Introduction to the InfiniBand(TM) Architecture

Document Information

<table>
<thead>
<tr>
<th>Document Category:</th>
<th>White Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Jeff Russell, John Steczkowski</td>
</tr>
<tr>
<td>Source Format:</td>
<td>MS Word 2000</td>
</tr>
<tr>
<td>Source File:</td>
<td>IntroToInfiniBand13</td>
</tr>
<tr>
<td>Template:</td>
<td>GeneralRpt.dot</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control Type:</th>
<th>Informal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Control:</td>
<td>None</td>
</tr>
<tr>
<td>Review Checklist:</td>
<td>None</td>
</tr>
<tr>
<td>Approval Authority:</td>
<td>Author</td>
</tr>
</tbody>
</table>

Revision History

0.1 Original review draft
1.0 Version sent for final review and approval
  1.1 Tech written and other minor edits
  1.2 Final comments
  1.3 Jeff Russell Comments

Table of Contents

ABSTRACT ................................................................................................................................. 1

TODAY'S I/O ARCHITECTURE ................................................................................................. 1
  PCI Characteristics .............................................................................................................. 2
  Reliability, Availability, Serviceability ........................................................................... 2

TOMORROW'S INFINIBAND(TM) ARCHITECTURE ................................................................. 3
  Reliability, Availability, Serviceability ........................................................................... 4

APPLICATIONS OF THE INFINIBAND(TM) ARCHITECTURE ............................................... 5
  Advanced Applications ......................................................................................................... 5
Abstract

This paper is designed to provide an introduction to the InfiniBand™ Architecture (IBA) and explain its purpose by describing a typical PCI based “standard, high-volume” server architecture and detailing how IBA will replace and improve upon PCI architecture. IBA is being promoted by a consortium of industry leaders organized into the Infiniband™ Trade Association.

In addition, a series of simple applications are described to illustrate the strengths of IBA, such as improved scalability, increased availability, and improved performance.

The goal of IBA is to replace PCI’s shared bus technology with a switched fabric architecture that will improve the Reliability, Serviceability, Availability (RAS) and performance characteristics of servers. This will be accomplished through standards that define port redundancy, fail-over characteristics and system management interfaces. Performance is increased by the switched fabric nature of the infrastructure so that applications don’t contend for bandwidth as in a shared bus environment. (An I/O unit provides the connection and interface between I/O peripherals and the InfiniBand* subnet; it is analogous to an HBA or NIC in a PCI-based system.)

Today’s I/O Architecture

In typical entry to mid-range servers, the standard I/O architecture uses a PCI bus as the communications medium to adapter cards that talk to storage devices and networks. An adapter card that interfaces to a storage protocol is called a Host Bus Adapter (HBA). Two storage protocols comprise the majority of server applications: Storage Area Networks (SAN) based on Fibre Channel and the parallel bus Small Computer Systems Interface (SCSI). An adapter card that interfaces to a Local Area Network (LAN) such as Ethernet is called a Network Interface Card (NIC). The basic goal of PCI is to provide a standard, modular interface for adapter cards.

![Server I/O based on a PCI bus architecture.](image-url)
As illustrated in Figure 1, the typical server architecture includes a memory controller that connects the CPU, memory, and the PCI bus bridge. The PCI bus provides the connection to one or more adapter cards. Each adapter card consists of a PCI interface and an I/O controller. These various PCI adapter cards give the system access to various media outside of the system chassis. Traditionally, the PCI adapter cards are viewed as the “access point” for getting data into or out of the system. The dashed line indicates the boundary of the typical server system.

**PCI Characteristics**

Although there is no doubt that the PCI bus has served the industry well for many years, it has several characteristics that limit performance as well as make implementation and administration more difficult.

PCI uses a shared memory architecture that prevents separation of the various I/O controllers’ address spaces. The failure of an adapter can potentially cause interference with the memory spaces of other adapters. Additionally, the shared memory architecture requires the CPU to interact directly with the I/O controller on the adapter card. This requires the CPU to slow down to bus speeds while manipulating data on the I/O controller, and this interaction impacts performance and exposes the CPU to any failures on the PCI bus.

Adapter cards use Direct Memory Access (DMA) to move blocks of information between system memory and the I/O medium. Each PCI adapter has its own DMA engine. This means each PCI adapter has its own register-based programming interface that complicates device driver development.

Since PCI uses a shared, parallel bus, there are electrical limitations on the speed at which the bus can run and limitations on the number of PCI adapters that can be used at any given bus speed. In the case of PCI-X, there is a limit of one adapter card on the bus at the highest speed. This also limits system design flexibility since the parallel bus can only be several inches long, therefore the PCI cards must be enclosed in the server.

Since the I/O bridge is an arbiter, system integrity relies on all of the adapters cooperating together on the bus. If one card misbehaves or malfunctions, there is no way for the system to diagnose and isolate the failing adapter card. This inability to isolate a fault may result in system failure. Besides fault isolation problems, the fair distribution of bus bandwidth depends on the cooperation of adapter cards instead of a central authority.

**Reliability, Availability, Serviceability**

Reliability describes the ability of a system or component to perform its required functions under stated conditions for a specified period of time. Reliability is expressed as the likelihood that a system will not fail, thus comparisons in reliability examine the probability of a failure. Availability describes the degree to which a system or component is operational or accessible when required for use. Serviceability describes the ability of a component to be installed, exchanged, or removed from a system.

The shared bus architecture of PCI has some distinct shortcomings when evaluated in terms of reliability, availability, and serviceability (RAS). Reliability can be examined at a detailed level and at a system level.

The PCI bus consists of over 100 signals, and if any one of those signals does not work, i.e. fails, the bus will not operate properly. The likelihood of failure of a single PCI bus signal is rather low. However, the probability of failure of a signal is increased where the adapter card attaches to a bus using a circuit board edge connector that does not have a positive seating mechanism. The likelihood of a PCI bus failure is proportional to the number of bus signals times the probability of a single signal failing. Examining reliability at a system level, the shared nature of the PCI bus means that the likelihood of a PCI bus failure is also proportional to the number of PCI adapter cards installed in the system. Thus, reliability of a system is decreased due to the shared bus and the large number of signals on the bus.
To improve availability, it is desirable that a single failure on an adapter card does not disable the bus and prevent the system from operating. PCI does not have any provisions for isolating any individual card so the system can continue to operate without that card. Furthermore, there is no way to create a redundant connection between the PCI bus bridge and the adapter card because of the shared nature of the bus.

PCI is not optimized for serviceability. Some systems have provisions for hot pluggable PCI cards, but these extensions to the standards are weak from an enterprise availability standpoint and for these reasons, most IT professionals do not use hot-pluggable PCI. Additionally, since the adapter cards are internal to the server, the bus must be powered down, which often requires the server to be shut down.

In summary, the shared nature of the PCI bus creates limitations in the level of reliability, availability, and serviceability that can be achieved. As mission critical applications are implemented in entry and mid-range servers, the next level of RAS is required.

**Tomorrow’s InfiniBand™ Architecture**

The purpose of IBA is the same as PCI. That is, to connect servers (CPU and memory) to external storage and LANs. In this respect, IBA is just a replacement of the PCI architecture, however it does have a series of system enhancements as outlined below. As with any new technology, there will be a migration to IBA as the system I/O interconnect. This will result in a coexistence of PCI based I/O and IBA based I/O.

As shown in Figure 2, the key difference between IBA and PCI is that a switched fabric in IBA replaces the PCI shared bus. The server interface to the InfiniBand fabric is called a Host Channel Adapter (HCA). What were PCI adapters are now called I/O Units, but the function is the same. In its simplest form, each I/O Unit consists of a an I/O controller and a Target Channel Adapter (TCA). The end node access points are called “channel adapters” because the IBA hardware supports high performance communication channels between the CPU and the I/O controllers. The switched fabric uses point-to-point links and switches to connect multiple servers and multiple I/O Units in a fabric configuration. The IBA fabric
provides the media for server clustering and the capability to connect multiple servers to storage devices or LANs connected through a single I/O controller.

The InfiniBand™ fabric is controlled by a fabric manager that discovers the physical topology of the fabric, assigns local identifiers, and establishes routing between end nodes. Furthermore, the fabric manager controls changes, such as adding or removing nodes.

The switched fabric architecture allows for the modularization of servers. Instead of having the I/O Units inside of the server, the server “box” can consist of one or more CPUs, memory, and an HCA that can connect to external IBA switches and I/O Units. In this configuration, servers can be replaced or upgraded as necessary without disturbing the peripheral configurations. As more connections are required, more switches can be added to build a larger fabric. InfiniBand™ Architecture implements a channel communication model between the server CPU and I/O controllers. Instead of shared memory architecture, the CPU and I/O controller are decoupled. A reliable delivery mechanism is implemented by IBA so that a CPU and I/O controller can exchange messages. This communication channel simplifies device driver design.

The channel based communication allows direct, protected access to host memory. This means that a misbehaving I/O unit cannot corrupt system memory and cause a system crash. Efficient host channel adapter implementation insures that the CPU is decoupled from the details of I/O communication resulting in improved performance and fault isolation should an I/O Unit fail.

The IBA specifies a means to construct redundant physical paths between end nodes and switches. Servers and I/O Units can have two or more ports for fabric connection that allow redundant configurations to be constructed. The fabric routing tables are aware of the redundant connections and can manage fail over should a fault occur.

Clustering will be able to be implemented as an evolution of the current technology. Vendors will not need to implement proprietary physical clustering mechanisms with IBA links available.

The IBA fabric infrastructure will provide scalable performance. In PCI, the bus bandwidth of approximately 500 MB/sec is shared by all of the devices on the PCI bus. With the IBA fabric, each link to a device has a minimum connection bandwidth of 500 MB/sec, and link bandwidth scales up to 2GB/sec for a single link.

The IBA packet format defines a global routing header. This will enable IBA fabrics to be connected across wide area networks through the use of edge routers on each fabric.

Basic IBA links have a distance limitation of many meters compared to PCI's limitation of several centimeters. This provides for flexibility in the placement of servers and I/O controllers in rack mounted configurations.

**Reliability, Availability, Serviceability**

One of the compelling reasons to use IBA for I/O connection is the RAS improvement compared to the PCI bus. Recall the short analysis of PCI reliability, or factors the contribute to the likelihood of system failure. At the detailed level, the numerous PCI signals and the printed circuit board edge connection contribute to the probability of failure. The shared bus contributed to the probability of system failure.

To address the reliability issues, IBA uses point-to-point links with fewer signals and a mechanical plug and jack. The basic link connection is implemented with four signal wires, resulting a smaller link failure multiplier, compared to the over 100 signals of a PCI bus. Individual signal failure is reduced by eliminating the card edge connector that provides a positive seating mechanism and prevents derating issues due to wear on the edge connectors. The IBA standard provides for multiple ports for each I/O Unit, this enhances the reliability of the system by providing multiple physical routes to a single device.

Availability is enhanced through redundancy and a standard fabric management technique. The point-to-point links allow problems to be diagnosed and new routes to be established should a device or link fail.
All IBA devices will have a common way of being managed, reducing both the learning curve and the amount of time necessary to remedy problems on the fabric. The IBA I/O units are designed from the ground up to be hot pluggable, enhancing serviceability. Some I/O units may be stand-alone devices. In those cases, replacement may be as simple as unplugging the IBA connector from one device and plugging in a new device. In other cases, the I/O Units may be adapter card format that can easily be installed or removed while the server is powered on. The IBA standard describes the adapter format, the physical signaling, and the management interface so that I/O Units can be hot swapped.

**Applications of the InfiniBandSM Architecture**

The IBA fabric allows data center scalability. Servers can be added as increases in computing power are needed and I/O units can be added as increases in storage and network and storage bandwidth are needed. The IBA fabric can be used for clustering servers and as a means to share I/O units. As an enhancement over the PCI paradigm, servers can share I/O units.

The IBA specification describes two basic form factors of I/O Units. Stand-alone units will have one or more external IBA ports and one or more media-specific ports. This implementation option will be more typical of advanced devices that provide functionality beyond a typical HBA or NIC. The second form factor is the adapter card format. These I/O Unit adapter cards can be inserted into a IBA standard chassis. The chassis implements a back plane that functions as a switch to interconnect the I/O Unit adapter cards.

**Advanced Applications**

The InfiniBandSM Architecture opens up many possibilities for advanced applications. The basic applications are the network attachment through a NIC and the storage connection through an HBA.

A traditional server connects to many LAN segments via a router and one or more NICs in the server. IBA enables a high speed, low latency connection between the server and local area networks by
embedding a TCA inside of a network router. This reduces the number of devices required to connect a server to its network media and consolidates this connection with the network router device.

A similar advanced application can be used to consolidate storage connections into a single storage router. Multiple HBAs are replaced by a single high speed, low latency connection to a storage router. The number of required devices is reduced by embedding the server connection into the storage router, which also provides the functionality of routing between SCSI buses and Fibre Channel storage area networks (SAN). The storage router serves as a central point from which to manage the storage area network.

Just as connecting the LAN and SAN to a server via a router provides a benefit, some small applications may benefit from direct attachment of a RAID subsystem using IBA.

The diagram in Figure 3 shows a variety of IBA products classified into families. Host connectivity products include the HCA and accompanying modifications to the server operating system (OS) and hardware abstraction layer (HAL). The basic infrastructure products include switches to build a fabric, an general InfiniBand router to provide inter-fabric connectivity, and an I/O chassis the provides connectivity for I/O Unit adapters. The I/O Unit adapters interface storage, network, and other media to the server through a server-resident device driver (DD). The advanced applications such as the LAN router, the storage router, and the RAID subsystem products are also shown in Figure 3.

Figure 3. InfiniBand™ Architecture product families.
This paper may contain forward-looking statements that involve risks and uncertainties. Among the important factors which could cause actual results to differ materially from those in the forward-looking statements are Crossroads’ ability to develop new and enhanced products that achieve market acceptance; the continuation of Crossroads’ successful relationships with its limited number of OEM customers; Crossroads’ ability to retain and recruit key personnel to manage its business successfully; and that Crossroads’ stock price could be volatile regardless of Crossroads’ actual financial performance and other factors detailed in the Crossroads’ most recent filings with the Securities and Exchange Commission.