GPU-InfiniBand Accelerations for Hybrid Compute Systems

Pak Lui
March 2013, GTC
Leading Supplier of End-to-End Interconnect Solutions

Server / Compute

Virtual Protocol Interconnect
56G IB & FCoIB

10/40/56GbE & FCoE

Switch / Gateway

Virtual Protocol Interconnect
56G InfiniBand

10/40/56GbE

Fibre Channel

Storage Front / Back-End

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs
Adapter Cards
Switches/Gateways
Host/Fabric Software
Cables

© 2013 Mellanox Technologies
- Mellanox Confidential -
Virtual Protocol Interconnect (VPI) Technology

**ConnectX-3** VPI Adapter

- Applications:
  - Networking
  - Storage
  - Clustering
  - Management

- Acceleration Engines:
  - Ethernet: 10/40/56 Gb/s
  - InfiniBand: 10/20/40/56 Gb/s
  - PCI EXPRESS 3.0

**SwitchX-2** VPI Switch

- Unified Fabric Manager
- Switch OS Layer
- 64 ports 10GbE
- 36 ports 40/56GbE
- 48 10GbE + 12 40/56GbE
- 36 ports IB up to 56Gb/s
- 8 VPI subnets

From data center to campus and metro connectivity
A new interconnect architecture for compute intensive applications
World’s fastest server and storage interconnect solution
Enables unlimited clustering (compute and storage) scalability
Accelerates compute-intensive and parallel-intensive applications
Optimized for multi-tenant environments of 100s of Virtual Machines per server
Connect-IB Performance Highlights

- World’s first 100Gb/s InfiniBand 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

Unparalleled Throughput and Message Injection Rates
GPUDirect History

- The GPUDirect project was announced Nov 2009
  - “NVIDIA Tesla GPUs To Communicate Faster Over Mellanox InfiniBand Networks”, http://www.nvidia.com/object/io_1258539409179.html

- GPUDirect was developed together by Mellanox and NVIDIA
  - New interface (API) within the Tesla GPU driver
  - New interface within the Mellanox InfiniBand drivers
  - Linux kernel modification to allow direct communication between drivers

- GPUDirect 1.0 was announced Q2'10
  - “Mellanox Scalable HPC Solutions with NVIDIA GPUDirect Technology Enhance GPU-Based HPC Performance and Efficiency”
  - “Mellanox was the lead partner in the development of NVIDIA GPUDirect”

- GPUDirect RDMA will be released Q2'13
GPU-InfiniBand Bottleneck (pre-GPUDirect)

- GPU communications uses “pinned” buffers for data movement
  - A section in the host memory that is dedicated for the GPU
  - Allows optimizations such as write-combining and overlapping GPU computation and data transfer for best performance

- InfiniBand uses “pinned” buffers for efficient RDMA transactions
  - Zero-copy data transfers, Kernel bypass
  - Reduces CPU overhead
GPUDirect 1.0

Receive

System Memory

CPU

Chip set

GPU

GPU Memory

Non GPUDirect

InfiniBand

InfiniBand

Transmit

System Memory

CPU

Chip set

GPU

GPU Memory

GPUDirect 1.0

InfiniBand

InfiniBand

Non GPUDirect

InfiniBand

InfiniBand

GPUDirect 1.0

System Memory

CPU

Chip set

GPU

GPU Memory

Non GPUDirect

InfiniBand

InfiniBand

Transmit

System Memory

CPU

Chip set

GPU

GPU Memory

GPUDirect 1.0

InfiniBand

InfiniBand

Non GPUDirect

InfiniBand

InfiniBand
GPUDirect 1.0 – Application Performance

- **LAMMPS**
  - 3 nodes, 10% gain

- **Amber – Cellulose**
  - 8 nodes, 32% gain

- **Amber – FactorIX**
  - 8 nodes, 27% gain
GPUDirect RDMA

**Receive**

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

---

**Transmit**

- CPU
- Chip set
- GPU

---

**InfiniBand**

---

**GPUDirect 1.0**

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

---

**GPUDirect RDMA**

---

© 2013 Mellanox Technologies

- Mellanox Confidential -
Initial Design of OSU-MVAPICH2 with GPU-Direct-RDMA

- Preliminary driver for GPU-Direct is under work by NVIDIA and Mellanox
- OSU has done an initial design of MVAPICH2 with the latest GPU-Direct-RDMA Driver
  - Hybrid design
  - Takes advantage of GPU-Direct-RDMA for short messages
  - Uses host-based buffered design in current MVAPICH2 for large messages
  - Alleviates Sandybridge chipset bottleneck
Preliminary Performance of MVAPICH2 with GPU-Direct-RDMA

Based on MVAPICH2-1.9b
Intel Sandy Bridge (E5-2670) node with 16 cores
NVIDIA Telsa K20c GPU, Mellanox ConnectX-3 FDR HCA
CUDA 5.0, OFED 1.5.4.1 with GPU-Direct-RDMA Patch
Preliminary Performance of MVAPICH2 with GPU-Direct-RDMA

GPU-GPU Internode MPI Uni-directional Bandwidth

Based on MVAPICH2-1.9b
Intel Sandy Bridge (E5-2670) node with 16 cores
NVIDIA Telsa K20c GPU, Mellanox ConnectX-3 FDR HCA
CUDA 5.0, OFED 1.5.4.1 with GPU-Direct-RDMA Patch
Preliminary Performance of MVAPICH2 with GPU-Direct-RDMA

Based on MVAPICH2-1.9b
Intel Sandy Bridge (E5-2670) node with 16 cores
NVIDIA Telsa K20c GPU, Mellanox ConnectX-3 FDR HCA
CUDA 5.0, OFED 1.5.4.1 with GPU-Direct-RDMA Patch
Remote GPU Access through rCUDA

**GPU servers**

- CUDA Application
  - Application
  - CUDA Driver + runtime
  - CUDA device

**GPU as a Service**

- Client Side
  - Application
  - rCUDA library
  - Network Interface

- Server Side
  - rCUDA daemon
  - Network Interface
  - CUDA Driver + runtime
  - CUDA device

rCUDA provides remote access from every node to any GPU in the system
GPU as a Service

- GPUs as a network-resident service
  - Little to no overhead when using FDR InfiniBand

- Virtualize and decouple GPU services from CPU services
  - A new paradigm in cluster flexibility
  - Lower cost, lower power and ease of use with shared GPU resources
  - Remove difficult physical requirements of the GPU for standard compute servers
Thank You