Advances in High Performance Computing

Brian Forbes – Sr. System Engineer
Mellanox HPC Technology

- **Performance**: 100Gb/s network injection rate, under 1us end-end latency
- **Scalability**: Flat, Fat and Fast. Powering 8 of the TOP20 Petascale Supercomputers
- **Advanced Features**: Collective Communication Offloads, Flexible Topologies
- **Big Data**: Boost Hadoop Performance with Direct InfiniBand RDMA Access

**Comprehensive End-to-End 10/40/56Gb/s Ethernet and 56Gb/s InfiniBand Portfolio**

<table>
<thead>
<tr>
<th>ICs</th>
<th>Adapter Cards</th>
<th>Switches/Gateways</th>
<th>Software</th>
<th>Cables</th>
</tr>
</thead>
</table>

**Scalability, Reliability, Power, Performance**
FDR InfiniBand New Features and Capabilities

**Performance / Scalability**
- >12GB/s bandwidth, <0.7μsec latency
- PCI Express 3.0
- InfiniBand Routing and IB-Ethernet Bridging

**Reliability / Efficiency**
- Link bit encoding – 64/66
- Forward Error Correction
- Lower power consumption
FDR Application Benchmarks

Computational Fluid Dynamics
Oil and Gas Reservoir Simulation
Molecular Dynamics
Weather and Earth Sciences

Up to 32% ROI on equipment and operating costs
SSDs, the storage hierarchy, In-Memory Computing…..

Remote I/O access needs to be equal to local I/O access

Native Throughput Performance over InfiniBand FDR
InfiniBand Verbs API

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

**SHMEM**
- Logical shared memory
- Memory

**PGAS**
- Logical shared memory
- Memory
- Memory

**MXM**
- Topology Aware Collective Optimization
- Hardware Multicast
- Separate Virtual Fabric for Collectives
- CoreDirect Hardware Offload

**FCA**
- Reliable Messaging Optimized for Mellanox HCA
- Hybrid Transport Mechanism
- Efficient Memory Registration
- Receive Side Tag Matching
Fabric Collective Accelerations Provide Linear Scalability

Barrier Collective

Reduce Collective

8-Byte Broadcast
Double Hadoop Performance with UDA

Terasort Benchmark*
(20GB file size, 16GB Data per node, 8 Mappers, 4 Reducers, 4 Disks)

Lower is better

Disk Writes 40%
Disk Reads 15%
CPU Utilization 2.5X

*TeraSort is a popular benchmark used to measure the performance of Hadoop cluster

~2X Faster Job Completion!
# Roadmap of Interconnect Innovations

<table>
<thead>
<tr>
<th>Year</th>
<th>InfiniHost</th>
<th>InfiniHost III</th>
<th>ConnectX (1,2,3)</th>
<th>Connect-IB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>World’s first InfiniBand HCA</td>
<td>10Gb/s InfiniBand PCI-X host interface 1 million msg/sec</td>
<td>20Gb/s InfiniBand PCIe 1.0 2 million msg/sec</td>
<td>The Exascale Foundation</td>
</tr>
<tr>
<td>2005</td>
<td>World’s first PCIe InfiniBand HCA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008-11</td>
<td>ConnectX (1,2,3)</td>
<td>World’s first Virtual Protocol Interconnect (VPI) Adapter</td>
<td>40/56Gb/s InfiniBand PCIe 2.0, 3.0 x8 33 million msg/sec</td>
<td></td>
</tr>
</tbody>
</table>
Announcing Connect-IB: The Exascale Foundation

- A new interconnect architecture for compute intensive applications
- World’s fastest server and storage interconnect solution providing 100Gb/s injection bandwidth
- Enables unlimited clustering scalability with new Dynamically Connected Transport service
- Accelerates compute-intensive and parallel-intensive applications with over 130 million msg/sec
- Optimized for multi-tenant environments of 100s of Virtual Machines per server

Enter the World of Scalable Performance
Mellanox HPC – Paving the way to Exascale Computing

**ScalableHPC**

- IMB Barrier - FDR

**Highest Performing Interconnect**

- <0.7usec latency
- 56Gb/s throughput
- Higher scalability
- Maximum Reliability

**Ultimate Scalability with Connect-IB**

- 100Gb/s throughput to network
- Over 130-million messages/second
- Dynamically Connected Transport service for unlimited inter-node scaling

**Accelerating Big Data**

- RDMA Plug-In
- Hadoop Platform
- Socket Lib
- TCP/IP
- IPoIB
- NIC Driver

© 2012 MELLANOX TECHNOLOGIES
Thank You

HPC@mellanox.com

PAVING THE ROAD TO EXASCALE

ADVANCING NETWORK PERFORMANCE, EFFICIENCY, AND SCALABILITY.