TECHNOLOGY BRIEF

May 2010

CORE-Direct
The Most Advanced Technology for MPI/SHMEM Collectives Offloads

To meet the needs of scientific research and engineering simulations, supercomputers are growing at an unrelenting rate. As supercomputers increase in size from mere thousands to hundreds-of-thousands of processor cores, new performance and scalability challenges have emerged. In the past, performance tuning of parallel applications could be accomplished fairly easily by separately optimizing their algorithms, communication, and computational aspects. However, as we continue to scale to larger machines, these issues become co-mingled and must be addressed comprehensively. Collectives communications have a crucial impact on the engineering and scientific application's scalability as they frequently are being used by applications for operations such as broadcasts for sending around initial input data, reductions for consolidating data from multiple sources and barriers for global synchronization. Collectives communications execute global communication operations to couple all processes/nodes in the system and therefore must be executed as quickly and as efficiently as possible. Current implementations of collectives operations are not scalable, suffer from systems noise and consume a major part of the CPU time.

Mellanox ConnectX®-2 adapters and InfiniScale® IV switches address the collectives communication scalability problems by offloading the communications to the adapters and switches. The technology named CORE-Direct (Collectives Offload Resource Engine) provides the most advanced solution for handling collectives operations, ensures maximum scalability, minimizes the CPU overhead and provides the capability for overlapping communications with computations.

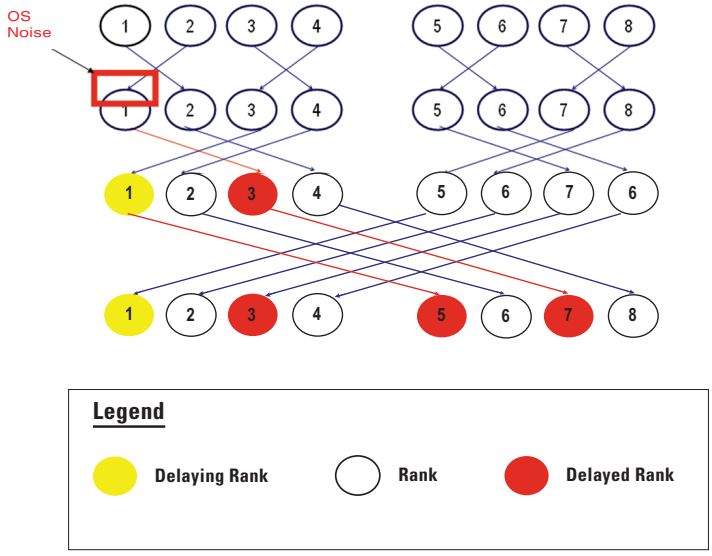
CORE-Direct Adapter Capabilities	CORE-Direct Switch Capabilities
Handle collectives operations in wire speed and in reliable manner, support unlimited in-flight collectives support	World highest Multicast operations/groups support
Handle local node and remote nodes communications	World lowest switch latency
Ability to perform floating–point operations	Programmable routing/adaptive routing
Flexibility to support different collectives topologies	
Support for blocking and non-blocking collectives operations (MPI-3)	

Mellanox adapters and switches are well suited to offload communication patterns and allow overlap of non-blocking collectives operations with application computation. Offloading communication processing to the HCA provides the capability to minimize the effects of system noise and to increase the communication–computation overlapping significantly.

System Noise Reduction

Operating system noise is one of the well known obstacles to large-scale application scalability. The CPU has to perform both computational and communication tasks, and it typically performs them sequentially. Collectives operations present a significant overhead to CPU run-time and therefore OS noise introduces additional delays to the overall execution time of CPU tasks.

The impact of OS noise on the application performance in an MPI operation is illustrated below. In the example, OS noise on rank 1 (or node 1) delays the arrival of the expected message from rank 2. This delay not only increases the execution time of rank 1, but it also increases the execution time of ranks 3, 5 and 7, which are indirectly affected by it, and therefore impact the entire application run time and reduces the entire system performance.



Communication and Computation Overlapping

Offloading collectives communication operations to the HCA relieves the CPU from the communication management workload, allowing an increase in collectives communication and computations overlapping/parallelism. Further gain in the overlap is achieved by having the HCA run part of the calculations instead of the CPU. ConnectX-2 includes a floating point operation unit, which enables offloading the data manipulation part of MPI_Reduce and MPI_Allreduce collectives operations to the HCA.

Mellanox CORE-Direct

Mellanox CORE-Direct technology provides the most complete and advanced solutions for offloading the MPI collectives operations from the software library to the network. CORE-Direct not only accelerates MPI applications but also solves the scalability issues in large scale systems by eliminating the issues of OS noise and jitter. Mellanox solutions offload the entire collectives communications – the collectives progress throughout in a reliable manner the collectives data manipulation by providing floating point engines within the adapters. Mellanox CORE-Direct is the only scalable, reliable and most efficient solution for MPI and SHMEM collectives operations.

Mellanox TECHNOLOGIES

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