Infiniband —— Fast Interconnect

Yuan Liu

Institute of Information and Mathematical Sciences

Massey University

May 2009

Abstract

Infiniband is the new generation fast interconnect provides bandwidths both “inside the box” and bandwidth “out of the box”. The InfiniBand Architecture (IBA) is mainly used by HPC and data centers. The architecture’s elements as well as the advantages are discussed in this report in the areas of bandwidth, CPU utilization, latency and RAS.

1 Introduction

Amdhal’s law states that to gain the best overall performance speed up, the improvement of different part in the system should be even and Moor’s Law states every 18 months, we get double performance from semiconductors. The imbalances between I/O speed and computation power results the industry and market yearn for new generation interconnect. Two major competing forces were then born, one was named Future I/O (FIO), an input/output architecture developed by IBM, Compaq and HP. it aims to increase server I/O throughput, the key problems anticipated for the next generation of computing [1], another was named Next Generation I/O (NGIO), developed by Intel Microsoft and Sun, its main focus is enabling relatively rapid PCI replacement in volume segments. [2] FIO and NGIO merged in 2000 named System I/O (SIO) with the best of the technologies from each side. SIO did not last long and renamed to InfiniBand. InfiniBand, a industry standard switch-based serial I/O interconnect architecture operating at a base speed of 2.5 Gb/s per link in each direction.[3] It is designed to provide high speed, low latency, cost efficient with enhanced reliability interconnect solution and it is mainly used in clusters and data centers, although it provides the bandwidth “inside the box” to replace bus structure and as well as the bandwidth “out of the box”. This report focuses on InfiniBand architecture elements and InfiniBand advantages.

2 InfiniBand Architecture

2.1 InfiniBand elements

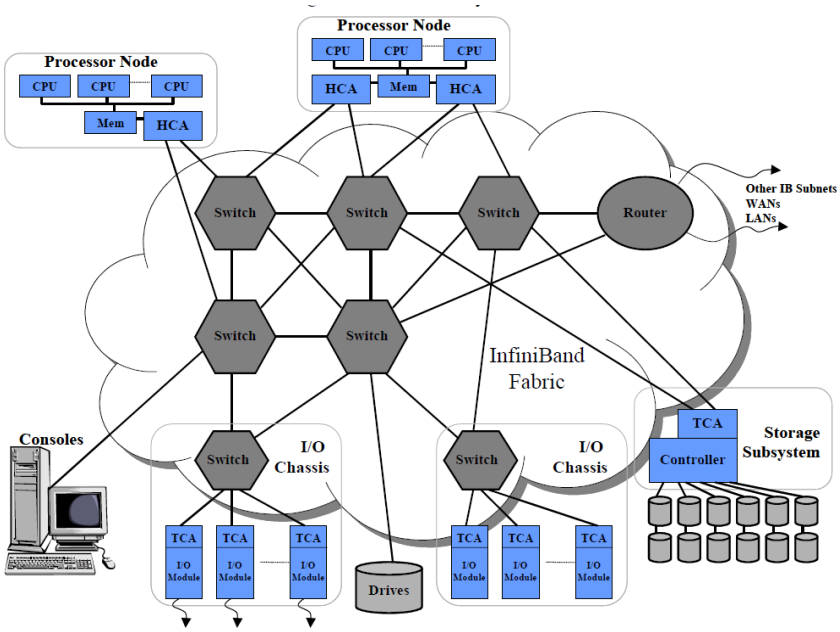

Figure 2.1.1 Infiniband structure [3]

As figure 2.1.1 shows, the key elements in InfiniBands are switches, adaptors and wirings.

InfiniBand switches

As Infiniband is switch based, the InfiniBand switches connect end nodes and forward packets from one port to another. They do not generate or consume packets and support 1X,4X and 12X mode and each mode can be single data, double data and quad data rate speeds shown in figure 2.1.2

	Single (SDR)	Double (DDR)	Quad (QDR)
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s

Figure 2.1.2 Effective theoretical throughput in different configurations[6]

HCAs

Host Channel Adapters are very intelligent, capable of handling large numbers of concurrent connections and typically have a large number of

send/receive buffers. They are used to connect from processor nodes to switches, they can have one or more ports and use local system bus interface, PCI-E for example.

TCAs

Target Channel Adapters, compare to HCAs, are not as much intelligence as HCAs due to the limited scope of their function. They only need to handle a small number of concurrent connections thus they do not have as much send/receive buffer space as HCAs.[3] They are used to connect from I/O nodes to switches. I/O nodes can be single or multiple storage devices.

Wirings

Over twisted pair copper wires, InfiniBand can be spanning up to 30 meters while using fiber cables allow 10kms distance from ordinary. Figure 2.1.3 shows that physical transport media may consist of 1,4,12 lanes, each bandwidth is 2.5Gb/s and representing a separate physical interface.[4]

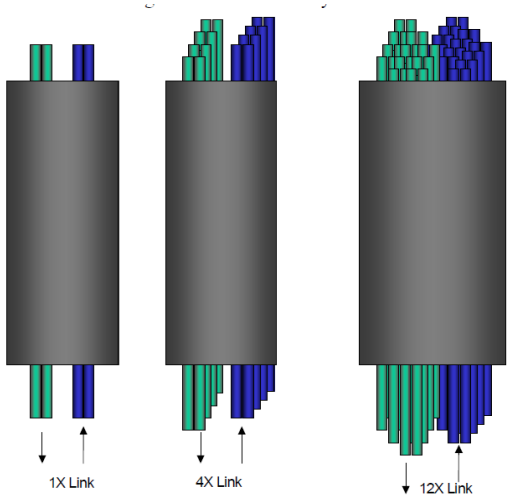

Figure 2.1.3 InifiniBand physical link [3]

2.2 Infiniband for clusters, Why InfiniBand is better than the others

Before InfiniBand was invented, Ethernet was widely used for cluster interconnect, in 9 years, from the birth of InfiniBand, many supercomputers have been adopted to InfiniBand because it provides high bandwidth, low CPU utilization, low latency, scalability as well as RAS. This section will discuss the benefits provided by InfiniBand compare with Ethernet and the classic Transmission Control Protocol

(TCP).

2.2.1.1 High bandwidth, the TCP bottleneck

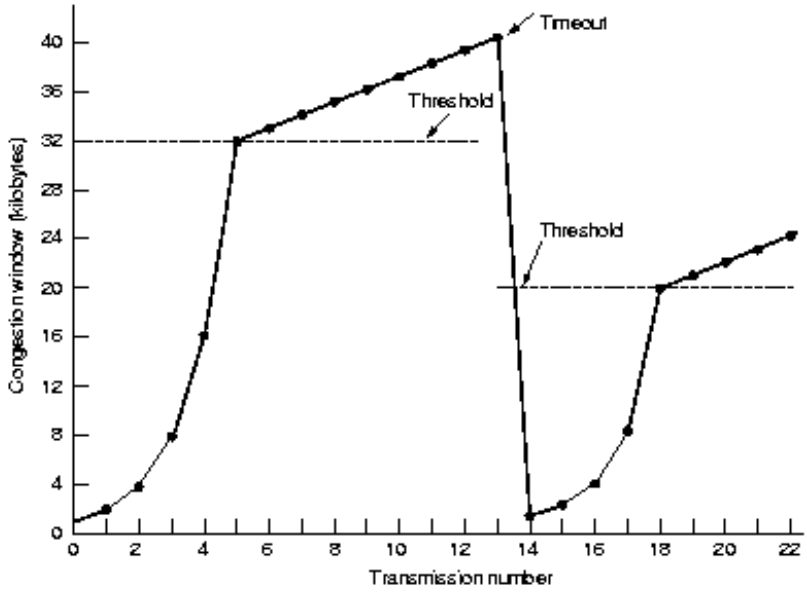

Figure 2.2.1 TCP congestion flow control

TCP uses slow-start algorithm. The behavior of the slow-start algorithm is to send a single packet, await an ACK, then doubles the number of packets needed to be sent, and await the corresponding ACKs until it reaches the slow-start threshold. Then the flow control will change the mode from slow-start to congestion avoidance, it increment number of packet to be sent by one according to last number of packets been successfully sent until router drops packet, i.e. when the router buffer is full [7]. Figure 2.2.1 shows how flow control and congestion control is achieved.
It is the bottleneck for high bandwidth in two ways. First, the slow-start algorithm works well for slow connections and/or continues transmition, it does not suit the behavior of message passing in clusters, which they need to send messages around as fast as possible. The time used in parallel computing should be the computation time. Thus, the slow-start algorithm has a high chance that the messages are sent before reaching the bandwidths limit. The congestion avoidance in TCP makes the case worse because of the slow increment for number of packet to be sent at once.

2.2.1.2 Infiniband eliminates TCP bottleneck plus more

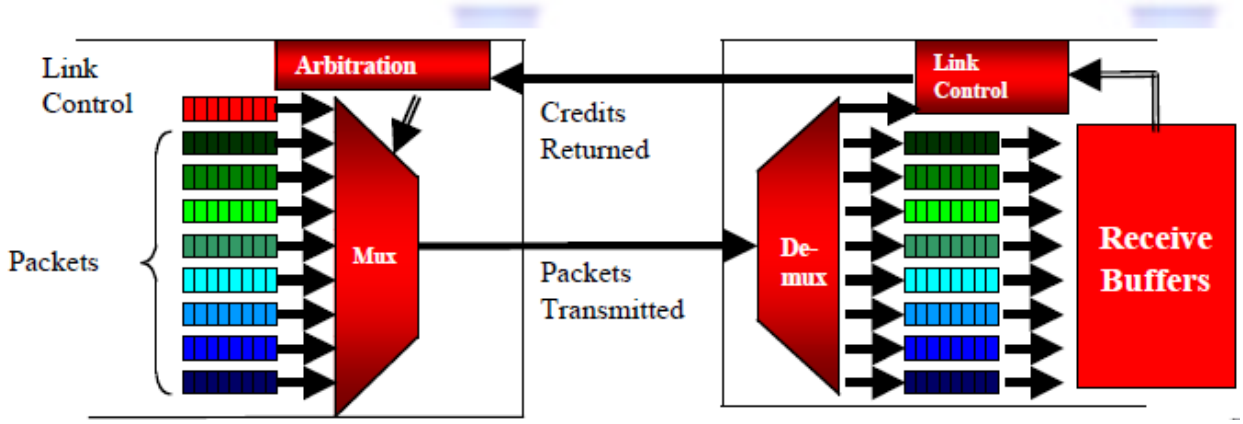

Figure 2.2.2 Infiniband flow control [9]

The InfiniBand link layer uses credit based flow control to avoid congestion, sender keeps track of number of free buffer slots in receiver so there is no slow-start and by knowing the free number of buffer slots, it can use as much of available bandwidth and more efficient. It does not loose data because data is not transmitted unless the receiver advertises credits indicating receive buffer space is available. Furthermore it is more powerful than an XON/XOFF (as twice of efficient) or CSMA/CD protocol (used by Ethernet).[8]

2.2.2.1 Head of line (HoL) blocking

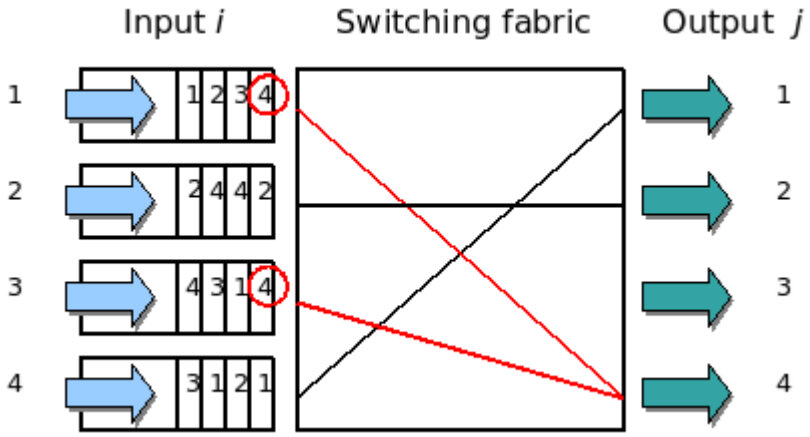

Figure 2.2.3 Head of Line Blocking

HoL blocking happens in buffered switches. Because of the first in first out nature of the input buffer, switches can only transfer packet in head of the buffer, so in the figure 2.2.3 case, for port number 1, packets go to output port 3 cannot be switched even that is free because packets have to wait until the packets to be consumed by port 4 first. Port 4 is busy consuming input from input port 1 and 4 at a time so

packets for port 3 have been blocked. Ethernet and all other network uses switches will have this problem unless switch builds the feature in to avoid HoL blocking. There is nothing specified in the Ethernet.

2.2.2.2 InfiniBand Virtual Lane (VL) alleviates HoL blocking

Each InfiniBand channel adapter, either HCA or TCA, shown in figure 2.2.2, has 1 or more ports, each port has multiple pairs of dedicated send and receive buffers, so it can send and receive concurrently. A Virtual Lane (VL) manages one pair of buffers and has its own flow control characteristics and priority. VL15 has the highest priority among all VLs is dedicated for management messages. Thus, each port has minimum of 2 (VL0 and VL15) and up to 16 VLs, the VL15 is dedicated for management of packets and using alternative paths, it not only alleviates HoL blocking, but also provides load-balancing [9] as well as QoS(see Appendix).

2.2.3 CPU utilization and Scalability

In parallel computing, we expect the system can be scaled as we needed and the processors should do computations on tasks rather than anything else. This is not 100% true in TCP. Ethernet is not an inherently reliable network[12], it runs TCP to ensure reliable communication. At the down side, TCP adds lot of overheads to CPU and network. This leads to leak full bandwidth by passing these overheads around and steals valuable CPU cycles to processing overhead, as the cluster scales, it becomes more and more noticeable and makes Ethernet an impediment to build big clusters.

At InfiniBand network level, InfiniBand provides at least ten times better CPU utilization vs. Gigabit Ethernet clusters by implementing the communications stack in hardware and taking advantage of RDMA capabilities.

2.2.4 RDMA

A significant advantage of InfiniBand is it is Remote Direct Memory Access (RDMA) capability. This allows processor nodes access memory without using heavy slow protocol like TCP.

In Ethernet, (figure 2.2.4.1) it takes 3 copies within system local bus to move data from a network processor node to the application over TCP. Figure 2.2.4.2 shows HCA read/write directly to/from subsystem in InfiniBand Architecture. As RDMA has been build into lowest levels of network interfaces, there is no need for a high over-

head protocol driver to verify integrity and de-multiplex messages to applications. This frees up CPU cycles and local bus cycles [10].

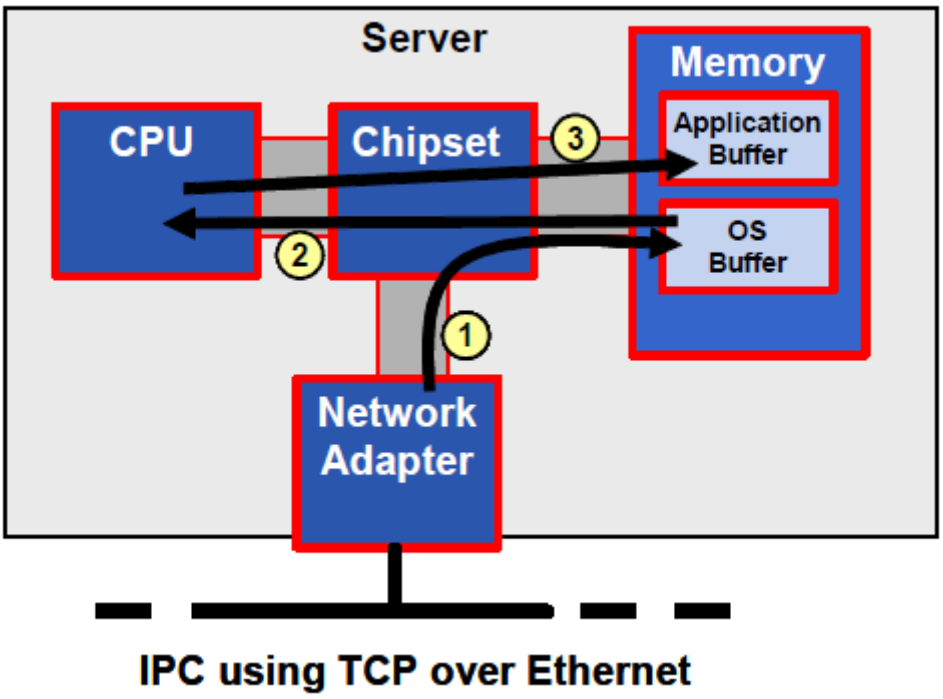

Figure 2.2.4.1 IPC using TCP over Ethernet [10]

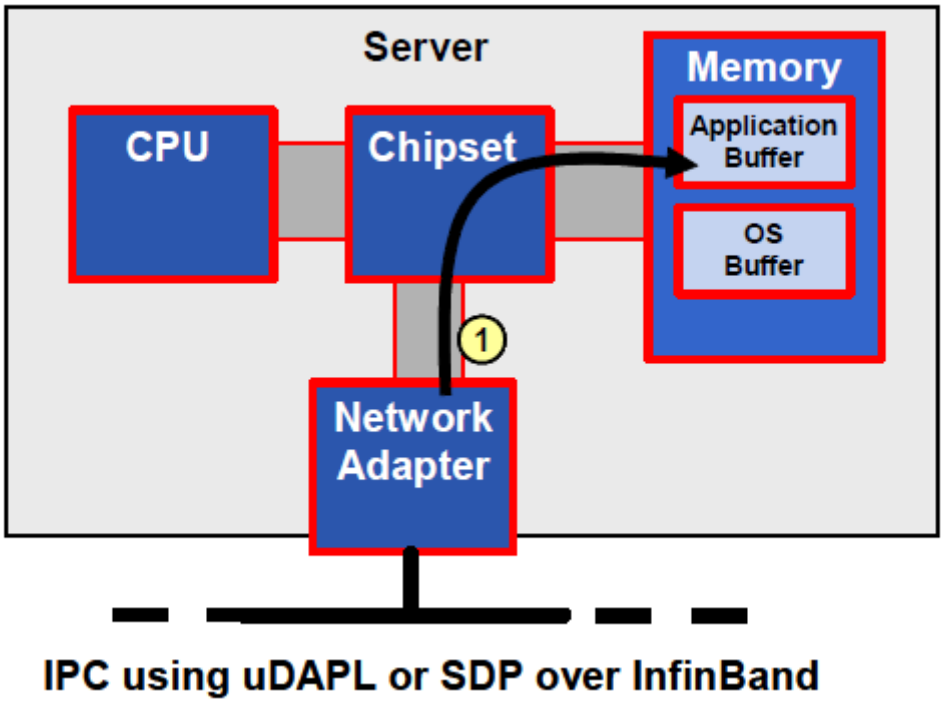

Figure 2.2.4.2 IPC using uDAPL(see Appendix) or SDP over InfiniBand [10]

2.2.5 Latency, Scalability, and RAS

A very important factor for HPC(see Appendix) is latency. In message passing, we expect message to be sent and received as fast as possible. In order to achieve this, system demands high brandwitdth which has been discussed in section 2.1, low latency and efficiently use of available bandwidth. Latency directly effects scalability of the cluster. In clusters using MPI, in most cases, synchronization is required between nodes, the faster they can synchronize, the larger system can scale. InfiniBand latency is at least 10 times faster than Ethernet according to Figure 2.2.5.1, the benchmark results are from Pallas MPI Benchmarks.

RAS stands for Reliability, Availability, and Serviceability. InfiniBand supports reliability with guaranteed reliable services which deliver in-order packet with 2 CRCs to ensure data integrity. The 2 CRC(see Appendix) packets, 16bit Variant CRC(VCRC) which covers the whole packet and recalculate from hop to hop and the 32 bit Invariant CRC(ICRC) which covers the fields do not change at link layer level while Ethernet uses only one single CRC. Availability is provided as it enables redundancy and supports fail-over by switching to an alternative path, should a link or device fail. Serviceability is achieved through hot swappability and special management functions, which can be either in-band or out-of-band.[12]

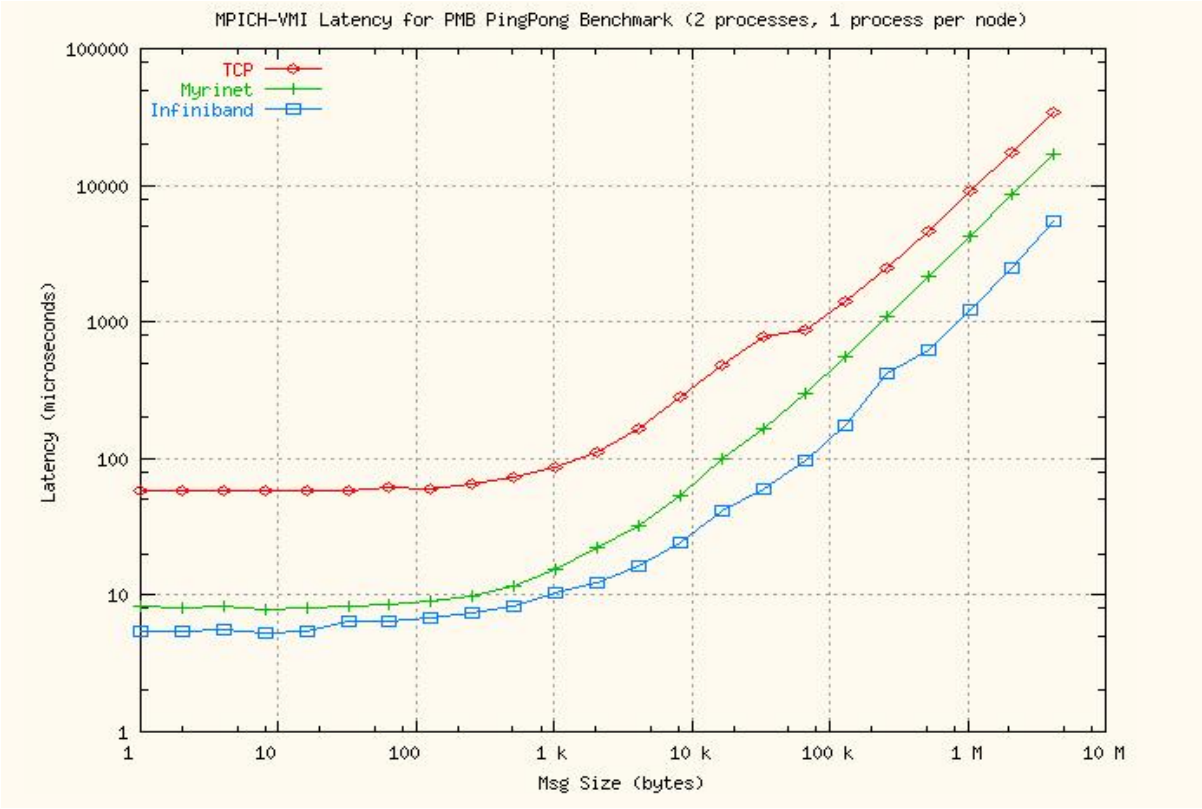

Figure 2.2.5.1 source: http://vmi.ncsa.uiuc.edu/performance/pmb_lt.php[11]

2.3 Current researches and Interconnects in the future

When NGIO and FIO was born, the reason was to eliminate I/O bottle neck. For bandwidth "out of the box" side, According to the InifiniBand roadmap (figure 2.3), in next 3 years, InfiniBand will push its brand width to near 1000Gb/s.[15]

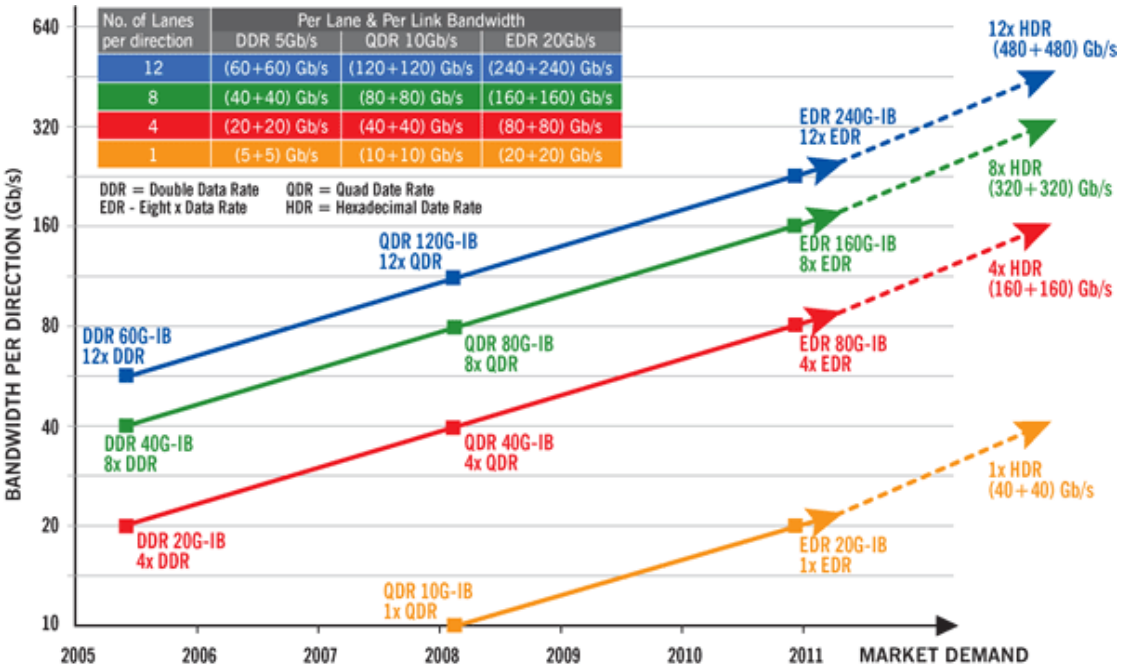

Figure 2.3 InifiniBand Roadmap

On the other hand, for bandwidth "inside of the box", the latest Intel Core i7 family processors integrated memory controller into CPU and replaced system front side bus (FSB) by QuickPath Interconnect (QPI) which provides up to 6.4GT/s speed. The researches and developments in related areas have proofed eliminate system I/O bottle neck is a trend, it is as important as improving processor computing power, thus Amdahl's law stands.

3 Conclusions

InfiniBand Architecture is industry standard fast interconnect, it can co-exist with current network and provides high bandwidth, low latency, high scalability, minimal CPU utilization, RAS and lots more. It can replace the shared bus structure in local system to switched link architecture but there are not many such boards been manufactured so far. It has been used as interconnect in the Top 500 supercomputers including the fastest IBM Roadrunner. Thus the main usage nowadays is use the bandwidth “out of the box” feature provided by InfiniBand to connect clusters and data centers. By seeing the fastest computers configuration currently available, we can predict more and more clusters and data centers will adapt to Infiniband Architecture.

4 References

[1] J.Cowan,C.Madison,G.Still,D.Garcia,M.Bradley & K.Potter, “PI”,*Proceedings of the The 6th International Conference on Parallel Interconnects,* p.238, 1999.

[2] T.Heil,“InfiniBand Adoption Challenges”, *InfiniBand: A Paradigm Shift From PCI.*1 Jun. 2000.

[3] Mellanax Technologies Inc,“Introduction to InfiniBand”,*White Paper*, 2003.

[4]T.C.Jepsen, “InfiniBand”,*Distributed Storage NetworksArchitecture,Protocols and Management*,p.159-174,2003.

[5]T.M. Ruwart, “InfiniBand – The Next Paradigm Shift in Storage”,*18th IEEE Symposium on Mass Storage Systems and 9th NASA Goddard Conference on Mass Storage Systems and Technologies*,17 Apr.2001

[6] Wikipedia, “InfiniBand”, *http://en.wikipedia.org/wiki/InfiniBand,* 18,Apr.2009

[7] G.Huston, “TCP Performance”, *The Internet Protocol Journal - Volume 3, No. 2,*2009

[8] H.T.Kung&R.Morris, “Credit-Based Flow Control for ATM Networks” *IEEE Network Magazine*, March 1995.

[9] C.Eddington, “InfiniBridgeTM: An Integrated InfiniBand Switch and Channel Adapter”. *http://mellanox.com/* 19 Apr. 2009

[10] Oracle,“Achieving Mainframe-Class Performance on Intel Servers Using InfiniBand Building Blocks” *An Oracle White Paper,* April 2003.

[11] NCSA, “Latency Results from Pallas MPI Benchmarks”*Virtual Machine Interface 2.1* 23 Mar. 2005

[12] Mellanax Technologies Inc, “InfiniBand™ Frequently Asked Questions”, *White Paper*, 2003.

[13] S.Shelvapille&V.Puri, “Encapsulation Methords for Transport of InfiniBand over MPLS Networks*”, Internet-Draft*, 12 Mar,2009

[14] Cisco Systems, “Quality of Service”,*Internetworking Technology Handbook(http://www.cisco.com/en/US/docs/internetworking/technology/handbook/QoS.html),* 21 Apr,2009

[15] InfiniBand Trade Association, “InfiniBand® Roadmap”, *InfiniBand® Roadmap (http://www.infinibandta.org/content/pages.php?pg=technology_overview),* 27 May,2009

Apendex A:

Definitions:

Invariant CRC	A CRC covering the fields of a packet that do not change from the source to the destination.[13]
Variant CRC	A CRC covering all the fields of a packet, including those that may be changed by switches.[13]
DAPL	Direct Access Programming Library is a high performance Remote Direct Memory Access API.
High-performance computing (HPC)	A branch of computer science that concentrates on developing supercomputers and software to run on supercomputers. A main area of this discipline is developing parallel processing algorithms and software: programs that can be divided into little pieces so that each piece can be executed simultaneously by separate processors.
QoS	Quality of Service refers to the capability of a network to provide better service to selected network traffic over various technologies, including Frame Relay, Asynchronous Transfer Mode (ATM), Ethernet and 802.1 networks, SONET, and IP-routed networks that may use any or all of these underlying technologies. The primary goal of QoS is to provide priority including dedicated bandwidth, controlled jitter and latency (required by some real-time and interactive traffic), and improved loss characteristics. Also important is making sure that providing priority for one or more flows does not make other flows fail. QoS technologies provide the elemental building blocks that will be used for future business applications in campus, WAN, and service provider networks.[14]