

Leading IT companies deliver the first interoperability demonstration of InfiniBand and iWARP

Data center managers contemplating the adoption of InfiniBand Architecture (IBA) or iWARP Ethernet now have a new element to consider in making their decision. The OpenFabrics Alliance (OFA) is creating a single, open sourced, interoperable software stack that supports both transports for Linux. InfiniBand is supported by the OpenFabrics Enterprise Distribution (OFED) software stack, which has been qualified to interoperate with multi-vendor hardware solutions. Ethernet is supported by an alpha version of an RDMA-over-Ethernet (iWARP) software stack. An application can use either the Host Channel Adapter (HCA) for InfiniBand or the RDMA-enabled Network Interconnect Card (RNIC) for iWARP without needing to know which PCI Express I/O device is connected to the server.

InfiniBand and the iWARP extensions to Ethernet are two transport technologies that support RDMA (Remote Direct Memory Access), though the actual underlying transport technologies differ. They offer other advanced features as well. Because the application program is the same for either transport technology, it makes it easier for the adopter of the technology to consider either or both options for their data center fabric choice.

OFA's creation

As part of its mission to support RDMA-capable transports, regardless of the actual transport technologies used, the OFA developed a single software stack for Linux that supports both technologies. Figure 1 below is a snapshot of the evolving software stack being developed by OFA:

Figure 1

OpenFabrics Software Stack

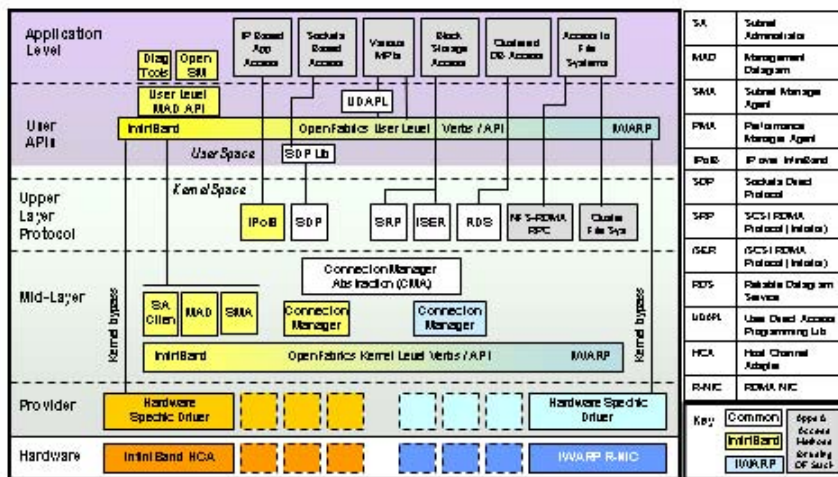


Figure 1 shows the InfiniBand and iWARP-specific components, as well as those that are common. The end result of this architecture is two-fold:

1. The application can use any common application-level access protocol to achieve transport-independence.
2. The application can add code specific to one of the lower-level access layers to take advantage of underlying, transport-specific features for improved performance.

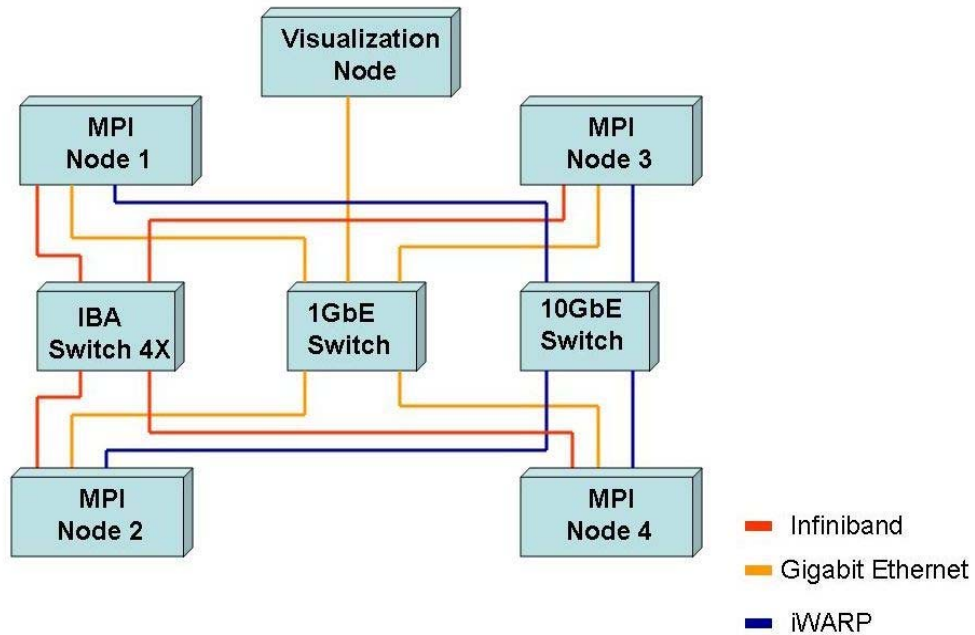
Intel® formed project teams to develop proof points of this new technology, originally shown at the Intel Developer Forum. The goal of the demonstration is to provide a vehicle that illustrates the unique and compelling benefits of RDMA-based transport to applications. At the highest level, the application sees a uniform transport neutral software stack. From that perspective, the application can utilize RDMA features such as zero copy and kernel bypass without regard to the underlying transport.

At the bottom of the OFA software stack are device drivers for the InfiniBand HCA and iWARP RNIC cards. Above the device drivers are the InfiniBand Connection Manager and iWARP Connection Manager. These layers provide transport independent connection management services that allow an MPI application to establish connections using IP addresses without regard to the underlying transport type. The OpenFabrics Verbs layer provides send/receive I/O services that allow applications to send and receive data in a transport independent fashion.

MPI-based demo

The demonstration hardware consists of four two-way Xeon nodes running the MPI demonstration application. The application Fluent runs over Intel MPI. In each node, there is a QLogic QLE7140 PCI Express InfiniBand HCA for InfiniBand connectivity and a NetEffect NE010e 10 Gigabit Ethernet (GbE) PCI Express ECA for iWARP connectivity. GbE connectivity will be provided through the on-board Ethernet interfaces. Three switches are utilized: a Cisco 4X InfiniBand switch, a Cisco 10GbE switch and a Cisco 1GbE switch. The InfiniBand and 10GbE switch ports use CX4 cabling, while the 1GbE switch uses standard Cat5 cabling. The InfiniBand interfaces on each MPI node are connected to the InfiniBand switch, and the iWARP interfaces on each MPI node are connected to the 10GbE switch. A fifth node is provided for visualization purposes. This node will be connected to the MPI cluster via the on-board 1GbE interfaces. Figure 2 illustrates the demonstration setup:

Figure 2



The MPI nodes run on an OpenFabrics-based software stack. The components of this stack are shown in Figure 3 below. At the bottom of the stack are device drivers for the QLogic InfiniBand HCAs and the NetEffect RDMA Network Interface Cards:

Figure 3

