InfiniBand: A New Era in Networking

Vivek D. Deshmukh
Assistant Professor,
S.B. Jain Institute of Technology, Management & Research, Nagpur

ABSTRACT

Now-a-days, the systems are coming with the high speed processors. With the development in the processors, the CPU performance is increasing day-by-day and making applications such as data mining, data warehousing, and e-business commonplace. This growth in computational power requires that the I/O subsystem should be able to deliver the data needed by the processor subsystem at the rate at which is it needed. In the past couple of years, it has become clear that the current shared bus-based architecture will become the bottleneck of the servers that host these powerful but demanding applications. The Peripheral Component Interconnect (PCI) bus, which is a dominant bus, commonly used in both desktop and server machines for attaching I/O peripherals to the CPU/memory Units. The most common configuration of the PCI bus used is as given in the Table**:**

Bit	Clock Rate (MHz)	Bandwidth (MB/s)
32	33	133
64	33	266
64	66	533

Today's desktop machines have lots of capacity available with the PCI bus in the typical configuration, but server machines are starting to hit the upper limits of the shared bus architecture. To resolve this limitation on the bandwidth of the PCI bus, a number of solutions are becoming available in the market as interim solutions such as PCI-X and PCI DDR. But these versions also fail to some extent. For ex: The PCI-X specification allows for a 64-bit version of the bus operating at the clock rate of 133 MHz, but this is achieved by ceasing some of the timing constraints. Because of the shared bus nature of these versions, the bus forces it to lower its fanout in order to achieve the high clock rate of 133 MHz. So, despite the temporary resolution of the PCI bandwidth limitation through these new upgrade technologies, there is a long term solution needed that cannot rely on shared bus architecture.

InfiniBand breaks through the bandwidth and fanout limitations of the PCI bus by migrating from the traditional shared bus architecture into switched fabric architecture.

The InfiniBand™ Architecture (IBA) is an industry standard that defines a new high-speed switched fabric subsystem designed to connect processor nodes and I/O nodes to form a system area network. These new interconnect method moves away from the local transaction-based I/O model across busses to a remote message-passing model across channels. The architecture is independent of the host operating system (OS) and the processor platform. IBA provides both reliable and unreliable transport mechanisms in which messages are enqueued for delivery between end systems. Hardware transport protocols are defined that support reliable and unreliable messaging (send/receive), and memory manipulation semantics (e.g., RDMA read/write) without software intervention in the data transfer path.

I. INTRODUCTION

InfiniBand is a switched fabric communication link primarily used in high-Performance computing. The architecture is based on a serial, switched fabric that in addition to defining link bandwidths between 2.5 and 30GBits/sec, resolves the scalability, expandability, and fault tolerance limitations of the shared bus architecture through the use of switches and routers in the construction of its fabric. The InfiniBand architecture specification defines a connection between processor nodes and high performance I/O nodes such as storage devices. Like Fiber Channel, PCI Express, SATA, and many other modern interconnect, InfiniBand is a point-to-point bidirectional serial link intended for the connection of processors with the peripherals such as disks.

II. INFINIBAND NETWORK

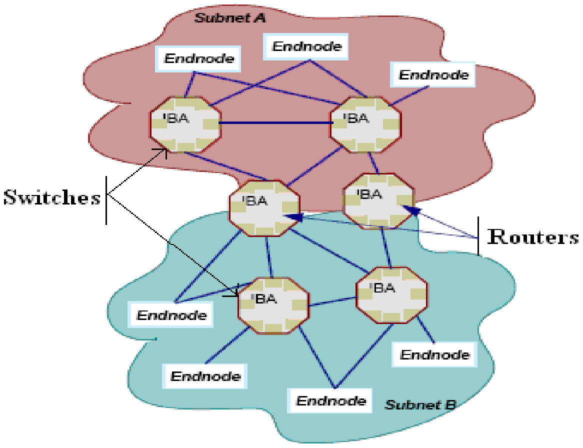

Figure 1: InfiniBand Network

The figure 1, above represents the simplest configuration of an InfiniBand. An Endnode represents either a host device such as a server or an I/O device such as a RAID subsystem. Two or more Endnodes connected through the switch form a Subnet.

Each node connects to the fabric through a channel adapter. There are two types of the channel adapters i.e. Host Channel adapter (HCA) and Target Channel Adapter (TCA). Each processor node contains a host channel adapter (HCA) and each peripheral node has a target channel adapter (TCA). Each channel adapter may have one or more ports providing multiple paths between a source and a destination. The fabric is able to achieve transfer rates at the full capacity of the channel, avoiding congestion issues that arise in the shared bus architecture. Furthermore it provides alternative paths which results in increased reliability and availability since another path is available for routing of the data in the case of failure of one of the link. Two or more subnets are connected using the routers. Each connection between nodes, switches, and routers is a point-to-point, serial connection, which provides number of benefits such as the connection provides the full capacity of the connection to the two endpoints as the link is fully dedicated to the two endpoints. This eliminates the contention for the bus as well as the resulting delays that emerge under heavy loading conditions in the shared bus architecture. The InfiniBand channel is designed for connections between hosts and I/O devices within a Data Center with the advantage of much higher bandwidth can be achieved.

Within a subnet, each port is assigned a unique identifier by the subnet manager called the Local Identifier (LID). In addition to the LID, each port is assigned a globally unique identifier called the GID. Switches make use of the LIDs and forwarding tables for routing packets from the source to the destination, whereas Routers make use of the GIDs for routing packets across the InfiniBand subnets.

III. INFINIBAND MANAGEMENT MODEL

IBA management model is based on following concepts.

1) Subnet Managers

Subnet Manager provides functionality for controlling and examining various aspects of subnet configuration and operation. There will be at least one Subnet Manager in each subnet. There can be multiple subnet managers in one subnet. In that case only one subnet manager will be master and the rest will act like standby subnet managers. If the master subnet manager goes down the passive subnet manager will be take over.

2) Subnet Management Agents

Each and every channel adapter, switch or router will have this low level functionality which responds to the commands and queries sent by Subnet Manager. It will set/query internal parameters of the channel adapter, switch or router.

3) Messaging scheme

There is a specific messaging scheme defined for communication between subnet managers and subnet agents or between two subnet managers. This messaging scheme specifies basic message types and interfaces through which subnet managers and subnet agents communicate.

4) A collection of specific messages including message content and related behaviors

Specific messages and message sequences are defined in terms of message contents and associated required behaviors. Messages are grouped into classes depending upon type of management activity the message support.

In order for an application to communicate with another application over the InfiniBand it must first create a work queue that consists of a queue pair (QP). In order for the application to execute an operation, it must place a work queue element (WQE) in the work queue. From there the operation is picked-up for execution by the channel adapter. Therefore, the Work Queue forms the communications medium between applications and the channel adapter, relieving the operating system from having to deal with this responsibility.

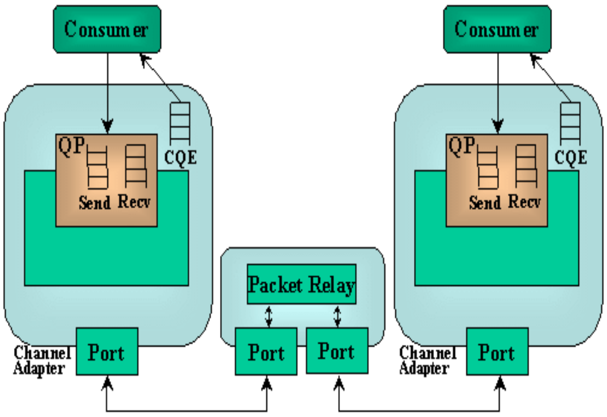

Figure 2: InfiniBand Communication Stack

The list of operations supported by the InfiniBand architecture at the transport level for Send Queues is as follows:

1) Send/Receive:

Supports the typical send/receive operation where one node submits a message and another node receives that message. One difference between the implementation of the send/receive operation under the InfiniBand architecture and traditional networking protocols is that the InfiniBand defines the send/receive operations as operating against queue pairs

2) RDMA-Write:

This operation permits one node to write data directly into a memory buffer on a remote node. The remote node must of course have given appropriate access privileges to the node ahead of time and must have memory buffers already registered for remote access.

3) RDMA-Read:

This operation permits one node to read data directly from the memory buffer of a remote node. The remote node must of course have given appropriate access privileges to the node ahead of time.

InfiniBand Link	Signal Pairs	Signaling Rate	Data rate	Full-Duplex Data Rate
1x	2	2.5 Gbits/s	2 Gbits/s	4 Gbits/s
4x	8	10 Gbits/s	8 Gbits/s	16 Gbits/s
12 x	24	30 Gbits/s	24 Gbits/s	48 Gbits/s

4) RDMA Atomics:

This operation name actually refers to two different operations that have the same effect but which operate different from one another. The Compare & Swap operation allows a node to read a memory location and if its value is equal to a specified value, then a new value is written in that memory location. The Fetch Add atomic operations reads a value and returns it to the caller and then add a specified number to that value and saves it back at the same address.

For Receive Queue the only type of operation is:

1) Post Receive Buffer:
Identifies a buffer into which a client may send to or receive data from through a Send, RDMA-Write, and RDMA-Read operation.

InfiniBand Provides following Transport Service Types:

1) Reliable Connection:

Reliable transfer of data between two entities.

2) Unreliable Connection:

Unreliable transfer of data between two entities. Like Reliable connection there are only two entities involved in the data transfer but message may be lost.

3) Reliable Datagram:

The QP can send and receive messages from one or more QPs using a reliable datagram channel between each pair of reliable datagram domains.

4) Unreliable Datagram:

The QP can send and receive messages from one or more QPs however the messages may get lost.

5) Raw Datagram: The raw datagram is a data link layer service which provides the QP with the ability to send and receive raw datagram messages that are not interpreted.

.Data Rates supported by InfiniBand:

The InfiniBand specification supports three data rates over both copper and fiber-optic cables. These are 1x, 4x and 12x. The comparison between these three is as shown in the following table.

The base data rate, 1X, is clocked at 2.5 Gbits/s and is transmitted over two pairs of wires i.e. Transmit and Receive which yields an effective data rate of 2 Gbits/s. The InfiniBand 4X and 12X interfaces use the same base clock rate, but uses multiple pairs, where each pair commonly referred to as a lane.

IV. PROTOCOLS SUPPORTED BY INFINIBAND

InfiniBand supports following protocols:

1) IPoIB:
IP over InfiniBand (IPoIB) allows TCP or UDP/IP applications to run over the InfiniBand transport and enables IP communications between InfiniBand attached servers or other IP devices. IPoIB also enables standard, sockets-based IP applications to be accessed on InfiniBand-attached servers.

2) SDP:
Sockets Direct Protocol (SDP) is an InfiniBand specific protocol which defines a standard wire protocol over IBA fabric to support stream sockets (SOCK_STREAM) networking over IBA. SDP utilizes various InfiniBand features (such as remote DMA (RDMA), memory windows, solicited events etc.) for high-performance zero-copy data transfers.

3) SRP:
SCSI RDMA protocol (SRP) is designed to take full advantage of the features provided by the InfiniBand Architecture. SRP allows a large body of SCSI software to be readily used on InfiniBand Architecture and is rapidly emerging as the protocol of choice for block-based storage.

4) iSER:
The iSCSI Extensions for RDMA (iSER) protocol maps the iSCSI protocol over a network that provides RDMA services like TCP with RDMA services (iWARP) or InfiniBand. This permits data to be transferred directly into SCSI I/O buffers without intermediate data copies.

5) iWARP:
The Internet Wide Area RDMA Protocol (iWARP) is an update of the RDMA Consortium's RDMA over TCP standard. iWARP is a superset of the Virtual Interface Architecture that permits zero-copy transmission over legacy TCP. It may be thought of as the features of InfiniBand applied to Ethernet.

V. KPIT INFINIBAND SRP INITIATOR DEVICE DRIVER DEVELOPMENT

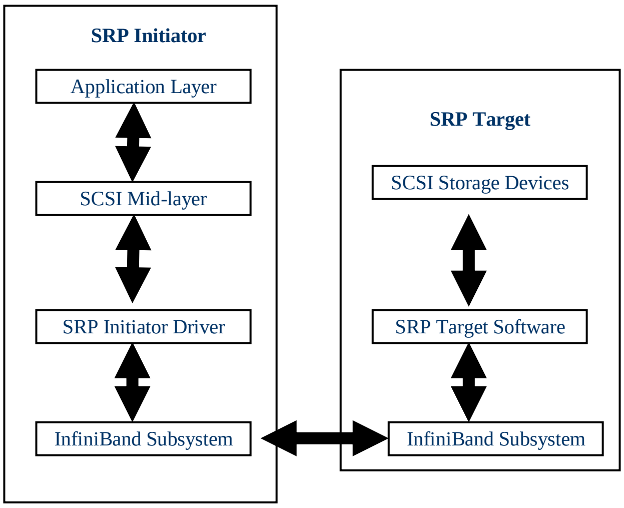

Figure 3: SRP Device Driver Block Diagram

KPIT AT-SSG group has developed InfiniBand SRP Initiator driver for UNIX operating system. As a part of device driver development activity, the performance analysis of the InfiniBand SRP driver was carried out. Following table shows number of I/Os performed per second. The IO Size column shows size of single IO (Device column 1k, 2k etc) and number of I/O’s performed per second over InfiniBand (IB) link and Fiber channel (FC) link. The MB/s column shows the speed of data transfer for IB and FC link. It can be easily seen from the below table that InfiniBand (IB) provides faster data transfer speed than FC.

CONCLUSION

InfiniBand is an interconnect technology that provides high throughput and low-latency transport for efficient data transfer between server memory and I/O devices, without CPU intervention. The InfiniBand specification defines the architecture for an interconnect that will pull together the I/O subsystems of the next generation of servers and will ultimately even move to the powerful desktops of the future. InfiniBand was originally envisioned as a comprehensive "system area network" that would connect CPUs and provide all high speed I/O for "back-office" applications. In this role it would potentially replace every datacenter I/O standard including PCI, Fiber Channel, and various networks like Ethernet.

REFERENCES

[1] InfiniBand Architecture Specification Volume 1 Release 1.2

[2] KPIT InfiniBand SRP Device Driver Project Documents

[3] SRP Protocol Specifications Version 1.6

IO Size			**MB/s**	
I/O Size	**No of I/Os on IB Link**	**No. of I/Os on FC Link**	**Speed IB Link**	**Speed FC Link**
1k	71903.31	17232.06	70.22	16.83
2k	69837.96	17069.75	136.40	33.34
4k	62239.80	16922.01	243.12	66.10
8k	24909.54	16290.71	194.61	127.27
16k	12268.94	9593.27	191.70	149.89
32k	4927.69	5491.01	153.99	171.59
64k	4411.44	2946.16	275.71	184.14
128k	3184.86	652.23	398.11	81.53
256k	2099.56	323.37	524.89	80.84
512k	1050.33	161.90	525.16	80.95
1024k	525.00	95.48	525.00	95.48