MELLANOX EDR UPDATE & GPUDIRECT
Leading Supplier of End-to-End Interconnect Solutions

Enabling the Use of Data

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

- ICs
- Adapter Cards
- Switches/Gateways
- Software and Services
- Metro / WAN
- Cables/Modules

At the Speeds of 10, 25, 40, 50, 56 and 100 Gigabit per Second
Entering the Era of 100Gb/s

**Adapters**
- ConnectX-4
  - 100Gb/s Adapter, 0.7us latency
  - 150 million messages per second
  - (10 / 25 / 40 / 50 / 56 / 100Gb/s)

**Switch**
- SwitchIB
  - 36 EDR (100Gb/s) Ports, <90ns Latency
  - Throughput of 7.2Tb/s

**Switch**
- Spectrum
  - 32 100GbE Ports, 64 25/50GbE Ports
  - (10 / 25 / 40 / 50 / 100GbE)
  - Throughput of 6.4Tb/s

**Interconnect**
- Copper (Passive, Active)
- Optical Cables (VCSEL)
- Silicon Photonics
End-to-End Interconnect Solutions for All Platforms

Highest Performance and Scalability for X86, Power, GPU, ARM and FPGA-based Compute and Storage Platforms

10, 20, 25, 40, 50, 56 and 100Gb/s Speeds

Smart Interconnect to Unleash The Power of All Compute Architectures
Mellanox Centers of Excellence

- Early access to new technologies (EDR, Multi-Host, HPC-X etc.)
- Co-Design effort to optimize and accelerate applications performance and scalability
- Participate in the Mellanox advisory board

Together We Can Develop the Solutions of Tomorrow
Technology Roadmap – One-Generation Lead over the Competition

Mellanox  ➔  20Gbs ➔  40Gbs ➔  56Gbs ➔  100Gbs ➔  200Gbs

Terascale

3rd

TOP500 2003
Virginia Tech (Apple)

Petascale

1st

“Roadrunner”
Mellanox Connected

Exascale

“Summit” System

“Sierra” System

EDR InfiniBand Performance – Commercial Applications

**OptiStruct Performance**
(Engine_Assy.fem)

- Altair

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Performance Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
</tr>
<tr>
<td>24</td>
<td>250</td>
</tr>
</tbody>
</table>

**RADIOSS 13.0 Performance**
(NEON1M11, MPP)

- Altair

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Performance Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
</tr>
<tr>
<td>16</td>
<td>800</td>
</tr>
</tbody>
</table>

**LS-DYNA Performance**
(neon_refined_revised)

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Performance Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
</tr>
<tr>
<td>32</td>
<td>250</td>
</tr>
</tbody>
</table>

© 2015 Mellanox Technologies
- Weather Research and Forecasting Model

- Optimization effort with the HPCAC

- EDR InfiniBand delivers 28% higher performance
  - 32-node cluster
  - Performance advantage increase with system size
### InfiniBand Adapters Performance Comparison

<table>
<thead>
<tr>
<th>Mellanox Adapters</th>
<th>ConnectX-4 EDR 100G</th>
<th>Connect-IB FDR 56G</th>
<th>ConnectX-3 Pro FDR 56G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Port Performance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uni-Directional Throughput</td>
<td>100 Gb/s</td>
<td>54.24 Gb/s</td>
<td>51.1 Gb/s</td>
</tr>
<tr>
<td>Bi-Directional Throughput</td>
<td>195 Gb/s</td>
<td>107.64 Gb/s</td>
<td>98.4 Gb/s</td>
</tr>
<tr>
<td>Latency</td>
<td>0.61 us</td>
<td>0.63 us</td>
<td>0.64 us</td>
</tr>
<tr>
<td>Message Rate</td>
<td>149.5 Million/sec</td>
<td>105 Million/sec</td>
<td>35.9 Million/sec</td>
</tr>
</tbody>
</table>
Mellanox Interconnect Advantages

- Proven, scalable and high performance end-to-end connectivity
- Flexible, support all compute architectures: x86, Power, ARM, GPU, FPGA etc.
- Standards-based (InfiniBand, Ethernet), supported by large eco-system
- Offloading architecture – RDMA, application acceleration engines etc.
- Flexible topologies: Fat Tree, mesh, 3D Torus, Dragonfly+, etc.
- Converged I/O– compute, storage, management on single fabric
- Backward and future compatible
- EDR InfiniBand delivers highest applications performance

Speed-Up Your Present, Protect Your Future
Paving The Road to Exascale Computing Together
Mellanox PeerDirect™ with NVIDIA GPUDirect™ RDMA

Native support for peer-to-peer communications between Mellanox HCA adapters and NVIDIA GPU devices
Industry Adoption of GPUDirect RDMA

- GPUDirect RDMA was released in May 2014 and is available for download from Mellanox
- Adoption and development continues to grow in various areas of technical disciplines
  - Leveraging RDMA and NVIDIA GPUs in today's energy-efficient datacenters
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

- Supports NVIDIA® GPUDirect RDMA with a separate plug-in
- Support for RoCE protocol over Mellanox VPI

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

- Based on Peer-to-Peer capability of PCIe
- Support for any PCIe peer which can provide access to its memory
  - NVIDIA GPU, XEON PHI, AMD, custom FPGA
Evolution of GPUDirect RDMA

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

With GPUDirect Shared Host Memory Pages
- Network and GPU could 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 efficiency between GPUs in remote nodes
- Based on PCIe PeerDirect technology
GPUDirect Sync (GPUDirect 4.0)

- GPUDirect RDMA (3.0) – direct data path between the GPU and Mellanox interconnect
  - Control path still uses the CPU
    - CPU prepares and queues communication tasks on GPU
    - GPU triggers communication on HCA
    - Mellanox HCA directly accesses GPU memory

- GPUDirect Sync (GPUDirect 4.0)
  - Both data path and control path go directly between the GPU and the Mellanox interconnect

Maximum Performance For GPU Clusters

2D stencil benchmark

- 27% faster
- 23% faster

<table>
<thead>
<tr>
<th>Number of nodes/GPUs</th>
<th>Average time per iteration (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>RDMA only 60</td>
</tr>
<tr>
<td>4</td>
<td>RDMA only 60</td>
</tr>
</tbody>
</table>
Note: A requirement on current platforms for GPUDirect RDMA to work properly is that the NVIDIA GPU and the Mellanox InfiniBand Adapter share the same root complex… Only a limitation of current hardware today, not GPUDirect RDMA.
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)
Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency

- Lower is Better

GPU-GPU Internode MPI Bandwidth

- Higher is Better

88% Lower Latency

10X Increase in Throughput

Source: Prof. DK Panda
Remote GPU Access through rCUDA

**GPU servers**

- CUDA Application
- Application
- CUDA Driver + runtime

**GPU as a Service**

- Client Side
  - Application
  - rCUDA library
  - Network Interface
- Server Side
  - rCUDA daemon
  - Network Interface
  - CUDA Driver + runtime

**rCUDA**

- Provides remote access from every node to any GPU in the system

**Throughput (MB/s)**

<table>
<thead>
<tr>
<th>Copy Size (MB)</th>
<th>CUDA</th>
<th>CUDA FDR</th>
<th>CUDA QDR</th>
<th>CUDA GBE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7000</td>
<td>6000</td>
<td>5000</td>
<td>4000</td>
</tr>
<tr>
<td>10</td>
<td>6000</td>
<td>5000</td>
<td>4000</td>
<td>3000</td>
</tr>
<tr>
<td>20</td>
<td>5000</td>
<td>4000</td>
<td>3000</td>
<td>2000</td>
</tr>
<tr>
<td>30</td>
<td>4000</td>
<td>3000</td>
<td>2000</td>
<td>1000</td>
</tr>
<tr>
<td>40</td>
<td>3000</td>
<td>2000</td>
<td>1000</td>
<td>0</td>
</tr>
</tbody>
</table>

**Time for matrix-matrix product (4096x4096)**

- CPU: 2.28 sec
- Remote GPU Over Ethernet: 0.65 sec
- Remote GPU Over InfiniBand: 0.52 sec
- Local GPU: 1.30 sec
- CPU only (MKL): 0.00 sec
Mellanox GPUDirect 기술을 통한 GPU기반 어플리케이션 가속화 솔루션

2X Data Center Performance

5X Higher Throughput
5X Lower Latency

2010
GPUDirect 1.0

2012
GPUDirect 2.0

2014
GPUDirect RDMA

2015
GPUDirect Sync

GPUDirect 1.0

GPUDirect 2.0

GPUDirect RDMA

Mem
CPU
GPUDirect RDMA
Mem
CPU
Mem
Mem
GPU
Mem
GPU
Mem
Mem
RDMA

© 2015 Mellanox Technologies - Mellanox Confidential - 22
THANK YOU

JOIN THE CONVERSATION

#GTC15