SOFTWARE

PRODUCT BRIEF

Fabric Collective Accelerator™ (FCA)

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 systems continue to scale to larger machines, these issues become co-mingled and must be addressed comprehensively.

Collective communications have a crucial impact on the scalability of large scale applications as they are frequently being used by high-performance computing (HPC) applications for operations such as broadcasts for sending around initial input data, reductions for consolidating data from multiple sources and barriers for global synchronization.

The Message Passing Interface (MPI) library or the Shared Memory (SHMEM) environments are two examples of libraries that provide implementations of collective communications for the usage of HPC applications.

Collective 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. Indeed, the scalability of most scientific and engineering applications is bound by the scalability and performance of the collectives routines employed. System noise increases the latency of collectives operations by amplifying the effect of small, randomly occurring OS interrupts during the collectives progression, therefore most current implementations of collectives operations will suffer from the effects of systems noise at extreme-scale. Collectives operations can also consume a significant amount of CPU cycles that could be otherwise spent doing meaningful computation further hampering application scalability.

Mellanox Technologies has addressed lost performance from the effects of system noise, by offloading the communications to the host channel adapters (HCAs) and switches. The fundamental technology, named CORE-Direct® (Collectives Offload Resource Engine), provides the most advanced solution available for handling collectives operations thereby ensuring maximum scalability, minimal CPU overhead, and providing the capability to overlap communication operations with computation allowing applications to maximize asynchronous communication.

Offloading the Collective Communications

Mellanox InfiniBand adapters and switches address the collective communication scalability problem by offloading the collective communication to the network. CORE-Direct® (Collectives Offload Resource Engine) is a hardware technology that is part of the Mellanox ConnectX®-2 or later adapters and which provides sophisticated hardware resources for complete offloading of the

HIGHLIGHTS

FEATURES

- Offload collectives communication from MPI process into Mellanox interconnect hardware
- Efficient collectives communication flow optimized to job and topology
- Monitor performance of collectives operations
- Support for blocking and nonblocking collectives
- Supports hierarchical communication algorithms (HCOL)
- Supports multiple optimizations within a single collective algorithm
- Thread safe

BENEFITS

- Significantly reduce MPI collectives runtime
- Increase CPU availability and efficiency for increased application performance
- Improved collectives function scalability beyond any proprietary interconnect

collectives communications. CORE-Direct-technology offloads the entire collectives communication in a reliable manner, as well as managing the data manipulations (reduction based collectives for example) that are part of the collectives communications using a fast, embedded within the adapter, floating point unit engines. CORE Direct technology is available for 3rd party software implementations through the CORE-Direct API that exists with Mellanox InfiniBand drivers. Mellanox CORE-Direct is the only scalable, reliable and most efficient solution for MPI and SHMEM collectives operations.

FCA – Easy to Integrate MPI Collectives Package

FCA is a Mellanox MPI-integrated software package that utilizes CORE-Direct technology for implementing the MPI collective communications. FCA can be used with all major commercial and open-source MPI solutions that exist and being used for high-performance applications. FCA with CORE-Direct technology accelerates the MPI collectives runtime, increases the CPU availability to the application and allows overlap of communications and computations with asynchronous collectives operations. Figure 1 shows the system configuration with FCA and CORE-Direct, and Figure 2 shows the software layers.

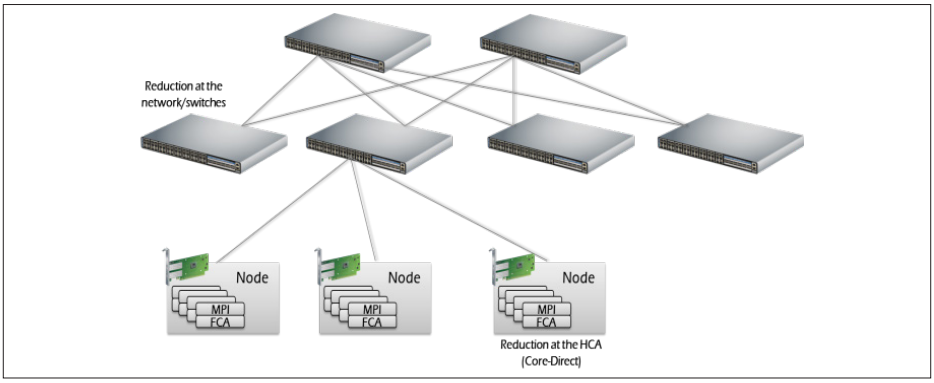

Figure 1. *HPC system architecture with FCA and CORE-Direct*

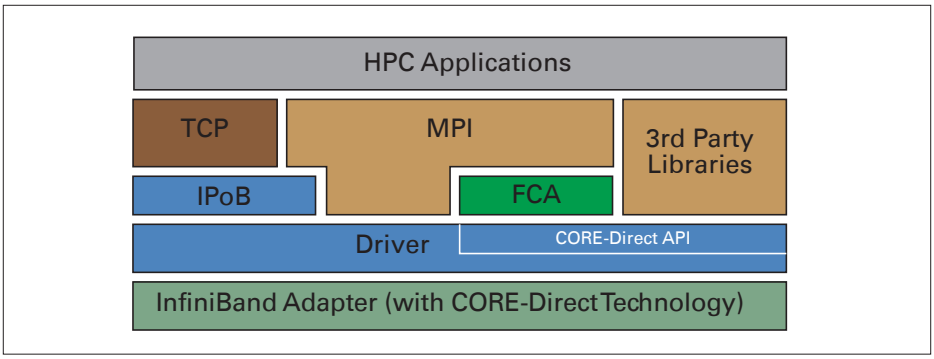

Figure 2. *Software architecture with FCA and CORE-Direct*

Benefits often can be seen immediately from CORE-Direct right out-of-the-box by simply specifying the necessary BCOL/SBGP combinations. In order to take maximum advantage of CORE-Direct, users may modify their applications to use MPI 3.0 non-blocking routines while using CORE-Direct to offload the collectives "under-the-covers", thereby allowing maximum opportunity to overlap communication with computation.

FCA 3.0 contains support to build runtime configurable hierarchical collectives. Also the ability to accelerate collectives with hardware multicast continues to be supported. Additional performance and scalability of Mellanox's advanced point-to-point library (MXM) is exposed that allows users to take full advantage of the new features with minimal effort.

FCA for Higher Return-on-Investment

Mellanox InfiniBand interconnect solutions for server and storage systems provide the highest performance and efficiency for HPC applications. FCA and CORE-Direct helps to further accelerate the application performance, increase the CPU efficiency and future proof the system architecture.

Figure 3 shows the FCA benefits even at low scale, and of course, the larger the system the effects that FCA will demonstrate will be even greater. Figure 3 shows how FCA accelerates OpenFOAM (open source CFD application) by 58% with as few as 16 nodes.

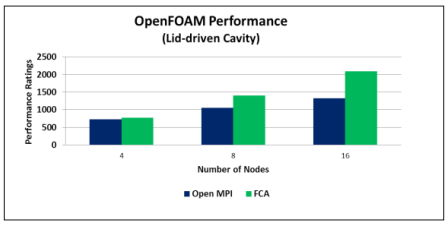

Figure 3. *OpenFOAM performance with FCA*

For more info please contact hpc@mellanox.com, or visit www.mellanox.com

Mellanox TECHNOLOGIES

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

15-3721PB Rev 1.1