Architecture and Application of Infortrend InfiniBand Design

Application Note

Version: 1.3

Updated: October, 2018

Abstract:

Focusing on the architecture and application of InfiniBand technology, this document introduces the architecture, application scenarios and highlights of the Infortrend InfiniBand host module design.

Contents

What is InfiniBand

Overview and Background

The InfiniBand Architecture originated in 1999 as the merger of two important proposals known as Next Generation I/O and Future I/O. These proposals and the InfiniBand Architecture are all rooted in the Virtual Interface Architecture (VIA), which is based on two concepts: ability for applications to exchange data directly between virtual buffers across a network and direct access to the network interface from the application space. The architecture enables data exchange between applications without involving the operation system to handle network processes. Therefore, in comparison to other network protocols, such as TCP, IP, and Ethernet and storage interconnects including Fiber Channel and iSCSI, InfiniBand's application-centric approach delivers excellent bandwidth performance.

Basics of InfiniBand

Hardware

The InfiniBand fabric is created with host channel adapters (HCA) and target channel adapters (TCA) that fit into servers and storage nodes and are interconnected by switches that tie all the nodes together over a high-performance network fabric. The InfiniBand Architecture defines a full set of hardware components necessary to deploy the architecture. Those components include:

- ✓ **HCA – Host Channel Adapter**

An HCA is the point where an InfiniBand end node, such as a server or storage device, connects to the InfiniBand network. InfiniBand supports a range of possible implementations without particular HCA functions implemented in hardware, firmware or software. In other words, the HCA provides the applications with full access to the network resources. A function, which is called address translation, allows an application to directly access the HCA without the assistance of operation systems.

- ✓ **TCA – Target Channel Adapter**

A TCA is not widely deployed today because a TCA is mostly used in an embedded environment. The applications in an embedded environment may be based on an embedded operating system or state machine logic and therefore may not require a standard interface for applications.

- ✓ **Switches**

An InfiniBand Architecture switch is conceptually similar to any other standard networking switch, but molded to meet InfiniBand's performance. They implement InfiniBand's link layer flow control protocol to avoid dropping packets. This is a key element of InfiniBand since it means that packets are never dropped in the network during normal operation. Compared to traditional TCP/IP Protocol, this "no drop" behavior makes InfiniBand the most efficient transport protocol.

- ✓ **Cables and Connectors**

The InfiniBand Architecture can support both active and passive copper cables as well as a wide range of optical cables and provide connectivity for 1x, 4x and 12x speeds covering all data rates: SDR, DDR, QDR, FDR and EDR.

Architecture

✓ **Direct Access**

InfiniBand provides applications a quick and easy messaging service which can be used in communication with other applications, networks and storages. Instead of making a request to the operating system for access to one of the server's communication resources, an application accesses the InfiniBand messaging service directly. Direct access means that an application need not rely on the operating system to transfer messages. This idea is in contrast to a standard network environment where the shared network resources, such as traditional TCP/IP network, are owned by the operating system and cannot be accessed by the user application. The messaging service provides a straightforward method for applications to access network resources, and therefore the complexity of network processes is eliminated and server message transferring is achieved without the operating system.

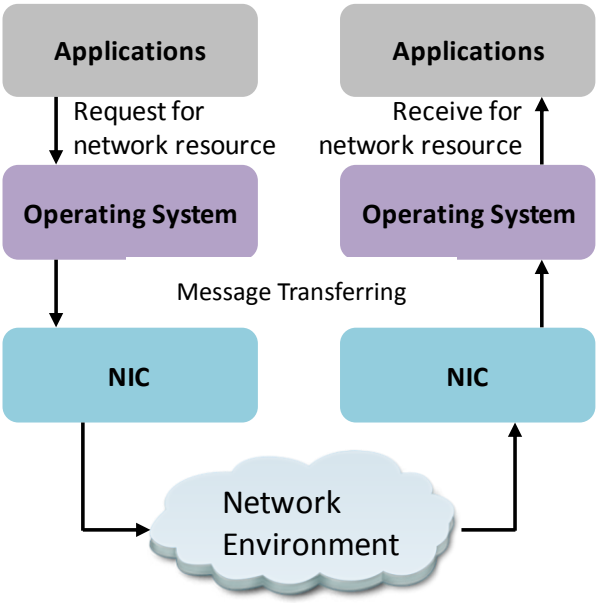


Figure 1 Traditional message flow between applications and network resources.

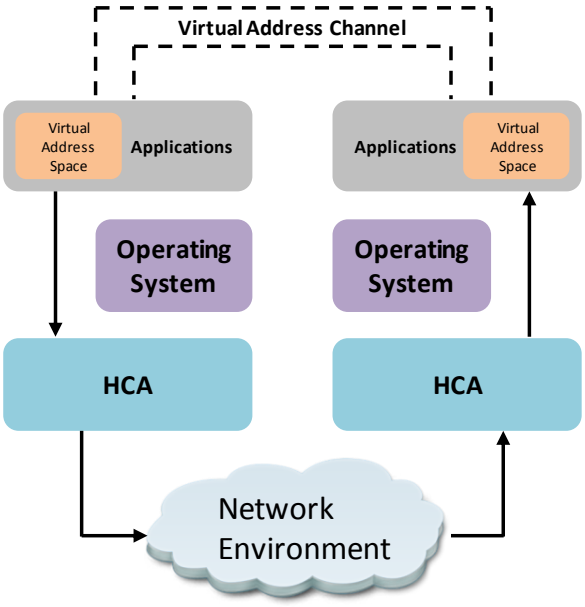


Figure 2 Messages bypassing the operating system via virtual address channels

✓ **Virtual Address Spaces**

InfiniBand provides the messaging service by creating a channel connecting an application to any other application, storage or service. Virtual address spaces are created for carrying messages and ensure that the channels are protected and unique. Those channels are the connections among virtual address spaces which are deployed in the huge network environment.

✓ **Queue Pairs**

To recognize each application and connection, the concept of Queue Pairs has to be introduced. Each Queue Pair consists of two components: Send Queue and Receive Queue and represents one end of the channel. One application is able to create multiple Queue Pairs depending on a variety of connection requirements. InfiniBand message service architecture is based on Queue Pairs and in order to ensure the unique connection and bypass the operation system, each application owns their virtual address space and keeps a mapping of the address book in the space. Accordingly, the application at each node of connections has a unique virtual channel to directly access another application.

✓ **SCSI RDMA Protocol (SRP)**

Infortrend storage systems support **SCSI RDMA Protocol** (SRP), also known as SCSI Remote Protocol published as an ANSI standard in 2002 and renewed in 2007. It allows one server to access SCSI storage attached to another server via remote direct memory access (RDMA). Compared to traditional TCP/IP communication Protocol, RDMA ensures higher throughput and lower latency.

The SRP protocol plugs into Linux using the SCSI layer. The SRP devices, such as an Infortrend InfiniBand Storage, can be physically located anywhere on the fabric.

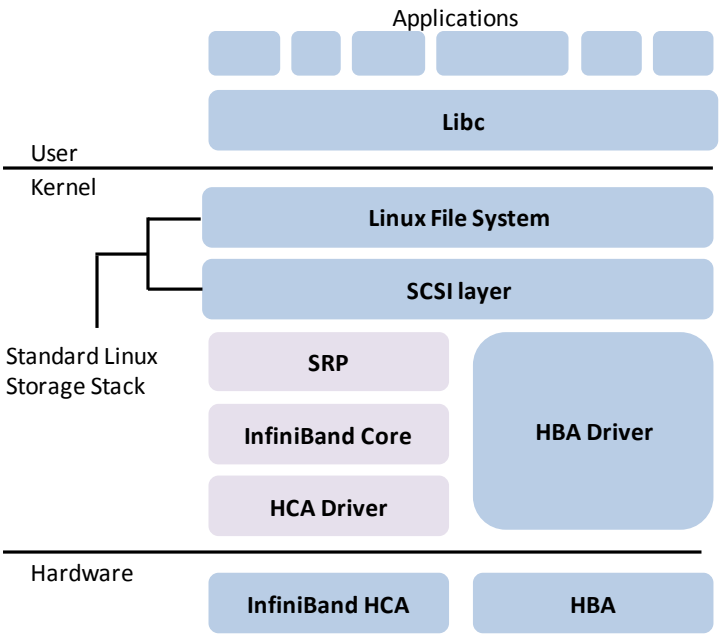


Figure 3 The SCSI RDMA protocol enables interfacing between InfiniBand HCA and SCSI layer in a Linux environment.

Application Scenarios for HPC

HPC systems rely on a combination of high-performance and low-latency storage systems to deliver performance and scalability to applications. An HPC application runs multiple processes in parallel. In an HPC cluster, these processes are distributed across the CPU cores or even a number of processors among servers. According to the purpose and configuration of a cluster system, the demand for bandwidth will increase as the whole cluster is scaled up. However, if we put HPC message passing traffic and storage traffic on a TCP network, it may not provide enough data throughput for either. Also, many HPC applications are IOPS-hungry and need a low-latency network for the best performance. It is recommended to use InfiniBand when storage and computational traffic is combined since 10GbE networks have 3-4 times the latency of InfiniBand. Many cases show 10GbE has limited scalability for HPC applications and InfiniBand proves to show better performance.

✓ **Shared Disk Cluster File System**

The shared disk cluster file system distributes its work in parallel to each cluster server node and the cluster server nodes share access to a common storage pool by creating their own channels to directly access the shared storage system. The load in the shared storage system is the sum of each node's workload.

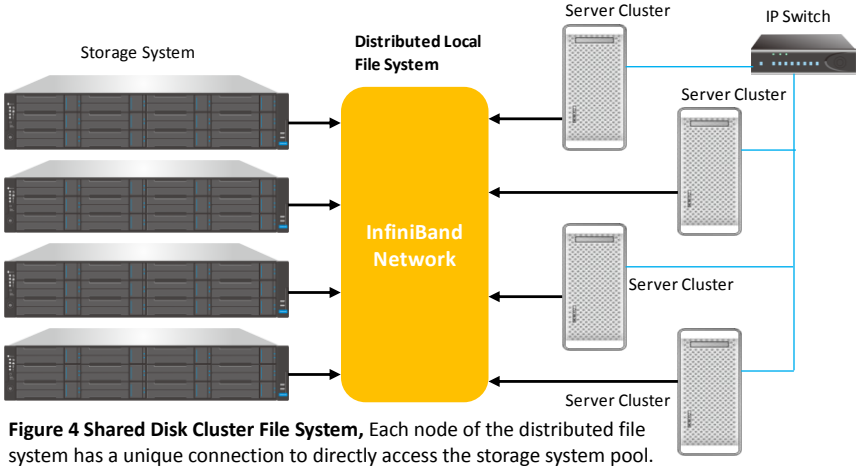


Figure 4 Shared Disk Cluster File System, Each node of the distributed file system has a unique connection to directly access the storage system pool.

✓ **Parallel File System**

A parallel file system is designed to satisfy storage demand from a number of clients in parallel, but independent of others. The file system is with the storage devices rather than being distributed. The metadata server and data server are usually separated. This algorithm reduces the frequency of accessing the file system metadata and allows the user data to be distributed among the storage pool.

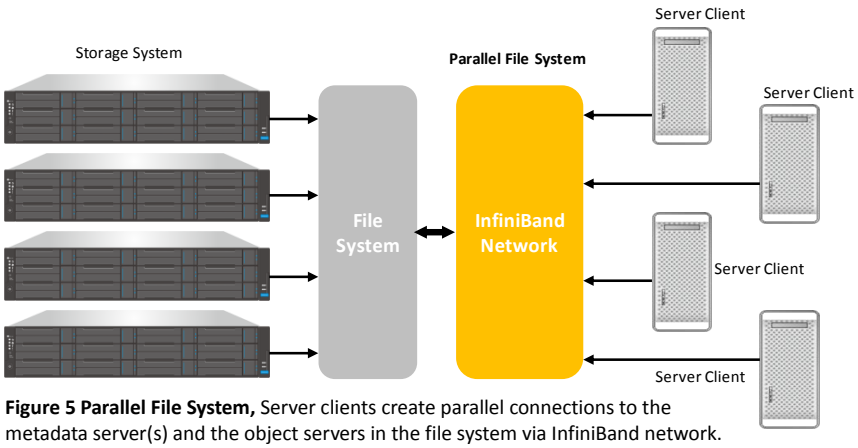
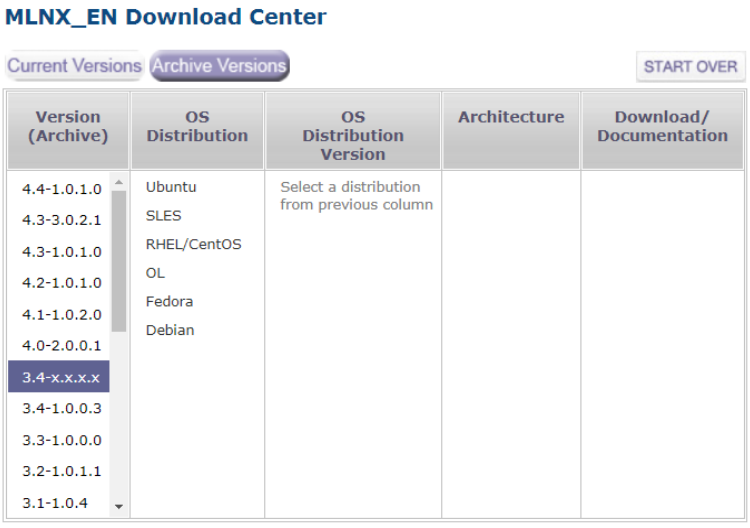


Figure 5 Parallel File System, Server clients create parallel connections to the metadata server(s) and the object servers in the file system via InfiniBand network.

Current Limitation

Before configuring InfiniBand connection between your host server and Infortrend storage system, you are required to install the Linux driver on your host server to ensure it can operate properly. Currently, Infortrend storage systems only support Mellanox HBA for InfiniBand host connection. Please visit their website and download the Linux driver that corresponds to your OS version.

[Note] Currently, Infortrend storage systems only support up to Mellanox Linux driver version **3.4-x.x.x.x**. Please DO NOT install the driver version after 4.0-2.0.0.1.



Mellanox Linux Driver Download Page

Infortrend InfiniBand Host Board Design

InfiniBand Host Board with Controller Module

1. The memory size per controller should be larger than 16GB if two InfiniBand host boards are setup in one controller.
2. The InfiniBand connectivity is only supported in a Linux environment.
3. Supported models:

Storage Family	Notes
EonStor GS Family	Available only for high-end models (GS: 2000 or higher/DS: 3000U or higher). Please refer to our latest product data sheet.
EonStor GSa Family	
EonStor GSc Family	
EonStor DS Family	

4. Only supported in block-level

Figure 6 EonStor GS 3000 controller rear view with InfiniBand host board

Performance Test for InfiniBand

Test Model: EonStor DS 4024RUB

Drive: 48x SAS 12Gb/s HDDs

Memory: 16GB DDR4 per controller

Benchmark Tool: FIO-2.1.7-1.el6.rf.x86_64

Block Size	I/O Behavior	Read	Write
1MB	Sequential	11,158 (MB/s)	6,300 (MB/s)

Figure 7 Rear view of Infortrend InfiniBand host board

Conclusion

Infortrend is dedicated to providing storage solutions of the highest quality to customers around the globe. This is made possible with stringent regulations within our in-house manufacturing process that monitors all details of our production lines. Infortrend products are suited for businesses and corporations of various sizes in a range of sectors including government, healthcare, IT, education, multimedia, database, data backup, surveillance and many more! For more details about Infortrend products and success stories, please visit www.infortrend.com or contact one of our representatives.