

Chapter 1

Introduction

InfiniBand architecture is a new server I/O technology that improves the way servers interconnect with I/O devices and with each other. It is one of the first new technologies of the millennium, a revolutionary new technology that transitions us into the 21st century.

I/O (which stands for Input and Output) refers to the ability to move data in and out of the processor/memory complex of a compute node. I/O takes place on several levels; InfiniBand architecture addresses the lower layers, defining an I/O interconnect and its transport protocols. Typically, high volume servers have used I/O technology, such as PCI, developed for desktop computers. PCI devices include storage and network adapters that enable the computer to attach to peripheral buses such as Fibre Channel, Parallel SCSI, and Ethernet. In this respect, InfiniBand architecture is a PCI replacement, promoting faster and more powerful adapters.

However, InfiniBand is much more than a PCI replacement. It provides many features and capabilities previously found only in mainframe architectures and it provides advanced capabilities necessary for server clusters. It also allows I/O adapters to move out of the server the same way that Fibre Channel devices reside in independent chassis and racks. Thus, storage arrays (RAID, etc.) can connect directly to InfiniBand fabric. In fact, some analysts believe InfiniBand could eventually become a storage interconnect like Fiber Channel and iSCSI, due to performance, cost, interoperability, etc.

The two primary goals driving this new technology are: overcoming PCI limitations (performance bottlenecks, expandability, scalability, etc.) and standardizing the proprietary technologies emerging in the clustering space (e.g., Servernet, Myricom, Giganet, etc.). Of course, it brings many new challenges and it changes the way we design, build, and incorporate server products and build enterprise infrastructures.

InfiniBand architecture is a rich and complex technology designed to overcome existing barriers, and at the same time, to provide the framework for new and enhanced features and capabilities. Not only

does InfiniBand provide an improved I/O interconnect and communication infrastructure, but it also changes the way we build and manage data centers.

Many senior architects from the world's leading computing companies collaborated on the specification to advance server I/O well into the future. With the specification process completed, the time has come to produce useful and valuable products and solutions.

Yes, it is a complex architecture, but it doesn't have to be difficult. The key is in understanding both the capabilities and the goals of the architecture. In particular, product vendors need to understand strengths and weaknesses so they can provide optimum products. End-users¹ must plan properly to take full advantage of InfiniBand products.

Most new technologies have underlying principles that are not obvious just from reading the specification, and InfiniBand architecture is no exception. First, industry specifications are significantly different from product specifications; a product specification tells exactly what is being built or purchased, while an industry specification tells what is or is not allowed. In many cases, what an industry specification doesn't say is just as important as what it does say.

While product engineers and site managers are concerned with "*where the rubber meets the road*,"² architecture has been described as "*where the rubber meets the sky*." It is important to understand the reason for this difference. In developing a new technology that will evolve and have a long lifetime, the architects must look into the future and build in features not necessarily relevant in today's products. Again, InfiniBand architecture is no exception. However, "evolve" is the key word. Thus, this book provides a road map identifying near-term and long-term goals of the architecture.

Another observation is that the scope of InfiniBand architecture covers a large range and spans many market segments (storage, communications, networking, server platforms, mainframe computing, etc), from standard high-volume server products to high-end server solutions. This book

¹ Data center managers, MIS managers, IT managers, network engineers, department managers, etc. are the end-users or customers of the technology and play an important role in how well InfiniBand architecture progresses.

² For those not familiar with this phrase, it refers to tires on a vehicle, and means where the real work is done.

discusses how the various features of the InfiniBand architecture apply to those market segments.

Scope

The goals of this book are:

- Explain the features and capabilities of InfiniBand architecture.
- Explain various deployment strategies and provide site managers with the knowledge they need to make intelligent decisions on how to adopt, purchase, deploy, and manage InfiniBand architecture based products and solutions.
- Identify various market segments and identify how InfiniBand architecture applies to them.
- Guide hardware and software developers through the InfiniBand architecture, tying architectural concepts together, setting practical expectations, and identifying responsibilities for hardware vendors, software vendors, operating system vendors, and platform vendors.
- Remove the mysteries of the architecture and promote the values that enable application developers to take full advantage of the technology.
- Educate product architects on how to make optimum use of InfiniBand features.
- Clarify the roles and expectations for the various product developers:
 - Channel adapter vendors.
 - Server platform vendors.
 - Operating System Vendors (OSVs).
 - Independent Hardware Vendors (IHVs).
 - Independent Software Vendors (ISVs).
 - Applications as an IPC³ client.
 - Management applications.
 - Network service providers.

³ IPC - Inter-Process Communication is the ability for processes on different nodes to directly communicate with each other including sending data and accessing each other's memory.

The bulk of this book focuses on typical solutions and common practices for the early adopter; i.e., the first few years of InfiniBand deployment in every-day, practical environments. InfiniBand architecture has potential that goes beyond these bounds.

The ultimate goal is enabling multi-vendor InfiniBand-based products that work and play well in a heterogeneous environment.

A Brief History

InfiniBand architecture resulted from two industry initiatives, Next Generation I/O (NGIO) and Future I/O. Intel started working on what eventually became InfiniBand as early as 1996, when a team of architects was asked to develop an I/O technology based on the Virtual Interface (VI) Architecture.⁴ The purpose of this effort was twofold:

1. Produce a serial interconnect that would permit I/O to move away from the CPU/memory complex.
2. Have that same serial interconnect provide a VI-compliant interface for inter-process communication (IPC) between applications on different servers.

The whole concept was based on VI architecture for communication between hosts and I/O pass-through to PCI cards in a remote chassis (i.e., a transparent extension to PCI). It was more than just serial PCI, but it still had many of the same characteristics and limitations.

After tedious evaluation and designing, someone asked the architects if they thought it was the right architecture. None of the architects thought so. For various reasons, they all thought that for such a significant effort, it just didn't provide solutions to the other I/O problems that servers faced, such as improving scalability, serviceability, and reliability. Additionally such an effort should provide a strong foundation for emerging concepts such as user level I/O, third party I/O, and device sharing.

To make a long story short, they were then asked to develop an I/O technology that would be worthy of the investment that Intel would ask

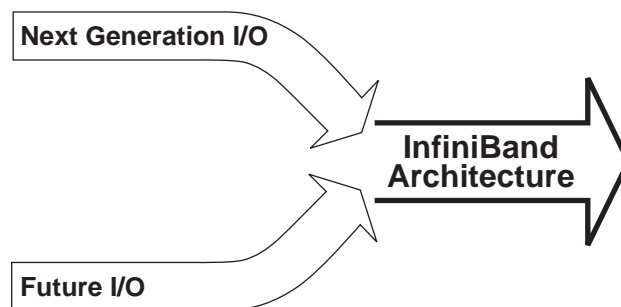
⁴ VI Architecture Specification V1.0. For more information, see the VI Developers' Forum (VIDF) web site <http://www.vidf.org>. The VI specification can be downloaded from <http://www.viarch.org>. The developer guide for VIPL, "Intel VI Architecture Developer's Guide V1.0," can be downloaded from http://developer.intel.com/design/servers/vi/developer/ia_imp_guide.htm.

the industry to make. Other companies also had similar experiences, recognized the same problems, came to the same conclusions, and wanted to collaborate in developing a common solution. This effort, soon named Next Generation I/O (NGIO), became an industry initiative led by Dell Computer, Hitachi, Intel, NEC, Fujitsu Siemens, and Sun Microsystems. NGIO grew to over 100 companies.

In this same time frame, another initiative known as Future I/O (FIO), led by IBM, Compaq, Adaptec, 3Com, Cisco, and Hewlett-Packard emerged. FIO also grew in size with many companies belonging to both initiatives.

These two initiatives, NGIO and FIO, had for the most part a common set of goals, but differed in some key areas. In general, NGIO focused on the standard high volume (SHV) server market while FIO targeted high-end platforms.

Both camps worked hard to be the first to produce a finished specification, knowing that the industry had no room for two new competing technologies. In the end, they both realized that fragmentation of the industry was the worst possible outcome, so over the summer of 1999, they worked out plans to merge the two initiatives. By October 1999, the InfiniBand Trade Association (IBTA) formed. Rumor has it that the name comes from infinite-bandwidth,⁵ a name that the architects on both sides equally disliked.



OM11735

Figure 1.1 Roots of the InfiniBand Architecture

⁵ InfiniBand is a trademark. It doesn't officially stand for anything except the name of the architecture and the name of the trade association that produced the architecture.

Both teams had a lot in common. Both initiatives were based on the same technologies and they both borrowed the same concepts from existing technologies – switched fabric, signaling characteristics (Fibre Channel, Ethernet), mainframe channel technology, proprietary cluster interconnect, and VI architecture. For example, VI architecture, which started in Intel labs and was jointly developed with Compaq and Microsoft, standardized the interface to cluster interconnects and promoted Inter-Process Communication (IPC). However, VI architecture is a software specification and both sides realized that VI architecture concepts needed to be pushed down as part of the transport and implemented in hardware.

The merge turned out better than anticipated, and the architecture ended up with far more capability than just the sum of the two original initiatives. It was the best of both technologies, and truly a collaborative, industry-wide effort. The InfiniBand Trade Association has seven steering committee members (Compaq, Dell Computer, IBM, HP, Intel, Microsoft, and Sun), 11 sponsor members (3Com, Adaptec, Agilent Technologies, Brocade Communication Systems, Cisco Systems, EMC, Fujitsu Siemens, Hitachi, Lucent Technologies, NEC, Nortel), and over 230 general member companies.

In October 2000, the InfiniBand Trade Association held its Fall Developers Conference announcing to the world that they had completed the *InfiniBand Architecture Specification* and it was available to the industry. That release (r1.0) consists of two volumes, architecture and electrical-mechanical. The specification was developed by 150 architects from over 30 companies and reviewed by more than 150 companies – an amazing feat to occur in just one year.

In June 2001, the IBTA released its first update of the specification (r1.0.a) that provides clarifications/errata updates. There are currently a number of Annexes under development as the architecture continues to evolve and grow.

That is the history of InfiniBand. The important point is that a significant number of senior architects from prominent companies collaborated to merge the best concepts into one common architecture. Now it's time to answer the question "Just what did they create?"