Experience with MVAPICH2 on Mellanox Hardware

Gilad Shainer
MUG 2013
Leading Supplier of End-to-End Interconnect Solutions

Comprehensive End-to-End Software Accelerators and Management

- **MPI**
  - Mellanox Messaging Acceleration (MXM)
- **SHMEM**
  - Fabric Collectives Acceleration (FCA)
- **PGAS**
  - Unified Fabric Management (UFM)

Management

- **Storage and Data**
  - **VSA**
    - Storage Accelerator (iSCSI)
  - **UDA**
    - Unstructured Data Accelerator

Comprehensive End-to-End InfiniBand and Ethernet Solutions Portfolio

- **ICs**
- **Adapter Cards**
- **Switches/Gateways**
- **Long-Haul Systems**
- **Cables/Module**

© 2013 Mellanox Technologies

- Mellanox Confidential -
Mellanox
Mellanox Interconnect Development Timeline

- QDR InfiniBand End-to-End
- GPUDirect Technology Released
- FDR InfiniBand End-to-End
- MPI/SHMEM Collectives Offloads (FCA), Scalable HPC (MXM), Open SHMEM, PGAS/UPC
- Connect-IB - 100Gb/s HCA Dynamically Connected Transport

2008

World’s First Petaflop Systems

2009

CORE-Direct Technology

2010

InfiniBand – Ethernet Bridging

2011

Long-Haul Solutions

2012

GPUDirect RDMA

Technology and Solutions Leadership, EDR InfiniBand Expected in 2014-2015
Mellanox InfiniBand Paves the Road to Exascale
Connect-IB

Architectural Foundation for Exascale Computing
World’s first 100Gb/s interconnect adapter
- PCIe 3.0 x16, dual FDR 56Gb/s InfiniBand ports to provide >100Gb/s

Highest InfiniBand message rate: 137 million messages per second
- 4X higher than other InfiniBand solutions

<0.7 micro-second application latency

Supports GPUDirect RDMA for direct GPU-to-GPU communication

New Innovative Transport – Dynamically Connected Transport Service

Supports Scalable HPC with MPI, SHMEM and PGAS/UPC offloads

Enter the World of Boundless Performance
GPUDirect RDMA

**Receive**

- System Memory
- GPU
  - GPU Memory
- CPU
  - Chip set
  - InfiniBand

**Transmit**

- System Memory
- GPU
  - GPU Memory
- CPU
  - Chip set
  - InfiniBand

**GPUDirect 1.0**

- CPU
  - InfiniBand
- GPU
  - Chip set
  - InfiniBand

**GPUDirect RDMA**

- CPU
  - InfiniBand
- GPU
  - Chip set
  - InfiniBand

© 2013 Mellanox Technologies

- Mellanox Confidential -
Preliminary Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency

Small Message Latency

- MVAPICH2-1.9
- MVAPICH2-1.9-GDR

Lower is Better

69% Lower Latency

Source: Prof. DK Panda

GPU-GPU Internode MPI Bandwidth

Small Message Bandwidth

- MVAPICH2-1.9
- MVAPICH2-1.9-GDR

Higher is Better

3X Increase in Throughput
Execution Time of HSG (Heisenberg Spin Glass) Application with 2 GPU Nodes

Source: Prof. DK Panda
Some Performance Data
Applications Examples

Quantum ESPRESSO Benchmark
(AUSURF112)

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Open MPI</th>
<th>MVAPICH2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>25</td>
</tr>
</tbody>
</table>

Iterations / hour
MVAPICH2 1.9 Performance
(osu_latency)
MVAPICH2 1.9 Performance
(osu_bw)

Bandwidth (MB/s)

Connect-IB Single Port
Connect-IB Dual Port
ISC’13 – Student Cluster Challenge Teams
ISC’13 Student Cluster Challenge
Submission for Participation is Open
Thank You