Interconnect Your Future

Scot Schultz, Director, HPC and Technical Marketing
HPC Advisory Council, European Conference, June 2014
Leading Supplier of End-to-End Interconnect Solutions

Virtual Protocol Interconnect
56G IB & FCoIB

10/40/56GbE & FCoE

Virtual Protocol Interconnect
56G InfiniBand

10/40/56GbE

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

<table>
<thead>
<tr>
<th>ICs</th>
<th>Adapter Cards</th>
<th>Switches/Gateways</th>
<th>Host/Fabric Software</th>
<th>Metro / WAN</th>
<th>Cables/Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2014 Mellanox Technologies
The Speed of Data Mandates the Use of Data

Mellanox Roadmap of Data Speed

- 10Gbs
- 20Gbs
- 40Gbs
- 56Gbs
- 100Gbs
- 200Gbs

Keeping You One Generation Ahead

Enabling the Use of Data
Connect-IB

Architectural Foundation for Exascale Computing
Mellanox Connect-IB The World’s Fastest Adapter

- The 7th generation of Mellanox interconnect adapters
- World’s first 100Gb/s interconnect adapter (dual-port FDR 56Gb/s InfiniBand)
- Delivers 137 million messages per second – 4X higher than competition
- Support the new innovative InfiniBand scalable transport – Dynamically Connected
Connect-IB Provides Highest Interconnect Throughput

Higher is Better

Unidirectional Bandwidth

Bidirectional Bandwidth

Gain Your Performance Leadership With Connect-IB Adapters

Source: Prof. DK Panda
Connect-IB Delivers Highest Application Performance

200% Higher Performance Versus Competition, with Only 32-nodes
Performance Gap Increases with Cluster Size
Mellanox PeerDirect™
What is PeerDirect™

PeerDirect is natively supported by Mellanox OFED 2.1 or later distribution.
Supports peer-to-peer communications between Mellanox adapters and third-party devices.

No unnecessary system memory copies & CPU overhead
- No longer needs a host buffer for each device
- No longer needs to share a host buffer either

- Provides native support for Intel Xeon PHI MPSS communication stack
- Supports NVIDIA® GPUDirect RDMA with a separate plug-in, **available at Mellanox.com**
- Support for RoCE protocol over Mellanox VPI

Supported with all Mellanox ConnectX-3 and Connect-IB Adapters
Mellanox PeerDirect™ Architecture

- A generic infrastructure to allow PCIe peer devices to communicate directly
- Enables direct RDMA transactions between PCIe device and Mellanox interconnect

**P2P Plugin**

**Peer Driver**

**HCA Driver**

**EXPORT Peer Device Memory Functions**

ib_umem_* functions are “tunneled” thru the p2p plugin module
GPUDirect RDMA

Native support for peer-to-peer communications between Mellanox HCA adapters and NVIDIA GPU devices
Before GPUDirect

- Network and third-party device drivers, did not share buffers, and needed to make a redundant copy in host memory.

With GPUDirect 1.0 Shared Host Memory Pages

- The network and GPU can share “pinned” (page-locked) buffers, eliminating the need to make a redundant copy in host memory.
Eliminates CPU bandwidth and latency bottlenecks
Uses remote direct memory access (RDMA) transfers between GPUs
Resulting in significantly improved MPI Send-Recv efficiency between GPUs in remote nodes
Note: A requirement for GPUDirect RDMA to work properly is that the NVIDIA GPU and the Mellanox InfiniBand Adapter share the same root complex.
Performance of MVAPICIH2 with GPUDirect RDMA

**GPU-GPU Internode MPI Latency**

- **Lower is Better**
- **67% Lower Latency**
- **5.49 usec**

**GPU-GPU Internode MPI Bandwidth**

- **Higher is Better**
- **5X Increase in Throughput**

Source: Prof. DK Panda

© 2014 Mellanox Technologies
Mellanox PeerDirect™ with NVIDIA GPUDirect RDMA

- HOOMD-blue is a general-purpose Molecular Dynamics simulation code accelerated on GPUs
- GPUDirect RDMA allows direct peer to peer GPU communications over InfiniBand
  - Unlocks performance between GPU and InfiniBand
  - This provides a significant decrease in GPU-GPU communication latency
  - Provides complete CPU offload from all GPU communications across the network
- Demonstrated up to 102% performance improvement with large number of particles

**HOOMD-blue Performance**  
(LJ Liquid Benchmark, 16K Particles)

**HOOMD-blue Performance**  
(LJ Liquid Benchmark, 512K Particles)
HPC-X
Mellanox Optimized Software Stack Components

Achieve Maximum Application Scalability and Performance
Future proofed software compatibility: SDR-DDR-QDR-FDR-EDR

Applications

Communication Libraries
MPI, SHMEM, UPC

Mellanox OFED®,
PeerDirect™, Core-Direct™, GPUDirect® RDMA

Linux / Windows Operating System

ConnectX-3, ConnectX-3 Pro
10/40/56 Gb/s Ethernet Support for RoCE

ConnectX-3 and Connect-IB
InfiniBand®

x86, Power8, ARM

© 2014 Mellanox Technologies
Mellanox HPC-X™ Components and Accelerators

Mellanox ScalableMPI - Message Passing Interface based on Open MPI
Mellanox ScalableSHMEM - One-sided Communications Library
Mellanox ScalableUPC – Berkeley UPC parallel programming language library
Mellanox MXM – Messaging Accelerator optimized for underlying hardware
Mellanox FCA – Fabric Collectives Accelerator supporting MPI-3
Mellanox OFED – Unlocks additional capabilities of the interconnect architecture

Profiling Tools …Benchmarking Tools….and MORE…
- Complete MPI/OpenSHMEM/PGAS/UPC package for HPC environments
  - Fully optimized for Mellanox InfiniBand and VPI interconnect solutions
  - Supports 3rd party solutions

- Components
  - Communication libraries: ScalableMPI, ScalableSHMEM, ScalableUPC
  - Acceleration libraries: MXM – Messaging Accelerator, FCA – Fabric Collectives Accelerator
  - Tools: Integrated Performance Monitoring Tool (IPM), benchmarks etc.

### 20%-70% Performance Improvement

**OpenFOAM Performance**
- Lid-driven Cavity

**LS-DYNA Benchmark**
- 3 Vehicle Collision

**CPMD Performance**
- S152 inp-1

- **CFD**
  - Performance Improvement
  - Number of Nodes: 4, 8, 16

- **CAE**
  - Performance Improvement
  - Number of Nodes: 4, 8, 16

- **Bio**
  - Performance Improvement
  - Number of Nodes: 1, 2, 4, 8, 16
Mellanox ScalableHPC Accelerates Parallel Applications

### InfiniBand Verbs API

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

**Mellanox FCA**
- Topology Aware Collective Optimization
- Hardware Multicast
- Separate Virtual Fabric for Collectives
- CORE-Direct Hardware Offload

---

Memory

MPI

P1 → P2 → P3

Memory

SHMEM

P1 → P2 → P3

Logical Shared Memory

Memory

PGAS

P1 → P2 → P3

Logical Shared Memory

Memory
Mellanox FCA Collective Scalability

- Increases CPU efficiency for increased scalability
- Support for blocking and nonblocking collectives
- Supports hierarchical communication algorithms (HCOL)
- Supports multiple optimizations within a single collectives algorithm
- Minimize the negative effect of system noise and jitter
Summary
End-to-end Interconnect Solutions to Deliver Highest ROI for all Applications

Maximum Application Scalability and Performance

World’s first 100Gb/s interconnect adapter

Increased Scalability with FCA

Dominant in Storage Interconnects