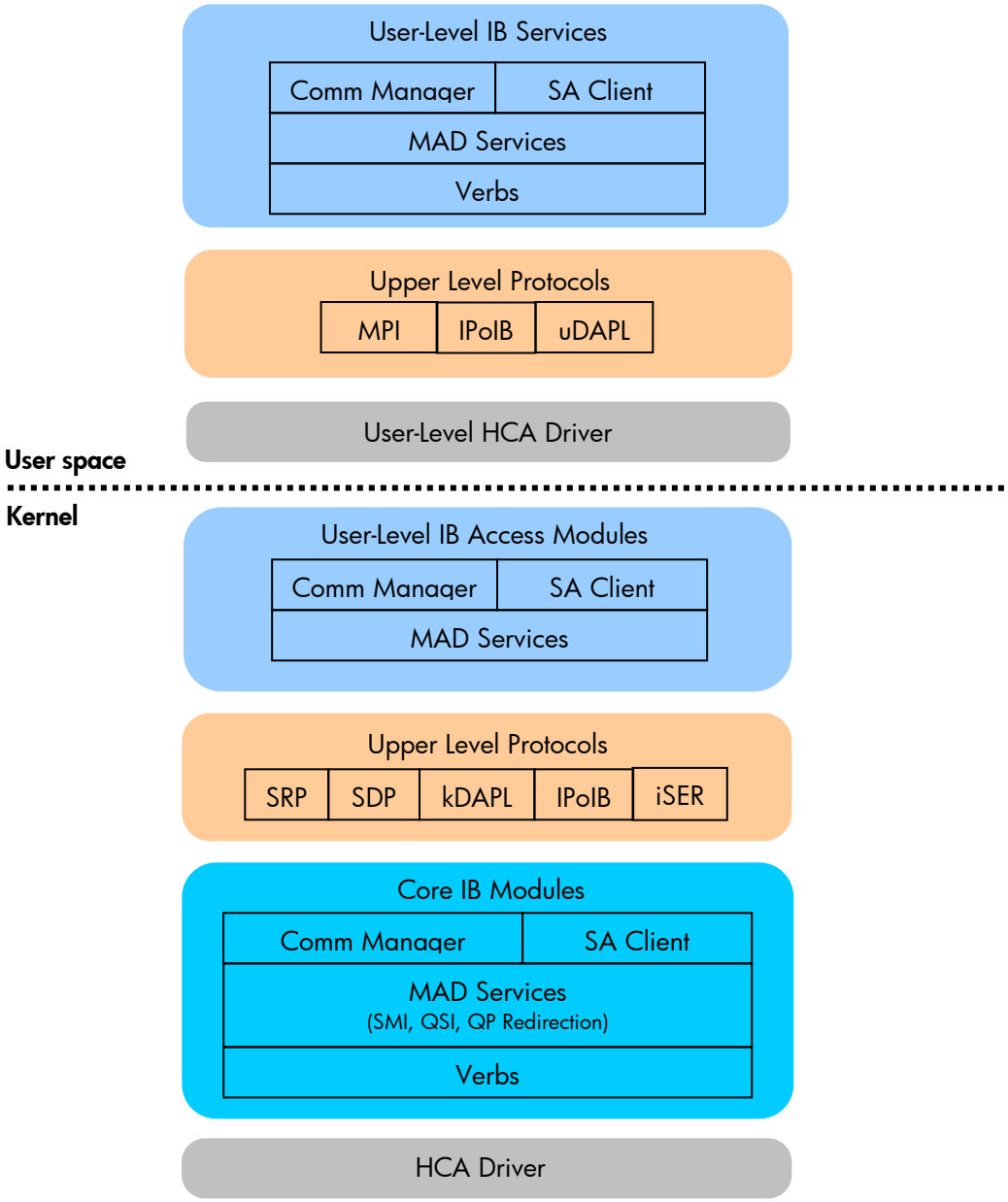
- TCP/IP offload engine (TOE), a NIC hardware enhancement that assumes TCP/IP processing duties
- RDMA over TCP/IP, a protocol enhancement that provides more efficient data transfers between nodes

InfiniBand, like Ethernet, uses a multi-layer processing stack to transfer data between nodes. However, InfiniBand architecture provides OS-bypass features such as the communication processing duties and RDMA operations as core capabilities and offers greater adaptability through a variety of services and protocols.

While the majority of existing InfiniBand clusters operate on the Linux® platform, drivers and HCA stacks are also available for Microsoft® Windows®, HP-UX, Solaris, and other operating systems from various InfiniBand hardware and software vendors.

The layered software architecture of the HCA allows code to be written without specific hardware in mind. The functionality of an HCA is defined by its verb set, which is a table of commands used by the application programming interface (API) of the operating system being run. A number of services and software protocols are available (Figure 4); and, depending on type, they may be implemented from user space or from the kernel.

Figure 4. InfiniBand HCA software layers



As indicated in Figure 4, InfiniBand supports a variety of upper level protocols (ULPs) that have evolved since the introduction of InfiniBand. These protocols are described in the following sections.

MPI

The message passing interface (MPI) protocol is a library of calls used by applications in a parallel computing environment to communicate between nodes. MPI calls are optimized for performance in a compute cluster that takes advantage of high-bandwidth and low-latency interconnects. In parallel computing environments, code is executed across multiple nodes simultaneously. MPI is used to facilitate the communication and synchronization among these jobs across the entire cluster.

To take advantage of the features of MPI, an application must be written and compiled to include the libraries from the particular MPI implementation used. There are several implementations of MPI on the market:

- HP-MPI
- Intel MPI
- Publicly available versions such as MVAPICH2 and Open MPI

MPI has become the de-facto IB ULP standard, and HP-MPI in particular has been broadly accepted by independent software vendors (ISVs) to eliminate the complexity of developing and running applications in parallel environments. By using shared libraries, applications built on HP-MPI can transparently select interconnects that significantly reduce the efforts for applications to support various popular interconnect technologies. HP-MPI is supported on HP-UX, Linux, True64 UNIX, and Microsoft Windows Compute Cluster Server 2003.

IPoIB

Internet Protocol over InfiniBand (IPoIB) provides support for all IP-based protocols by allowing the use of any TCP/IP-based application over the InfiniBand media. IPoIB does not support the RDMA features of InfiniBand. IPoIB supports IPv4 or IPv6 protocols and addressing schemes. In the operating system, an InfiniBand HCA is configured as a traditional network adapter and can use all of the standard IP-based applications such as PING, FTP, and TELNET.

DAPL

The Direct Access Programming Library (DAPL) allows low-latency RDMA communications between nodes. The uDAPL provides user-level access to RDMA functionality on InfiniBand. Applications need to be written with a specific uDAPL implementation to use RDMA for data transfers between nodes. This interface is being replaced in the open source world by the RDMA API.

SDP

Sockets Direct Protocol (SDP) is an industry standard that allows stream socket applications to run transparently over an RDMA interconnect (iWARP or InfiniBand).

SRP

SCSI RDMA Protocol (SRP) is a data movement protocol.

iSER

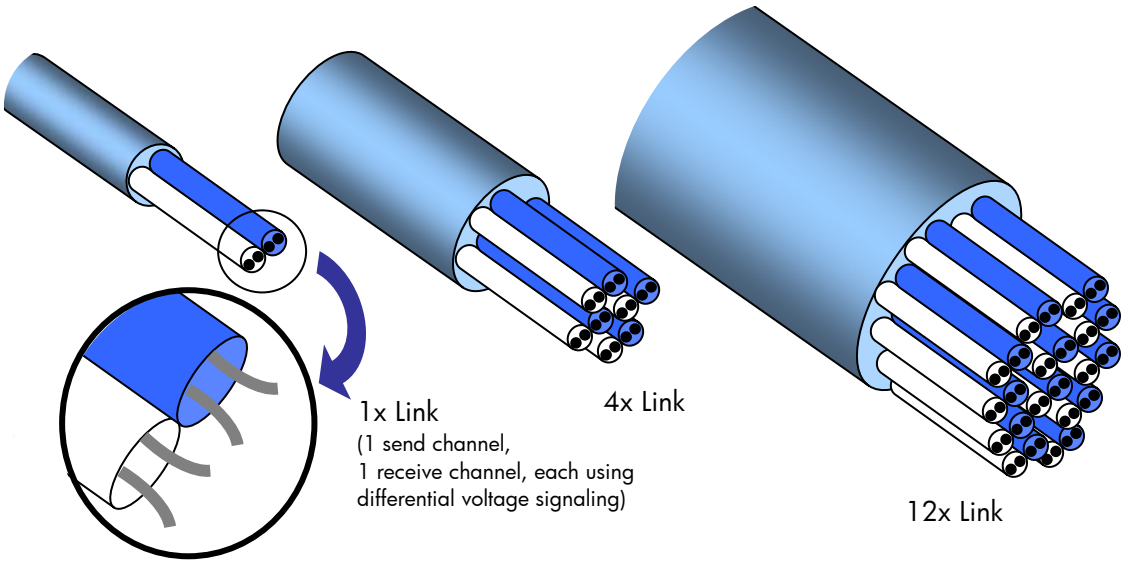
iSCSI Enhanced RDMA (iSER) is a storage standard originally specified on the iWARP RDMA technology and now officially supported on InfiniBand. The iSER protocol provides iSCSI manageability to RDMA storage operations.

InfiniBand physical architecture

InfiniBand fabrics use high-speed, bi-directional serial links between devices. The bi-directional links contain dedicated send and receive lanes for dual-simplex operation. For single data rate (SDR) operation, each lane has a signaling rate of 2.5 Gbps. Bandwidth is increased by adding to the number of lanes per link and using double data rate (DDR) operation.

InfiniBand interconnect types include 1-, 4-, or 12-wide full-duplex links (Figure 5), providing theoretical full-duplex SDR bandwidths of 5 Gbps, 20 Gbps, and 60 Gbps, respectively.

Figure 5. InfiniBand link types



Encoding overhead in the data transmission process limits the maximum data bandwidth per link to approximately 80 percent of the signal rate. However, the switched fabric design of InfiniBand allows bandwidth to grow or aggregate as links and nodes are added. Double data rate (DDR) operation also increases bandwidth significantly (Table 1). Quad data rate (QDR) operation, which may be available in 2008, provides a 10-Gbps signal rate *per lane*.

Table 1. InfiniBand bandwidth

Link	SDR half duplex signal rate	SDR half duplex usable bandwidth [1]	DDR half duplex signal rate	DDR half duplex usable bandwidth [1]
1x	2.5 Gbps	2 Gbps (250 MBps)	5 Gbps	4 Gbps (500 MBps)
4x	10 Gbps	8 Gbps (1 GBps)	20 Gbps	16 Gbps (2 GBps)
12x	30 Gbps	24 Gbps (3 GBps)	60 Gbps	48 Gbps (6 GBps)

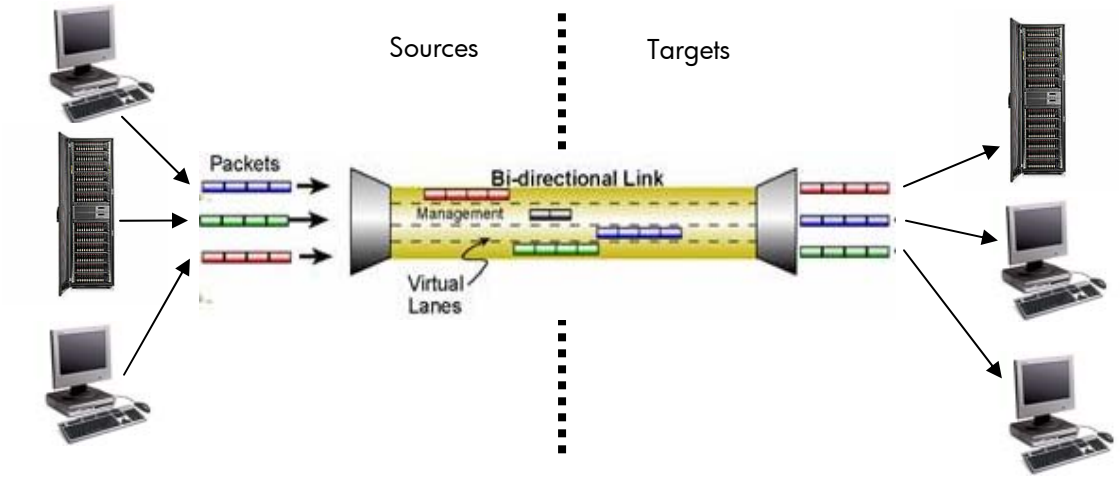
NOTE: [1] Usable bandwidth may actually be less due to protocol overhead, which can vary depending on packet size.

Most InfiniBand deployments have been on 4X SDR products, but the adoption of 4X DDR products is increasing rapidly. To accommodate the various physical designs, the InfiniBand specification defines both copper and fiber optic links. Current operating distances for copper cable is up to 17 meters at the 4X SDR and about 10 meters at 4X DDR. The fiber optic links can be up to 300 meters at SDR and about 100 meters at 4X DDR.

Link operation

Each link can be divided (multiplexed) into a set of virtual lanes, similar to highway lanes (Figure 6). Each virtual lane provides flow control and allows a pair of devices to communicate autonomously. Each link accommodates a minimum of two and a maximum of sixteen virtual lanes. One lane is reserved for fabric management and the other lane(s) are used for packet transport. The virtual lane design allows an InfiniBand link to share bandwidth between various sources and targets simultaneously. For example, if a 10-Gb/s link were divided into five virtual lanes, each lane would have a bandwidth of 2 Gb/s. The InfiniBand architecture defines a virtual lane mapping algorithm to ensure inter-operability between end nodes that support different numbers of virtual lanes.

Figure 6. InfiniBand virtual lane operation



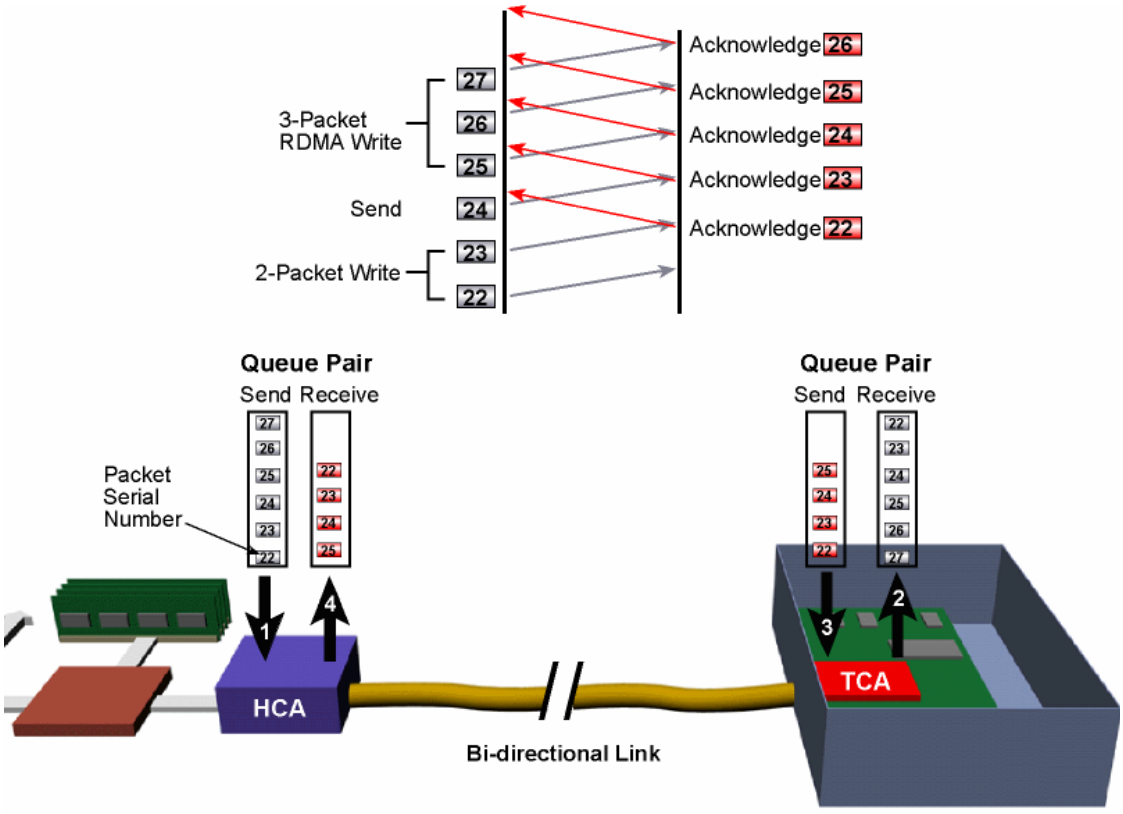
When a connection between two channel adapters is established, one of the following transport layer communications protocols is selected:

- Reliable connection (RC) – data transfer between two entities using receive acknowledgment
- Unreliable connection (UC) – same as RC but without acknowledgement (rarely used)
- Reliable datagram (RD) – data transfer using RD channel between RD domains
- Unreliable datagram (UD) – data transfer without acknowledgement
- Raw packets (RP) – transfer of datagram messages that are not interpreted

All protocols can be implemented in hardware; some protocols are more efficient than others. The UD and Raw protocols, for instance, are basic datagram movers and may require system processor support depending on the ULP used.

During reliable connection operation (Figure 7), hardware at the source generates packet sequence numbers for every packet sent; and the hardware at the destination checks the sequence numbers and generates acknowledgments for every packet sequence number received. The hardware also detects missing packets, rejects duplicate packets, and provides recovery services for failures in the fabric.

Figure 7. Link operation using reliable connection protocol



The programming model for the InfiniBand transport assumes that an application accesses at least one Send and one Receive queue to initiate the I/O. The transport layer can handle four types of data transfers for the Send queue:

- Send/Receive – typical operation where one node sends a message and another node receives the message.
- RDMA Write – operation where one node writes data directly into a memory buffer of a remote node
- RDMA Read – operation where one node reads data directly from a memory buffer of a remote node
- RDMA Atomics – allows atomic update of a memory location from an HCA perspective

The only operation available for the receive queue is Post Receive Buffer transfer, which identifies a buffer that a client may send to or receive from using a Send or RDMA Write data transfer.

InfiniBand summary

Networks using TCP/IP (such as Ethernet) have adapted zero-copy (RDMA) protocols into their communications protocol stack to enhance performance. RDMA is a core capability of InfiniBand architecture. Flow control support is native to the HCA design, and the latency time for InfiniBand data transfers is generally less—typically 3 to 4 microseconds for MPI pingpong latency—than that for 10Gb Ethernet.

InfiniBand provides structured communications between nodes allowing simultaneous connections between multiple node pairs. This gives the fabric the ability to scale or aggregate bandwidth as more nodes and/or additional links are connected to it.

Parallel compute applications that involve a high degree of message passing between nodes benefit significantly from DDR IB. HP BladeSystem c-Class clusters and similar rack-mounted clusters support IB DDR HCAs and switches.

InfiniBand is further strengthened with HP-MPI becoming the leading solution among ISVs for developing and running MPI-based applications across multiple platforms and interconnect types. Software development and support becomes simplified since interconnects from a variety of vendors can be supported by an application written to the HP-MPI protocol.

InfiniBand and HP BladeSystem c-Class products

In the past few years, cluster computing has become a mainstream architecture for high performance computing. As the technology becomes more mature and affordable, clusters have been deployed in practically every HPC vertical. The trend in this industry is toward using space- and power-efficient blade systems. HP BladeSystem c-Class solutions offer significant savings in power, cooling, and data center floor space without compromising performance. Figure 8 shows some of the HP c-Class products[1].

Figure 8. HP BladeSystem c-Class server products

HP c7000 enclosure with blades

HP BL460c Server Blade

HP 4x DDR IB Interconnect Switch

HP 4x DDR IB HCA mezzanine card

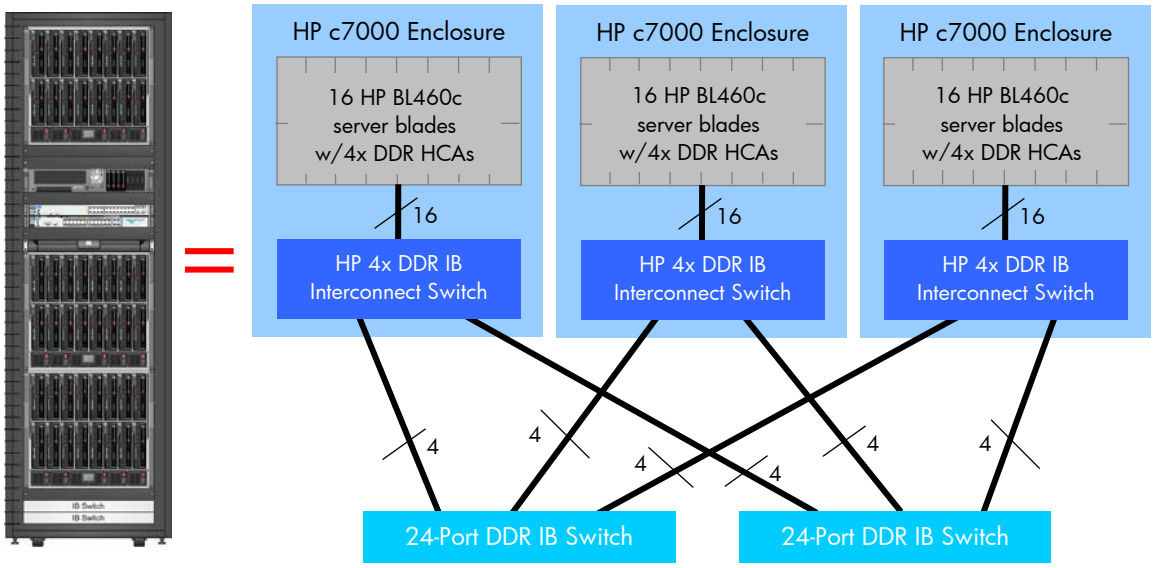
The c7000 enclosure supports up to 16 half-height or 8 full-height server blades and includes rear mounting bays for management and interconnect components. Each server blade includes both PCIe x8 and x4 mezzanine connectors for I/O options such as the HP 4x DDR IB mezzanine card. The HP 4x DDR IB mezzanine card features a memory-free design economical in both cost and power consumption.

[1] HP BladeSystem c-Class solutions include comprehensive offerings in server blades, storage blade, networking products, software, services, and data center solutions. For more details on HP BladeSystem c-Class, please visit www.hp.com/go/blades.

NOTE:
The HCA mezzanine card should be installed in a PCIe x8 connector for maximum InfiniBand performance.

Figure 9 shows a rack configured with HP c-Class components supporting 48 nodes in a cluster configuration. HP manufactures, markets, and supports HPC cluster products under its Unified Cluster Portfolio branding, with Cluster Platforms 3000BL and 4000BL being built from the HP BladeSystem c-Class product family.

Figure 9. HP c-Class 48-node cluster configuration



As an increasingly-adapted industry-standard interconnect, InfiniBand will continue to evolve and develop. Revisions to the InfiniBand specification have included quad-signaling rates that will extend bandwidth to 40 Gbps in each direction on a 4x link, raising InfiniBand performance to new levels.

Future InfiniBand products are likely to include IB storage boxes, lower-cost HCAs, and motherboards with embedded IB interfaces. Off-the-shelf application support from software vendors will increase as InfiniBand continues to gain acceptance into mainstream IT.

Conclusion

The decision of using Ethernet or InfiniBand should be based on the interconnect performance requirement and cost consideration. InfiniBand leads in performance in both bandwidth and latency, and in general has a lower cost than the 10Gb Ethernet products that have started coming to market. Vendors also have aggressive roadmaps to roll out quad data rate InfiniBand products in the future.

The HP Unified Cluster Portfolio includes a range of hardware, software, and services that provide customers a choice of pre-tested, pre-configured systems for simplified implementation. HP Cluster Platforms are built around specific hardware and software platforms and offer a choice of interconnects. For example, the HP Cluster Platform CL3000BL cluster platform uses the HP BL460c blade server as the compute node with a choice of GbE or InfiniBand interconnects.

No longer relegated to Linux or HP-UX environments, HPC clustering is now supported through Microsoft Windows Compute Cluster Server 2003, with native support for HP-MPI. HP is committed to supporting both InfiniBand and Ethernet infrastructures, and to helping customers choose the most cost-effective fabric interconnect solution for their needs.

For more information

For more information, refer to the resources listed below:

Subject	Hyperlink to resources
HP products	www.hp.com
HPC/IB/cluster products	www.hp.com/go/hptc
HP InfiniBand products	http://h18004.www1.hp.com/products/servers/networking/index-ib.html
InfiniBand Trade Organization	http://www.infinibandta.org
Open InfiniBand Alliance	http://www.openib.org/
RDMA Consortium	http://www.rdmaconsortium.org
Technology brief discussing iWARP RDMA	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00589475/c00589475.pdf

Call to action

Send comments about this paper to TechCom@HP.com.

TC070101TB, January 2007