SOLUTION BRIEF

Connect-IB™: Architecture for Scalable High Performance Computing

The capabilities and scale of high performance computing (HPC) systems continues to grow. New systems are providing the ability to run simulations of scientific theory and industrial design at a far higher level of resolution and complexity. In addition, these systems can solve many problems that were not within reach even a few years ago, such as hurricane tracking and prediction, human heart and brain modeling or advanced problems in nuclear fission.

As these systems grow, the challenges placed on interconnect that provides the communication for the compute and storage of these systems continue to expand. With the ever increase of node counts, as well as the explosion of multi-core capabilities on a single node, the amount of system concurrency continues to increase, and with this the demands placed on the interconnect fabric also increase. For instance, the amount of system concurrency in 2011 was ~500,000 compute cores, while projections for Exascale systems in the 2018 timeframe are expected to be 1 billion cores, an increase of over 2000 times. This kind of scale requires interconnect that provides very high throughput, low latency and high message injection rates into the fabric. In addition the endpoints must be able to scale appropriately to handle the hundreds of thousands of connections that applications will need to communicate between these compute cores.

The Connect-IB architecture has been developed specifically with these kinds of requirements and characteristics in mind. It's highly parallel design enables high performance for high concurrent core counts at the endpoints, and its unique scalability features allow these end points to scale at tens to hundreds of thousands of nodes. This solution brief will highlight some of the key features incorporated in the Connect-IB family of adapters to meet the high demands of today's and future high performance computing requirements.

High Throughput

Connect-IB is the first InfiniBand adapter on the market that enables 100Gb/s uni-directional throughput (200 Gb/s bi-directional throughput) by expanding the PCI Express 3.0 bus to 16-lanes and through dual 56Gb/s FDR InfiniBand network ports. In addition, the internal data path of the device can also deliver over 100Gb/s data throughput. Thus, MPI and other parallel programming languages can take advantage of this high throughput, utilizing the multi-rail capabilities built into the software. While Mellanox ConnectX®-3 adapters provided applications running on the Intel Sandy Bridge systems to realize the full capabilities and bandwidth of the PCI Express 3.0 x8 bus, Connect-IB adapters increase these capabilities, which is important for bandwidth sensitive applications, and even more critical with the advent of increased CPU cores such as the new Intel Ivy Bridge systems. In addition, this level of throughput will be required to satisfy the needs to new heterogeneous environments such as GPGPU and Intel Xeon PHI based endpoints.

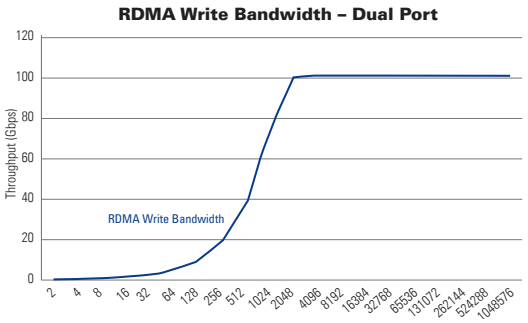

Figure 1. *100Gb/s RDMA Write Performance*

High Message Rate

Many HPC applications are based on communications patterns that use many small messages between parallel processes within the job. It is critical that the interconnect used to transport these messages provides low latency and high message rate capabilities to assure that there are no bottlenecks to the application. The new Connect-IB architecture provides an increase in the message rate of previous InfiniBand offerings by over 4 times. Connect-IB can deliver over 137 million single packets (non-coalesced), native InfiniBand messages per second to the network. This increase assures that there are no message rate limitations for applications and that the multiple cores on the server will communicate to other machines as fast as the cores are capable, without any slowdown from the network interface.

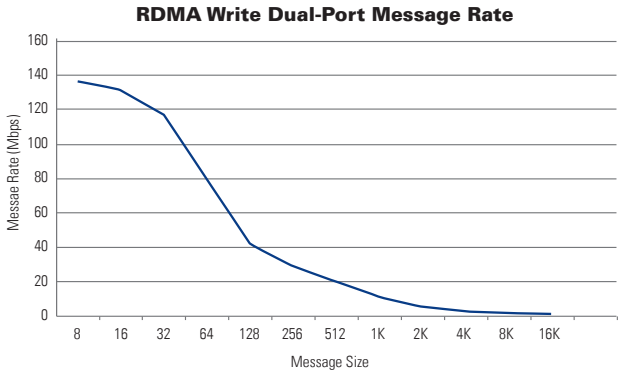

Figure 2. *Message Rate Example*

Dynamically Connected Transport Service

One of the major strengths of InfiniBand is the capabilities it enables with Reliable Connection (RC) Transport Services. These provide a number of advantages for parallel computing communications including end-to-end reliability performed by the adapter hardware, full transport offload, large send/receive messages, and the capability of remote memory access through RDMA. Because the RC Transport service requires connections to be established between the two endpoints of the connection, context for these connections must be established and stored on the endpoints. The amount of context grows with the size of the job, and thus the amount of connections that need to be established grows. At extreme large scale this can cause a higher amount of memory consumption on the endpoint host's memory. It can also affect performance of the endpoint at large scale when the adapter resources become heavily used and context retrieval from the system memory to adapter happens at a higher frequency.

As InfiniBand has evolved, and as cluster sizes have grown, new transport mechanisms have been introduced by Mellanox Technologies to enable the size of connections to grow and be handled more efficiently within the adapter. This includes Shared Receive Queue (SRQ) support, introduced in 2004, which enables the receive size connections to share a single receive queue for data buffers on the receive size, allowing for more efficient handling of buffer space in system memory.

With RC connections, a single connection is established between every CPU core and every other CPU core within the running application. This means that each endpoint will hold P2*N connections (where P=PPN or processors per node, and N=number of nodes participating in the job). Thus, a 16-core machine running across 256 nodes would require 65,536 connections on each endpoint. The Extended Reliable Connected Transport Service or XRC, introduced in 2007, provides changes in the transport layer mechanism to reduce the number of connections to P*N, so in the same example above the number of connections per endpoint is reduced down to 4096. However, with all of these optimizations, the adapter resources and host memory consumption are still related to the system size in the matter of number of compute nodes.

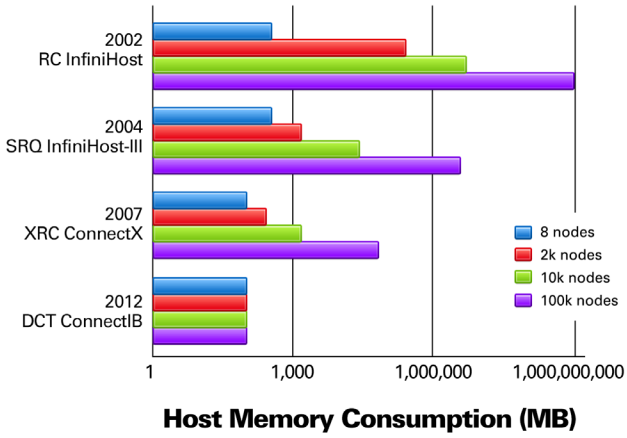

Figure 3. *Memory Resource Usage for various Transport Types*

The Connect-IB architecture solves any limitations posed by the number of connections needed for a given job by using a new transport mechanism called Dynamically Connected (DC) Transport Service. The concept with DC Transport is that connections are not nailed down at the beginning of the job, but instead are setup and torn down dynamically by the adapter hardware on an 'as needed' basis. A simplified way of looking at it is that the adapter device holds a pool of 'connection resources' to use on an 'as needed' basis whenever the upper layer software wishes to send data to another process within the cluster. Once the software queues the data to be sent, the hardware grabs a context set from the pool of resources and establishes a connection with the receive side device 'in-band' with the data being sent (so no latency overhead is taken). Once the data packets for this particular message

have been sent and the data has been acknowledged, the connection is torn down by the hardware, and the connection context entry is added back into the pool of connection resources for other data transfers to use in the future. What this means is that the number of connection resources needed by the adapter is dependent only on the number of messages a single server can hold at a given point of time and is completely decoupled from the size of the cluster. Now the size of the cluster is irrelevant to connection resource usage, allowing for basically unlimited scalability for connection based communications.

Summary

With the performance throughput gains, outstanding message rate capabilities, and the new scalable Dynamically Connected Transport service, Connect-IB is poised to solve the interconnect challenges of today's and tomorrow's toughest supercomputing requirements. The architecture is built from the ground up to remove bottlenecks and provide large scale interconnect for the largest sized and most demanding clusters in the world today and in the future.

CONNECT-IB: PRODUCT FEATURES

INFINIBAND
- IBTA Specification 1.2.1 compliant
- FDR 56Gb/s InfiniBand
- Hardware-based congestion control
- 16 million I/O channels
- 256 to 4Kbyte MTU, 1Gbyte messages

ENHANCED INFINIBAND
- Hardware-based reliable transport
- Extended Reliable Connected transport
- Dynamically Connected transport service
- Signature-protected control objects
- Collective operations offloads
- GPU communication acceleration
- Enhanced Atomic operations

STORAGE SUPPORT
- T10-compliant DIF/PI support
- Hardware-based data signature handovers

FLEXBOOT™ TECHNOLOGY
- Remote boot over InfiniBand

HARDWARE-BASED I/O VIRTUALIZATION
- Single Root IOV*
- Up to 16 physical functions, 256 virtual functions
- Address translation and protection
- Dedicated adapter resources
- Multiple queues per virtual machine
- Enhanced QoS for vNICs and vHCAs
- VMware NetQueue support

PROTOCOL SUPPORT
- OpenMPI, IBM PE, Intel MPI, OSU MPI (MVAPICH/2), Platforms MPI, UPC, Mellanox SHMEM
- TCP/UDP, IPoIB, RDS
- SRP, iSER, NFS RDMA, SMB Direct
- uDAPL

Ordering Part Number	InfiniBand Ports	PCI Express
MCB191A-FCAT	Single FDR 56Gb/s	3.0 x8
MCB192A-FCAT	Dual FDR 56Gb/s	3.0 x8
MCB193A-FBAT	Single FDR 56Gb/s	2.0 x16
MCB193A-FCAT	Single FDR 56Gb/s	3.0 x16
MCB194A-FCAT	Dual FDR 56Gb/s	3.0 x16

Table 1. *Connect-IB Product Information*

*Future Support

Mellanox TECHNOLOGIES
350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com