Mellanox Co-Design Strategy

Mellanox Co-Design Strategy

Server/Storage

Application

Communication Libraries

Network

Server/Storage

Application

MPI SHMEM/PGAS

Bottlenecks

Throughput Latency Message Rate Hardware Offloads

© 2012 MELLANOX TECHNOLOGIES

- MELLANOX CONFIDENTIAL -
Mellanox Co-Design Strategy

Server/Storage

Application

Communication Libraries

Network

Server/Storage

Application

Extension of I/O communications (RDMA, collectives, synchronization etc)

Throughput
Latency
Message Rate
Hardware Offloads

MPI
SHMEM/PGAS
Mellanox ScalableHPC Accelerate Parallel Applications

**MXM**
- Reliable Messaging Optimized for Mellanox HCA
- Hybrid Transport Mechanism
- Efficient Memory Registration
- Receive Side Tag Matching

**FCA**
- Topology Aware Collective Optimization
- Hardware Multicast
- Separate Virtual Fabric for Collectives
- CoreDIrect Hardware Offload

InfiniBand Verbs API
Mellanox Scalable SHMEM Solution with FCA

**Barrier Collective**

- **Latency (μs)**
- Processes (PPN=8)
  - Without FCA
  - With FCA

**Reduce Collective**

- **Latency (μs)**
- Processes (PPN=8)
  - Without FCA
  - With FCA

**8-Byte Broadcast**

- **Bandwidth (KB/processes)**
- Processes (PPN=8)
  - Without FCA
  - With FCA
Mellanox ScalableSHMEM

Overview

The SHMEM programming library is a one-side communications library that supports a unique set of parallel programming features including point-to-point and collective routines, synchronizations, atomic operations, and a shared memory paradigm used between the processes of a parallel programming application.

Mellanox ScalableSHMEM 2.0 is based on the API defined by the OpenSHMEM.org consortium. The library works with the OpenFabrics RDMA for Linux stack (OFED), and also has the ability to utilize Mellanox Messaging libraries (MXM) as well as Mellanox Fabric Collective Accelerators (FCA), providing an unprecedented level of scalability for SHMEM programs running over InfiniBand.

Features
- Provides a programming library for shared memory communication model extending use of InfiniBand to SHMEM applications
- Seamless integration with MPI libraries and job schedulers allowing for Hybrid programming model
- Maximum collective scalability through integration with Mellanox Fabric Collective Accelerator (FCA)
- High message rate performance with integration and Mellanox Messaging Accelerator (MXM)
- Thanks to Steve Poole, ORNL and UT-Battelle for support of this project.
Thank You
HPC@mellanox.com

PAVING THE ROAD TO EXASCALE
ADVANCING NETWORK PERFORMANCE, EFFICIENCY, AND SCALABILITY.