Dominique A. Heger, Fortuitous Technology, (dom@fortuitous.com), Austin, TX, 2006

InfiniBand – Introduction, Performance, & Technology

In 2000, a new high-speed interconnect (InfiniBand) was announced that could and would replace all the other internal and external system interconnects (or at least that was the goal). The announcement was met with overwhelming silence. Then, on the heels of the announcement came the burst of the dot-com bubble, and InfiniBand was seemingly left for dead. Thankfully, the InfiniBand community was able to outlast the slow adoption rate, and quietly developed a very robust solution. Today, InfiniBand is widely deployed and provides high-speed (from 10Gbit/sec., to 60Gbit/sec. and potentially up to 120Gbit/sec.) interconnects. InfiniBand supports interconnections within or among servers (server-to-server connectivity), as well as among storage systems (server-to-storage and storage-to-storage connectivity). Major corporations are now introducing InfiniBand based solutions. To illustrate, Sun Microsystems announced that SUN's Utility Computing Grid will be based on InfiniBand, and IBM has presented an InfiniBand based solution for the IBM blade center (in collaboration with Topspin). Further, IBM supports InfiniBand for their General Parallel File System (GPFS) based clusters. In addition, SGI, integrated InfiniBand in its Altix servers. The user community can now take advantage of the provided InfiniBand features such as high bandwidth, extremely low latency periods, simple implementation, and low costs to build sound cluster interconnects or robust SAN solutions.

InfiniBand - Definition

InfiniBand represents a high-performance, switch-based interconnect architecture that is designed to operate within a server (component-to-component communication, replacing existing bus technologies), as well as an external interconnect solution (frame-to-frame communication for server or storage components). It offers a single interconnect for clustering, communication and storage purposes. Some of InfiniBand's key benefits are:

- Backed by industry standards, basically in the same way as T11 is responsible for the Fibre Channel standards, and the Internet Engineering Task Force (IETF) is looking over the iSCSI and Internet standards, the InfiniBand Trade Association is governing the InfiniBand standards.

- Max performance with minimal latency overhead. The current InfiniBand specification ranges from 10Gbit/sec. to 60G bit/sec. with very minimal latency overhead, which is typically measured in nanoseconds. To illustrate, nanoseconds are also used to measure read() and write() access times to physical memory.

- The InfiniBand *hardware transport* implements communication functionality's (that are traditionally performed by an operating system and the CPU) in the firmware of the InfiniBand device. This drastically reduces CPU overhead, and allows the host processor to allocate its CPU cycles to the user applications rather than on communication entities.

- Remote Direct Memory Access (RDMA). This capability allows systems to transfer data in main memory without involving the processor, cache or operating system of either compute node. RDMA transfers drastically reduce CPU overhead and again, allow the CPUs to spend their cycles serving application requests and not having to process communication requests.

InfiniBand – Usage & Application

Besides enabling higher performance within a server (replacing PCI buses with InfiniBand fabrics), InfiniBand introduces bandwidth capacities that are normally used for CPU-to-CPU and CPU-to-memory communications outside of the server. It allows physically separated systems to communicate with each other as if they would represent a single, large symmetric multiprocessing system. Traditional application clusters, such as Oracle RAC, rely on proprietary interconnects or TCP/IP to manage the complex nature of the inter-cluster traffic. The slow inter-process communication mechanism among the server nodes limits most cluster sizes to only a few nodes, as the overhead of a distributed lock manager on

an active cluster may cause the database processes to take a significant performance hit. TCP/IP over Gigabit Ethernet (GigE) introduces a rather significant load on the CPU (next to providing a limited bandwidth and higher latency than InfiniBand). Each additional node in the cluster introduces more overhead and traffic, limiting the aggregate performance potential of the cluster. IP-over-InfiniBand (IPoIB) eliminates this bottleneck. Application servers communicating with the cluster have similar issues. TCP/IP communications with the cluster may have a negative impact on performance (this is of course workload based). It is hard to justify a high-performance cluster setup if the application servers have only limited access (performance wise) to the data. InfiniBand connections between the cluster and the application servers eliminate the bottleneck. The RDMA feature of InfiniBand, using the Message Passing Interface (MPI), allows database servers in the cluster to directly read and write to each other's memory. This eliminates the TCP/IP and operating system overhead, which increases (1) the performance of each node and (2) the aggregate performance of the entire cluster. The IPoIB inter-processor communication allows the cluster to work on a single application similar to a large SMP system, without the notorious scalability issues fond on SMP systems.

Many of today's servers require three different network adapters to efficiently and effectively operate. (1) A GigE card for the LAN, (2) a Fibre Channel card for the SAN, and (3) a dedicated server-to-server clustering card (either proprietary or another GigE card). And in some instances, the cluster nodes may require an additional dedicated GigE card to connect to a backup network. In a blade server environment, providing each blade system with 3 cards introduces some issues, including increased power consumption, additional space requirements within the server, higher costs, greater complexity, and an increased heat issue. A blade server with an InfiniBand backplane eliminates the three cards. The backplane further enables RDMA capabilities to allow the blade servers to act as one unit (if necessary).

The adoption of InfiniBand in blade server systems and cluster solutions has been growing for some time. Clusters are designed to cost-effectively provide massive processing power. To truly deliver on that, they have to have high-performance and low-latency communication features (especially to the storage components). In order for InfiniBand-based cluster systems to utilize high-performance Fibre Channel storage, enhanced solutions have to be developed. One solution is to create a separate Fibre Channel SAN. And even though SAN bandwidth has recently doubled to 4Gbit/sec., separate Fibre Channel switches and host adapters create multiple network fabrics, increasing the cost and the complexity. The other option is an InfiniBand-to-Fibre Channel gateway. Such a solution has its drawbacks though. Gateways can become a performance bottleneck, they still are rather expensive, and they increase latency due to added protocol translation scenarios. As a result, the next step in the InfiniBand evolution is to tackle the storage subsystems. Native InfiniBand interfaces enable storage systems to attach directly to the existing InfiniBand fabric switches that are in use by the cluster, simplifying the network, and providing significant cost savings compared to Fibre Channel or gateway based solutions. Additionally, native InfiniBand interfaces require no InfiniBand-to-Fibre Channel protocol translations (within the controller), enabling higher performance and lower latency scenarios. InfiniBand's high (per-channel) bandwidth (which is more than double the 4Gbit/sec. Fibre Channel bandwidth) delivers maximum throughput across a minimum number of InfiniBand connections. This saves on switch ports and lowers the acquisition and service costs, respectively.

References

Alfaro, Sanchez, Duato, Das. "A Strategy to Compute the InfiniBand Arbitration Tables". In Int'l Parallel and Distributed Processing Symposium, April 2002.

Banikazemi, Govindaraju, Blackmore,Panda. "MPI-LAPI: An Efficient Implementation of MPI for IBM RS/6000 SP Systems". IEEE Transactions on Parallel and Distributed Systems, October 2001.

Carrera, Rao, Iftode, Bianchini. "User-Level Communication in Cluster-Based Servers." In Proceedings of the Eighth Symposium on High-Performance Architecture, February 2002.

Jni, "Introduction to InfiniBand" Jni Corporation, 2003

Liu, Wu, Panda, "High Performance RDMA-Based MPI Implementation over InfiniBand", Ohio State University, 2004