High Speed I/O Server Computing with InfiniBand

José Luís Gonçalves

Dep. Informática, Universidade do Minho
4710 - 057 Braga, Portugal
zeluis@ipb.pt

Abstract: High-speed server computing heavily relies on low latency and high bandwidth interconnection technologies and fast I/O channels at each node. InfiniBand Architecture is one of the emerging technologies in this area. A brief analysis of the architecture key features is presented. Concurrent approaches and technologies are mentioned.

1 Introduction

Over the past decade, tremendous technical advances have been made in high volume servers for business computing. However, one part of the server —the I/O (input/output) subsystem — was left behind, and is the weak link in the server architecture.

The InfiniBand Trade Association (IBTA) is the driving force behind the development of InfiniBand Architecture (IBA). This association was founded in September 1999 with the merger of two other associations: The Future I/O (founded by HP, Compaq and IBM) and The Next Generation I/O forum (founded by Intel and integrating Dell, Sun and some others key players of the computer industry). Today IBTA has more than 220 companies working in the technology specification. If maturity of a proposed standard specification is measured by weight of the specification, InfiniBand is a very mature specification. This is a very large, complicated specification requiring nearly 2,000 pages. The first specification of the architecture was released in September 23rd 2000.

Some of those companies have announced products with this technology, but apart from some prototypes, there are no products available to the end user.

This communication overviews the key features of IBA, points out some of the most innovative features and briefly compares it to concurrent technologies [1], [2].

2 InfiniBand Architecture Overview

Computer evolution continues to follow Moore's law in what regards processor's speed: every 18 months or so, processors clock speed doubles. In the meanwhile bus systems continue to evolve very slowly compared to processors. InfiniBand intends to level that situation.

IBA is directed to high speed I/O needs, bus systems cannot provide the scalability, performance, reliability and availability required by most high performance systems.

IBA defines a switched, point-to-point communications system, commonly addressed as a fabric. This approach differs from the one used by other architectures such as Fibre Channel or iSCSI in a key feature which is that IBA implements the communication processing and management in hardware right next to the system memory subsystem instead of being implemented by a device attached to the system bus controlled by the system main processor.

IBA differs from other technologies in the way that faces the problem of the bottleneck that I/O represents in the high performance computing scenario. This approach brings I/O right to the core of the computing system [3].

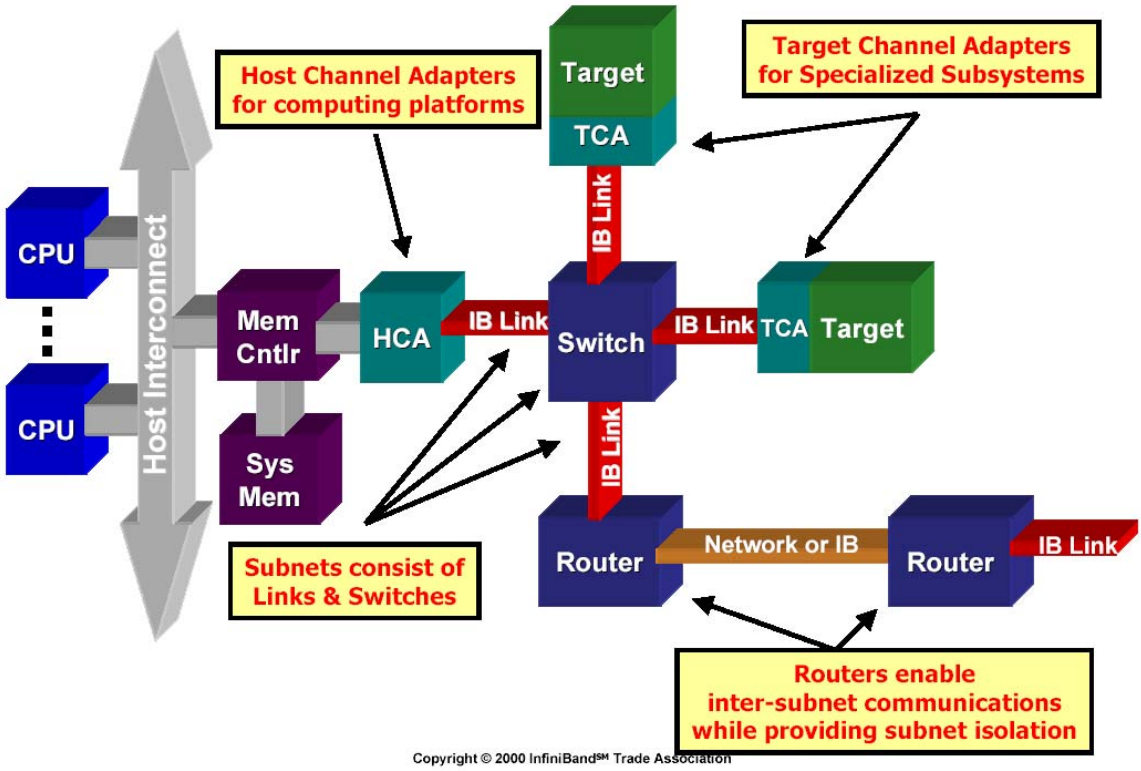

Fig. 1. InfiniBand architecture Schema.

In the above picture we can see the layout of the architecture and some of its key elements. IBA works by connecting Host Channel Adapters (HCA) to Target Channel Adapters (TCA). HCA are usually located right next to the system's CPU and memory, while the TCA are usually attached to storage and peripherals devices. The switch is located between TCA and the HCA to direct traffic between the HCA and different TCA. The router connects the all the InfiniBand fabric to other InfiniBand subnets, LAN/WAN networks or other devices.

2.1 IB link

IB links represent the connections between InfiniBand components, creating this way the InfiniBand fabric.

2.2 Host Channel Adapter (HCA)

This acts like a terminator/concentrator to the InfiniBand links. The representation seen in the picture is just from illustrations proposes, IBA specifications do not explicitly specify how the HCA should be attached inside the server. The HCA resides in a host processor node, connects the node to the IBA fabric, and acts as the interface to processes that require I/O services from devices connected to the InfiniBand fabric.

2.3 Target Channel Adapter (TCA)

The TCA will be the end point of the IBA, the device that will connect the IB link to the target device. The TCA resides in an I/O node. It connects the node to the IBA fabric and acts as the interface to an I/O device. The TCA includes an I/O controller that is specific to its particular I/O device protocol, such as SCSI, Fibre Channel (FC), or Ethernet.

2.4 Switch

The switch handles intrasubnet traffic, passing packets across the fabric from one end node to another within the subnet, based on the packet destination addresses. A subnet manager configures switches with forwarding tables that are used in conjunction with packet destination addresses to ensure that the packet is routed correctly.

2.5 Routers

This interconnects different IBA subnets, providing both scalability and isolation to all InfiniBand fabric. Connections between IBA routers may be made by InfiniBand link or other mechanism such as MAN or WAN. With the encapsulation of InfiniBand over IP provided by InfiniBand routers it is possible to connect InfiniBand Fabrics in different locations.

3 InfiniBand Layer Structure

As all modern communication architectures InfiniBand implements all the seven OSI model layers. At the physical layer, IBA defines both electrical and mechanical characteristics for the system. These include cables and receptacles for fibre and copper media, backplane connectors and hot swap characteristics are also defined. Each copper link consists of 2 wires in each direction, each link has a 2.5 Gbaud signalling rate.

The distance between links in not defined, although an attenuation limit of 15 db is imposed. As a result copper links should not be longer than 17m and optical links longer than 10Km.

The link and transport layers are the key of the InfiniBand Architecture. The link Layer key features are its packet layout, point-to-point link operations, and switching within a local subnet. There are two types of packets within the link layer, management and data packets. Management packets are used for link configuration and maintenance. Data packets carry up to 4k bytes of a transaction payload.

All devices within a subnet have a 16 bit Local ID (LID) assigned by the Subnet Manager.

QoS is supported by InfiniBand through Virtual Lanes (VL). These VLs are separate logical communication links, which share a single physical link. Each link can support up to 15 standard VLs and one management lane (VL 15).

VL15 is the highest priority and VL0 is the lowest. As a packet traverses the subnet, a Service Level (SL) is defined to ensure its QoS level. Each link along a path can have a different VL, and the SL provides each link a desired priority of communication. Each switch/router has a SL to VL mapping table that is set by the subnet manager to keep the

proper priority with the number of VLs supported on each link. Therefore, the IBA can ensure end-to-end QoS through switches and routers.

Credit based flow control is used to manage data flow between two point-to-point links. Flow control is handled on a per VL basis allowing separate virtual fabrics to maintain communication utilizing the same physical media. Each receiving end of a link supplies credits to the sending device on the link to specify the amount of data that can be received without loss of data.

At the transport layer IBA defines both reliable and unreliable communication mechanisms. For reliable communications IBA defines that the sequencing and retransmission should be implemented in hardware. A key based mechanism is used to identify InfiniBand links, so multiple operating systems and network topologies can coexist in the same IBA fabric. There is a significant improvement that the IBA offers for the transport layer: all functions are implemented in hardware.

IBA defines the routing requirements of IBA equipment; the vendor provides algorithms and management tools. IBA also defines interfaces between hardware and software; these interfaces are called verbs.

Verbs are neither a software API nor Hardware Adaptation layer. Verbs are mixture of both. IBA specifications leave some latitude to vendors who wish to design and manufacturer IBA equipment. To guarantee compliance between different equipment from different manufactures IBTA defined verbs to influence/guide design of hardware software and interfaces.

4 InfiniBand Native Protocols

4.1 Remote Direct Memory Access (RDMA).

InfiniBand is largely based on a concept known as Remote DMA. RDMA provides a way to reliably communicate between servers—as well as server to I/O—without the need for heavy protocols like TCP/IP. RDMA is built into the lowest levels of network interfaces that can read and write data directly into the memory subsystem of a server or I/O device without the need for a high overhead network protocol driver to verify integrity and demultiplex messages to applications. Instead, messages are moved directly into or out of the memory that is owned by an application. Also, communications reliability is built directly into the underlying network protocol. Because of RDMA, InfiniBand allows servers and applications to reliably transfer small messages across the network with extremely low latency, and large bulk data transfers with extremely high throughput, with minimal burdens on the application servers.

Low latency and high bandwidth allow large mission-critical databases to be deployed on commodity platforms—for a fraction of the current cost. Smaller enterprises also finally gain access to enterprise-class database technology that was previously out of reach due to cost.

4.2 Direct Access File System (DAFS)

DAFS provides Local file sharing, using memory-to-memory (RDMA) mechanisms, gaining high throughput and extremely low-overhead access to shared data. DAFS offloads file system processing and meta-data I/O from the application/database servers and eliminates protocol-processing overhead, while preserving the advantages of file access.

Using protocols like DAFS, network attached storage (NAS) could have the latency of block storage (SAN), or even better with lower TCO usually associated with file based storage compared to block storage. It remains to be seen if DAFS and alike will change the NAS/SAN equilibrium.

4.3 Socket Direct Protocol (SDP)

SDP maps standard socket (Winsock, BSD, etc') API's to InfiniBand architecture, keeping the same API's towards the application in a way that the application still thinks it is working over a TCP/IP based network, but with much greater performance and lower latency.

SDP is used for IPC between InfiniBand nodes and for session based networking with elements outside the InfiniBand fabric residing in the LAN/WAN. This efficient connectivity is achieved by using Intelligent Routers that terminate TCP/IP traffic translating it to SDP.

SDP works in conjunction with the IP over IB standard, which enables routing and non TCP Traffic (ICMP, UDP) to traverse between the nodes.

4.4 IP over InfiniBand (IPoIB)

IPoIB is an IETF standard being developed, which defines the mapping of IP traffic over InfiniBand messaging mechanisms. It enables applications and routers to use InfiniBand like an Ethernet medium. It defines topics like ARP, Multicasting, and MIB's mainly using unreliable InfiniBand transport. It can be used alone or in conjunction with SDP.

A few interesting advancements are planned in the future enabling higher availability, performance, QoS by using some of InfiniBand special mechanisms.

5 Differences between IBA and Other Interconnection Technologies

There are numerous standards today in interconnection technologies, each one of them with unique features. Examples of some of those technologies are PCI-X, Fibre Channel, 3GIO, Gigabit Ethernet, and RapidIO.

The next table briefly compares some of the features supported in hardware by InfiniBand and by other interconnecting hardware [4].

6 InfiniBand and High Performance Embedded Computing (HPEC)

As originally designed, InfiniBand was primarily cables connecting boxes. However, the standards have evolved so that a backplane and a packaging definition have been defined as well. [5]

Feature	InfiniBand™	PCI-X	Fibre Channel	1Gb & 10Gb Ethernet	Hyper-Transport™	Rapid I/O
Bus/Link Bandwidth	2.5/10/30Gb/s[a]	8.51 Gb/s	1/2.1 Gb/s[b]	1 Gb, 10Gb	12.8, 25.6, 51.2 Gb/s[f]	16/32 Gb/s[c]
Bus/Link Bandwidth (Fully Duplexed)	5/20/60Gb/s[a]	Half-Duplex	2.1/4.2 Gb/s[b]	2 Gb, 20Gb	25.6, 51.2, 102 Gb/s[f]	32/64 Gb/s[c]
Pin Count	4/16/48[d]	90	4	4, Fiber	55,103,197[f]	40/76[c]
Maximum Signal Length	Km	Inches	Km	Km	Inches	Inches
Transport Media	PCB, Fiber and Copper Cable	PCB only	Copper and Fiber Cable	Copper and Fiber Cable	PCB only	PCB only
Simultaneous Peer to Peer communication	15 VLs + Mngt Lane			X		3 Transaction Flows
Native Hwd Transport Support with Memory Protection	X					
In-Band Management	X		Uses out-of-band mngt	Not native, can use IP		
RDMA Support	X					
Native Virtual Interface Support	X					
End-to-End Flow Control	X			X	X	X
Memory Partitioning[e]	X		X			
Quality of Service	X		X	Limited		X
Reliable	X		X		X[f]	X
Scalable	X		X	X	X	X
Maximum Packet Payload	4 KB	Not Packet Based	2 KB	1.5KB (10Gb no jumbo support)	64 bytes	256 bytes

Fig. 2. This table illustrates de difference between some of the most popular interconnection technologies including InfiniBand.

6.1 High Bandwidth

The most obvious benefit of InfiniBand is bandwidth and the scalability of the bandwidth available. With debits ranging from 2.5 Gbps to 60 Gbps and with the possibility of putting together several links, bandwidth is no longer a restrain in HPEC systems.

6.2 Latency

An equally important parameter in this type of systems is latency. There are a number of factors that can affect end-to-end latency. The first is the processor overhead required to initiate a data transfer. InfiniBand is a channel-based architecture, meaning that the host will create a channel command list to describe the transfer to be performed. The Host Channel Adapter (HCA) will then execute the command list using its DMA engine. Traditional PCI devices with DMA engines require a system call to perform this operation, which is very expensive in terms of processor overhead. Even more significant is that the

operating system will often copy the data to internal buffers before initiating the DMA. This is one of the primary performance limitations for such software stacks as TCP/IP. InfiniBand was able to draw on concepts from Intel's VI Architecture to minimize processor overhead. The application is able to build its own transfer requests, and initiate the transfer via the HCA without any system calls.

A well-designed packet switch is critical to low latency. InfiniBand was designed so that a switch can cut through a packet from input port to output port by looking at the first 8 bytes of the header, before the packet has even fully arrived. Other switch architectures will buffer the entire packet before deciding what to do with it, resulting in increased latency compared to InfiniBand.

6.3 Reliability, Availability, Serviceability (RAS)

Reliability, Availability, Serviceability is a requirement for this class of system, and was designed in from day 1. There will be InfiniBand installations that may never come down once they are started. Individual components will be repaired/upgraded/replaced without bringing down the fabric and its applications. Enabling this requires a number of capabilities and is more than just error checking.

Certainly, data transfer reliability is important. InfiniBand defines reliable and unreliable data delivery. There are applications where it is ok for data to not be delivered.

Quality of Service capabilities are also defined. On a per connection basis, the fabric can provide minimum priority or bandwidth. InfiniBand goes beyond data reliability to define several security features. Before a data transfer can be initiated, a memory region must be registered with the HCA. The HCA then knows how to translate application's virtual addresses to physical address. The application then communicates to the HCA in terms of logical addresses, or offsets within the registered memory region.

An application can not transfer data outside this memory region, protecting an errant application from transferring data it does not own. Equally important, an application can not read or write data in another node which that node has not allowed. This protection exists for each end-to-end connection, with multiple connections possible between each pair of nodes.

InfiniBand defines a protection domain, which is a group of nodes that can communicate with each other. This is another reliability feature, since nodes in different protection domains are not allowed to talk with each other. A node is allowed to be in multiple protection domains, enabling bridging between domains. Live insertion/removal is supported. Not only is the electrical interface defined so that it will not be physically damaged during insertion/removal, but hardware/software mechanisms are defined so that the operator and/or application can be notified of state changes.

Also defined are mechanisms to monitor the health of the system. Baseboard management functions watch voltage levels, fan speeds, and other system parameters.

6.4 Server Blades

The notion of blade-based servers is a relatively recent phenomenon with both start-ups and the well-established server vendors having already announced server blade products or the intent to develop such products. Typically the early products have focused on power and density as the primary benefits of server blades. This focus on form factor rather than function misses the true benefit of blade based technology. Focusing only on the server

component also misses the other critical parts of a system area networks: I/O connectivity and switching. In fact the true benefit that server blade and more importantly I/O blade technology provides is the ability to deliver a highly available, easy to manage and scalable infrastructure for both computing and I/O. IBA specification sets all the necessary parameters to manufacturers build server blades compatible between systems and backplanes.

7 Conclusions

All the major players in IT business are involved in the development of InfiniBand specification. This fact alone tells us the impact that IBA can have in High Performance Computing (HPC). Parallel busses have been evolving for a long time, but they are now the bottleneck in HPC. Serial interconnects will provide the leap in performance and capability to complement today's processor performance to enable new applications and new ways to approach the HPC paradigm. Some studies claim that by 2006 80% of the servers installed will have support for InfiniBand. Although production products are not available at the time, the interconnection of some prototypes was entirely successful.

The major advantage of InfiniBand from a user's point of view is the ability to dimension the systems accordingly with their needs. Everything now can be arranged in an orderly and efficient fashion. InfiniBand will not replace interconnection technologies such as Gigabit Ethernet or 10 Gigabit Ethernet, and is not intended to. InfiniBand is more than extra bandwidth. InfiniBand is object oriented programming methodology applied to HPC.

References

[1] Infiniband Trade Association: InfiniBand Architecture Specification Volume I, Release 1.1ahttp://www.infinibandta.org, (June 2001).

[2] Infiniband Trade Association: InfiniBand Architecture Specification Volume II, Release 1.1ahttp://www.infinibandta.org, (June 2001).

[3] Pfister, G. F. : Aspects of the InfiniBand Architecture. Proceedings of the 2001 IEEE International Conference on Cluster Computing, (2002).

[4] Mellanox Technologies: Introduction to InfiniBand White Paper. http://www.mellanox.com

[5] Bozman, J.S., Turner, V.: InfiniBand Architecture. IDC White Paper, (January 2001)