Using OpenFabrics InfiniBand for HPC Clusters

The OpenFabrics Alliance was formed to resolve issues with hardware and software interoperability and to deliver open source software for Remote Direct Memory Access fabric technologies like the InfiniBand architecture. Major initiatives by this alliance have resulted in the creation of an open source Linux® OS–based software stack for InfiniBand, which is rapidly gaining acceptance in the high-performance computing field.

BY MUNIRA HUSSAIN, RINKU GUPTA, AND TONG LIU

Effective server resource utilization requires a balance among CPU speed, memory bandwidth, and I/O performance. However, traditional bus-based architecture is limited in its ability to meet fast-growing CPU performance. To eliminate the I/O bus architecture bottleneck, the InfiniBand Trade Association (IBTA), consisting of leading computer technology companies like Dell, Intel, Sun, and AMD, was formed to develop specifications for the InfiniBand architecture. By the end of 2000, the IBTA had defined InfiniBand technology, which is a switched fabric I/O architecture that replaces the traditional server PCI bus with a high-speed, low-latency serial I/O interconnect to deliver a point-to-point, scalable interconnect infrastructure. In contrast to proprietary interconnects, InfiniBand is based on industry standards, enabling it to evolve and scale.

InfiniBand has shown itself to be a compelling high-performance interconnect technology for performance-focused cluster environments. Because of its performance and scalability, InfiniBand has been adopted in high-performance computing (HPC) cluster environments for communication-intensive applications, including those in the fields of aerospace and defense, oil and gas exploration, fluid dynamics, weather modeling, and biology and life sciences. According to the June 2006 TOP500 Supercomputer Sites list of the world’s most powerful computers, InfiniBand is rapidly gaining acceptance as a high-performing cluster interconnect, and the number of InfiniBand-based supercomputers has grown 33 percent since November 2005.

In addition to its benefits for Interprocess Communication (IPC), InfiniBand has been successfully recognized and utilized in the storage systems field. By delivering a converged communication and storage interconnect, InfiniBand can help simplify data center management and help reduce total cost of ownership. The adoption of

---


InfiniBand in the server and storage market, in turn, is driving the demand for development of a complete software stack to support InfiniBand products from different vendors. To fulfill this demand, the OpenFabrics Alliance (formerly the OpenIB Alliance) was formed in June 2004 to create a standard open source InfiniBand software stack. The availability of a standard open source software stack can help promote wide adoption and interoperability across heterogeneous computing architectures.

**InfiniBand hardware architecture**

Established as an industry standard, the InfiniBand architecture develops a different approach toward I/O efficiency by replacing the traditional shared bus architecture with a switched fabric. As Figure 1 shows, the unified InfiniBand I/O fabric includes three basic components: host channel adapter (HCA), target channel adapter (TCA), and InfiniBand switches. HCAs are connection interfaces to CPUs; TCAs are links to storage, Fibre Channel networks, and other I/O nodes. InfiniBand switches provide the switched connections between HCAs and TCAs. InfiniBand routers can be used to connect multiple switched subnets.

The InfiniBand protocol is an OS bypass protocol; it provides direct access to the InfiniBand HCA and can reduce the number of user-kernel context switches and memory copies. It off-loads data movement from the server CPUs to the InfiniBand HCA, enabling communication between devices and hosts without the traditional system resource overhead associated with network protocols. With Remote Direct Memory Access (RDMA) enabled, InfiniBand adopts a zero-copy approach to transferring data between the sender’s memory and the receiver’s memory without involving the host CPUs. RDMA has the potential to be useful in communication-intensive applications.

InfiniBand offers traffic management through virtual lanes (VLs), creating multiple virtual links within a single physical link. This mechanism allows a pair of linked devices to isolate communication interference from other connected devices. Each InfiniBand link can accommodate 2 to 16 VLs, which includes 1 VL for traffic management and others for packet transmission.

InfiniBand delivers high bandwidth by utilizing full duplex bidirectional links between devices. It is designed to support three data transfer rates: 1X, 4X, and 12X, with a base data rate (1X) of 2.5 Gbps, thus theoretically providing full duplex single data rate (SDR) bandwidths of 5 Gbps, 20 Gbps, and 60 Gbps. Ongoing development of InfiniBand technology with double date rate (DDR) and quad data rate (QDR) operations may help enable significantly increased throughput.

**OpenFabrics software stack**

The InfiniBand architecture is a complex and hierarchical structure comprising many components, libraries, programming interfaces, and modules along with multiple protocols that can be used based on application requirements. The InfiniBand interconnect gained popularity in the open source community as research and scientific computing labs adopted it. As a result, multiple versions of the software stack became available. To standardize the InfiniBand software stack, the OpenFabrics Alliance (then named the OpenIB Alliance) was formed to create a standardized Linux OS–based InfiniBand software stack that would be hardware and vendor independent. Key members of the alliance include Dell, Intel, Cisco Systems, and AMD; InfiniBand-focused companies like Mellanox Technologies, SilverStorm Technologies, and Voltaire; and research and scientific computing labs such as Lawrence Livermore National Laboratory, Los Alamos National Laboratory, and Sandia National Laboratory.

The maturity of the InfiniBand stack and the efforts of this alliance have resulted in the adoption of InfiniBand into the upstream Linux kernel. As a result, core components of InfiniBand have been accepted in the Linux 2.6 kernels (2.6.11 and later). Since its inception, the OpenFabrics Alliance has broadened its charter to include other RDMA-capable data center fabrics. The OpenFabrics Alliance now supports iWARP (RDMA on Ethernet) along with InfiniBand.

**InfiniBand software architecture**

The InfiniBand software architecture (see Figure 2) consists of both kernel-level and user-level components, which complement each other to help provide an end-to-end solution. In the software stack, the low-level InfiniBand kernel driver module is hardware specific and ported on top of the associated hardware. The rest of the architecture is hardware agnostic and is divided into kernel-space software and user-space software.

---

1 Additional details about the InfiniBand architecture can be found in the InfiniBand specification at www.infinibandta.org/specs.
Kernel-space software
The kernel-space software consists of the core InfiniBand modules and upper-layer protocols.

Core InfiniBand modules. The core InfiniBand module layer consists of the following:

- **InfiniBand verbs**: These are defined semantically in the InfiniBand specification and describe the action or function to take place. The actual translation of the semantics of a verb to the form of an application programming interface (API) is carried out on top of the hardware driver in the kernel-level software.
- **Management Datagram (MAD) services and agents**: These define an entry point for the upper layers to interact with the HCA driver and hardware.

The InfiniBand specification defines the Subnet Manager, which is responsible for topology discovery and management. It also defines a group of managers called General Services Managers, which are responsible for management and operation of the InfiniBand hardware; examples of such managers include the Performance Manager, which is responsible for retrieving performance statistics from devices, and the Communication Manager, which is responsible for managing connections between devices. Each General Services Manager typically has a corresponding agent on the InfiniBand device (for example, the Performance Manager Agent or Device Manager Agent) and uses MAD packets to request information or operations.

The MAD services layer in the kernel includes components like the Subnet Management Interface (SMI) and General Services Interface (GSI). The SMI is a unique queue pair (QP, the InfiniBand mechanism to send and receive data), called QP0, associated with InfiniBand ports to send and receive information and operation request-response packets between the Subnet Manager and its agents. Similarly, the GSI is a unique QP1 associated with InfiniBand ports to communicate with the various General Services Managers and their agents.

Upper-layer protocols. The rest of the kernel-space software consists of upper-layer protocols like IP Over InfiniBand (iPoIB), Socket Direct Protocol (SDP), SCSI RDMA Protocol (SRP), and Internet SCSI Extensions for RDMA (iSER) in the kernel space.

User-space software
The user-space software consists of access to certain user-level InfiniBand services. InfiniBand allows user-level access to the InfiniBand hardware by providing user-space APIs. The user-level verbs API (semantically similar to the kernel-space verbs) allows upper-level middleware like Message Passing Interface (MPI) software and other applications to bypass the kernel and OS to obtain direct access to the InfiniBand hardware.

OpenFabrics Enterprise Distribution
The OpenFabrics Enterprise Distribution (OFED) is an effort by the OpenFabrics Enterprise working group—a group composed of select hardware vendors providing products based on the OpenFabrics stack—to create a commercial-quality, solution-level OpenFabrics software distribution for end users. It enables end users to deploy InfiniBand devices from various hardware vendors by using a generic software stack. The OFED distribution is a superset of the OpenFabrics software, and contains a snapshot of the OpenFabrics software and components such as MPI software (which falls outside the scope of OpenFabrics) along with the latest kernel modules taken from the OpenFabrics repository.

As of June 2006, the first release of the OFED distribution includes a comprehensive list of kernel modules and software packages such as iPoIB, SDP, SRP, User Direct Access Programming Library (uDAPL), MPI implementations like Open MPI and MVAPICH, subnet managers like OpenSM, and various performance and diagnostic tools.

InfiniBand interoperability and manageability
The OpenFabrics software stack enables hardware and software interoperability and manageability between heterogeneous InfiniBand clusters. Such capabilities have helped this software stack to contribute immensely toward the adoption of InfiniBand. For more information, visit www.openfabrics.com.

Munira Hussain is a systems engineer adviser in the High-Performance Computing Group at Dell.

Rinku Gupta is a systems engineer and consultant in the High-Performance Computing Group at Dell.

Tong Liu is a systems engineer in the High-Performance Computing Group at Dell.