Advancing Applications Performance With InfiniBand

Pak Lui, Application Performance Manager

September 12, 2013
Mellanox Overview

- **Leading provider of high-throughput, low-latency server and storage interconnect**
  - FDR 56Gb/s InfiniBand and 10/40/56GbE
  - Reduces application wait-time for data
  - Dramatically increases ROI on data center infrastructure

- **Company headquarters:**
  - Yokneam, Israel; Sunnyvale, California
  - ~1,200 employees* worldwide

- **Solid financial position**
  - Record revenue in FY12; $500.8M, up 93% year-over-year
  - Q2’13 revenue of $98.2M
  - Q3’13 guidance ~$104M to $109M
  - Cash + investments @ 6/30/13 = $411.3M

* As of June 2013
Providing End-to-End Interconnect Solutions

Comprehensive End-to-End Software Accelerators and Management

- **MPI**
- **SHMEM**
- **PGAS**

MXM
Mellanox Messaging Acceleration

FCA
Fabric Collectives Acceleration

**UFM**
Unified Fabric Management

Management

Storage and Data

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

Comprehensive End-to-End InfiniBand and Ethernet Solutions Portfolio

<table>
<thead>
<tr>
<th>ICs</th>
<th>Adapter Cards</th>
<th>Switches/Gateways</th>
<th>Long-Haul Systems</th>
<th>Cables/Modules</th>
</tr>
</thead>
</table>

© 2013 Mellanox Technologies
Virtual Protocol Interconnect (VPI) Technology

**ConnectX®-3** VPI Adapter

- redhat
- Novell
- Microsoft
- VMware
- OpenStack

**SwitchX®-2** VPI Switch

- Unified Fabric Manager
- Switch OS Layer

**Applications**
- Networking
- Storage
- Clustering
- Management

**Acceleration Engines**
- Ethernet: 10/40/56 Gb/s
- InfiniBand: 10/20/40/56 Gb/s

**From data center to campus and metro connectivity**

- 64 ports 10GbE
- 36 ports 40/56GbE
- 48 10GbE + 12 40/56GbE
- 36 ports IB up to 56Gb/s
- 8 VPI subnets

**Applications**
- Networking
- Storage
- Clustering
- Management

**Features**
- 3.0
- Adapter Card
- Mezzanine Card

**Networking**
- From data center to campus and metro connectivity
MetroDX™ and MetroX™

- MetroX™ and MetroDX™ extends InfiniBand and Ethernet RDMA reach
- Fastest interconnect over 40Gb/s InfiniBand or Ethernet links
- Supporting multiple distances
- Simple management to control distant sites
- Low-cost, low-power, long-haul solution

40Gb/s over Campus and Metro
Data Center Expansion Example – Disaster Recovery
Key Elements in a Data Center Interconnect

- **Servers**
- **Application Icons**
- **Storage**
- **Switch and IC**
- **Cables, Silicon Photonics, Parallel Optical Modules**

**Adapter and IC**

© 2013 Mellanox Technologies
Recent Acquisitions of Kotura and IPtronics Enable Mellanox to Deliver Complete High-Speed Optical Interconnect Solutions for 100Gb/s and Beyond
Mellanox InfiniBand Paves the Road to Exascale
Meteo France

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Centre National de Calcul Météopole, Toulouse</td>
<td>522 TFlops peak performance 56 racks built DLC 1008 nodes Fat Tree InfiniBand FDR Lustre 2 Po, 69 GB/s Disks storage 209 TB</td>
<td>2,85 PFlops peak performance 56+45 racks built DLC 1800 nodes Fat Tree InfiniBand FDR Lustre 3,57 Po, 188 GB/s Disks storage 400 TB</td>
<td>Operational Espace Clément Ader Montaudran</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2013 Mellanox Technologies
20K InfiniBand nodes
Mellanox end-to-end FDR and QDR InfiniBand
Supports variety of scientific and engineering projects
- Coupled atmosphere-ocean models
- Future space vehicle design
- Large-scale dark matter halos and galaxy evolution

Asian Monsoon Water Cycle

High-Resolution Climate Simulations
NCAR (National Center for Atmospheric Research)

- “Yellowstone” system
- 72,288 processor cores, 4,518 nodes
- Mellanox end-to-end FDR InfiniBand, CLOS (full fat tree) network, single plane
Applications Performance (Courtesy of the HPC Advisory Council)

COSMO RAPS 5.1 Performance (COSMO_EU)

ICON RAPS 2.0 Performance (test_hat_jww, Ivy Bridge)
Applications Performance (Courtesy of the HPC Advisory Council)

WRF Benchmark (conus12km)

NEMO Benchmark

© 2013 Mellanox Technologies
Dominant in Enterprise Back-End Storage Interconnects

SMB Direct

Microsoft

NetApp

xyratex

EMC²

DataDirect Networks

IBM

Oracle
Leading Interconnect, Leading Performance

Bandwidth

- 10Gb/s (2001)
- 40Gb/s (2008)
- 56Gb/s (2010)
- 100Gb/s (2013)
- 200Gb/s (2016)

Latency

- 5 usec (2001)
- 2.5 usec (2004)
- 1.3 usec (2008)
- 0.7 usec (2010)
- 0.6 usec (2013)
- <0.5 usec (2016)

Same Software Interface

- 2001 to 2017
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

Unmatchable Storage Performance
- 8,000,000 IOPs (1QP), 18,500,000 IOPs (32 QPs)

New Innovative Transport – Dynamically Connected Transport Service

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

Enter the World of Boundless Performance
Connect-IB Memory Scalability

Host Memory Consumption (MB)

InfiniHost, RC 2002
InfiniHost-III, SRQ 2005
ConnectX, XRC 2008
Connect-IB, DCT 2012

- 8 nodes
- 2K nodes
- 10K nodes
- 100K nodes
Dynamically Connected Transport Advantages

Dynamically Connected Transport Advantages
(MPI Latency Reduction %)

Message Size (Byte)

- DCT Advantage for IMB Allreduce Latency
- DCT Advantage for IMB Alltoall Latency
FDR InfiniBand Delivers Highest Application Performance

**OpenFOAM Performance**
(Lid-driven Cavity)

- **QDR InfiniBand**
- **FDR InfiniBand**

**Message Rate**

- **QDR InfiniBand**
- **FDR InfiniBand**

**NAMD Benchmark**
(Platform MPI, ApoA1)

- **10GbE**
- **40GbE**
- **InfiniBand FDR**

**RADIOSS Benchmark**
(NEON1M11, MPI)

- **1GbE**
- **10GbE**
- **40GbE**
- **FDR InfiniBand**
Scalable Communication

MXM, FCA
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
- CORE-Direct Hardware Offload

**InfiniBand Verbs API**
MXM v2.0 - Highlights

- Transport library integrated with OpenMPI, OpenSHMEM, BUPC, Mvapich2
  - More solutions will be added in the future
  - Utilizing Mellanox offload engines

- Supported APIs (both sync/async): AM, p2p, atomics, synchronization

- Supported transports: RC, UD, DC, RoCE, SHMEM

- Supported built-in mechanisms: tag matching, progress thread, memory registration cache, fast path send for small messages, zero copy, flow control

- Supported data transfer protocols: Eager Send/Recv, Eager RDMA, Rendezvous
Mellanox FCA Collective Scalability

**Barrier Collective**

- Horizontal axis: Processes (PPN=8)
- Vertical axis: Latency (us)
- Graphs show latency comparison between Without FCA and With FCA.

**Reduce Collective**

- Horizontal axis: Processes (PPN=8)
- Vertical axis: Latency (us)
- Graphs show latency comparison between Without FCA and With FCA.

**8-Byte Broadcast**

- Horizontal axis: Processes (PPN=8)
- Vertical axis: Bandwidth (KB/processes)
- Graphs show bandwidth comparison between Without FCA and With FCA.
FDR InfiniBand Delivers Highest Application Performance

ICON 2.0 Performance
(test_hat_jww)

Performance (Jobs/Hour)

Number of Nodes

<table>
<thead>
<tr>
<th>Nodes</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open MPI (Untuned)</td>
<td>50</td>
<td>75</td>
<td>100</td>
<td>125</td>
<td>150</td>
</tr>
<tr>
<td>Open MPI (Tuned by FCA)</td>
<td>100</td>
<td>125</td>
<td>150</td>
<td>175</td>
<td>200</td>
</tr>
</tbody>
</table>

CESM Performance
(B1850CN, ATM)

model-years/wall-day

Nodes

<table>
<thead>
<tr>
<th>Nodes</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>No FCA</td>
<td>5</td>
<td>7.5</td>
<td>12</td>
</tr>
<tr>
<td>FCA</td>
<td>10</td>
<td>15</td>
<td>20</td>
</tr>
</tbody>
</table>
GPU Direct
GPUDirect RDMA

Receive

System Memory → CPU ← Chip set → GPU → CPU

GPUDirect 1.0

InfiniBand → InfiniBand

Transmit

System Memory → CPU ← Chip set → GPU → CPU

GPUDirect RDMA

InfiniBand → InfiniBand

© 2013 Mellanox Technologies
Preliminary Performance of MVAPICH2 with GPU Direct RDMA

GPU-GPU Internode MPI Latency

**Small Message Latency**

- **MVAPICH2-1.9**
- **MVAPICH2-1.9-GDR**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.78</td>
</tr>
<tr>
<td>4</td>
<td>6.12</td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td></td>
</tr>
<tr>
<td>1K</td>
<td></td>
</tr>
<tr>
<td>4K</td>
<td></td>
</tr>
</tbody>
</table>

**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**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Bandwidth (MB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td></td>
</tr>
<tr>
<td>1K</td>
<td></td>
</tr>
<tr>
<td>4K</td>
<td></td>
</tr>
</tbody>
</table>

**Higher is Better**

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

Source: Prof. DK Panda
Thank You