WHITE PAPER

FDR InfiniBand is Here

All About Lowest Latency and Highest Scalability

The high-speed InfiniBand server and storage connectivity has become the de facto scalable solution for systems of any size – ranging from small, departmental-based compute infrastructures to the world's largest PetaScale systems. The rich feature set and the design flexibility enable users to deploy the InfiniBand connectivity between servers and storage in various architectures and topologies to meet performance and or productivity goals. These benefits make InfiniBand the best cost and performance solution when compared to proprietary and Ethernet-based options.

According to the June 2011 TOP500 List of the world's top supercomputers, InfiniBand (all Mellanox based solutions) is used as the server interconnect solution for five of the top 10 supercomputers, all of which are Petascale systems, as well as the storage interconnect solution for some of the top 10 proprietary-based systems (including the fastest supercomputer on the list). Furthermore, InfiniBand connects the majority of the systems in the top 100 supercomputers, top 200, 300 and even 400. See Figure 1.

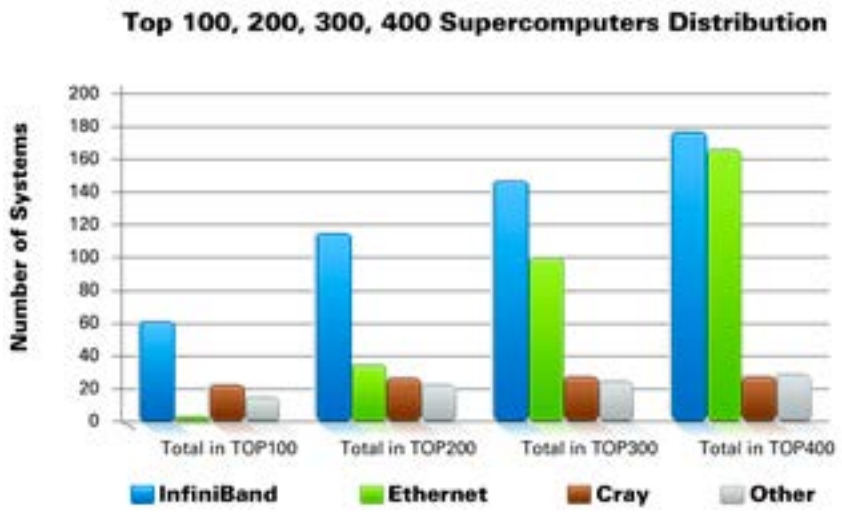

Figure 1. Top 100, 200, 300, 400 supercomputers distribution according to their interconnect as reported by the June 2011 TOP500 list (www.top500.org)

The InfiniBand technology is developing rapidly. See Figure 2. SDR InfiniBand (10Gb/s) solutions were introduced to the market in 2002, DDR InfiniBand (20Gb/s) solutions were introduced in 2005, QDR InfiniBand (40Gb/s) solutions were introduced in 2008 and now FDR InfiniBand (56Gb/s) solutions have been introduced in June 2011. This translates to a new InfiniBand speed with enhanced capabilities set every three years to support the compute (CPU/GPU) development over time and the increasing performance demands from HPC applications and users.

As a standard-based solution, InfiniBand offers backward compatibility between the different speeds. This allows users to expand their system and to connect newer parts to older existing ones. For example, NASA Ames (USA) deployed one of the world's PetaScale systems which consist of more than 11,000 nodes installed in multiple phases. Their system utilizes InfiniBand DDR and QDR technologies from Mellanox.

Figure 2. InfiniBand technology development over time

The newest edition to the InfiniBand technology is FDR InfiniBand 56Gb/s. In June 2011, Mellanox became the first company to announce and offer FDR InfiniBand solutions. As anticipated, Mellanox also announced multiple design wins with the new InfiniBand technology. Newer generations of HPC systems can now reap the benefits from utilizing the fastest and most scalable InfiniBand interconnect solution on the market.

InfiniBand FDR is all about lower latency and higher scalability and reliability. From a technology perspective, Mellanox FDR InfiniBand includes the following new capabilities:

- Link speed increase from 10Gb/s per InfiniBand lane to 14.0625Gb/s per InfiniBand lane, or 56Gb/s per InfiniBand port
- Data link encoding modification from 8/10 bits to 64/66 bits
- Forward Error Correction (FEC) between switches and between switches and adapters
- Integration of InfiniBand to Ethernet gateway within the InfiniBand switch
- Lower power consumption

The new data link encoding increases the InfiniBand network efficiency by more than 21%. However, from an engineering perspective, the 64/66 data link encoding causes a slight latency increase and the FDR switch latency is 200 nanoseconds. Nonetheless, this is still faster than any other switch vendor solution and in particular faster than any non-Mellanox QDR (with 8/10 data bits encoding) switch solution by more than 10%. FDR's accelerated speed enables faster data delivery by more than 70%, resulting in a dramatic reduction in the overall InfiniBand fabric latency.

For data integrity and guaranteed reliable data transfer between end-nodes (servers and storage), InfiniBand uses an end-to-end hardware reliability mechanism. Each InfiniBand packet contains two Cyclic Redundancy Checks (CRCs). The Invariant CRC (ICRC) covers all fields which do not change as the packet traverses the fabric. The Variant CRC (VCRC) covers the entire packet. The combination of the two CRCs allows switches and routers to modify appropriate fields and maintain end-to-end data integrity. If a data corruption occurs due to Bit Error Rate (BER), the packet will be discarded by the switch or the adapter, and will be re-transmitted by the source to the target. In order to accelerate the data retransmission, a new mechanism was added to InfiniBand FDR - Forward Error Correction (FEC). FEC allows the InfiniBand devices (adapters and switches) to fix bit errors throughout the network and reduce the overhead for data re-transmission between the end-nodes. The FDR InfiniBand FEC mechanism utilizes redundancy in the 64/66-bits encoding to enable error correction with no bandwidth loss and has the ability to work over each link independently, on each of the link lanes. The new mechanism delivers superior network reliability, especially for large scale data centers, high-performance computing or web 2.0 centers, and ensures a predictable low-latency characteristic, critical for large scale applications and synchronizations. See Figure 3.

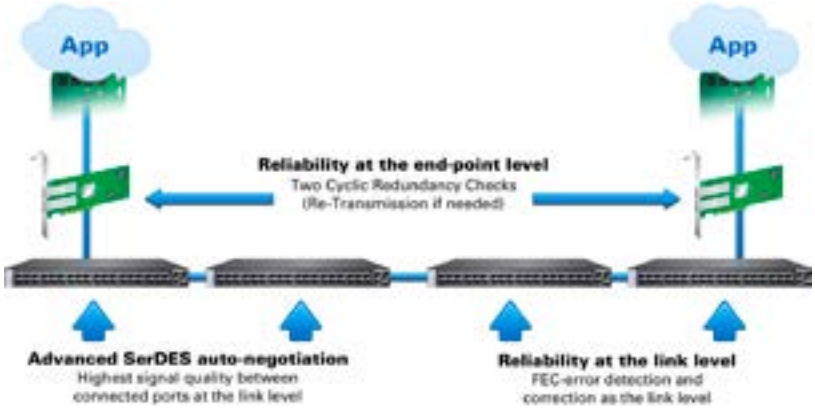

Figure 3. Mellanox InfiniBand reliability mechanisms

The Mellanox FDR InfiniBand switch product line introduces new consolidated fabric elements for higher scalability and fabric consolidation. The seamless integration of the bridges yields a simplified, yet high-performance connectivity from the InfiniBand network to legacy or other networks, resulting in substantial CAPEX and OPEX savings.

The newly introduced FDR InfiniBand switches and cables brings forth low latency, high-performance, scalable, efficient and reliable interconnect solutions for connecting servers and storage. The performance, scalability and economical advantages delivered by FDR InfiniBand increase applications productivity and optimize the return on investment. FDR InfiniBand is the best interconnect solution for high-performance and data center clusters today.

Mellanox TECHNOLOGIES

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