InfiniBand In-Network Computing Technology and Roadmap

Gilad Shainer, MUG, August 2019
The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale

CPU-Centric (Onload) vs. Data-Centric (Offload)

- **CPU-Centric (Onload)**: Must Wait for the Data, Creates Performance Bottlenecks
- **Data-Centric (Offload)**: Analyze Data as it Moves! Higher Performance and Scale

In-Network Computing

© 2019 Mellanox Technologies
Accelerating All Levels of HPC / AI Frameworks

Application

- Data Analysis
- Real Time
- Deep Learning

Communication

- Mellanox SHARP In-Network Computing
- MPI Tag Matching
- MPI Rendezvous

Network

- Network Transport Offload
- RDMA and GPU-Direct RDMA
- SHIELD (Self-Healing Network)
- Enhanced Adaptive Routing and Congestion Control

Connectivity

- Multi-Host Technology
- Socket-Direct Technology
- Enhanced Topologies
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations

- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications

- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16/32/64 bits
SHARP AllReduce Performance Advantages (128 Nodes)

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency
SHARP AllReduce Performance Advantages
1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology

SHARP Enables Highest Performance

MPI AllReduce Latency
1500 Nodes, 1PPN

MPI AllReduce Latency
1500 Nodes, 40PPN, 60K MPI Ranks
NCCL-SHARP Delivers Highest Performance

Mellanox SHARP Plug-in for NCCL 2.4 (Bandwidth)

- NCCL-TREE
- NCCL-RING
- NCCL-SHARP

4 system nodes - (32) NVIDIA V100 16GB SXM2 with NVLINK

© 2019 Mellanox Technologies
SHARP Performance Advantage for AI

- SHARP provides 16% Performance Increase for deep learning, initial results
- TensorFlow with Horovod running ResNet50 benchmark, HDR InfiniBand (ConnectX-6, Quantum)

P100 NVIDIA GPUs, RH 7.5, Mellanox OFED 4.4, HPC-X v2.3, TensorFlow v1.11, Horovod 0.15.0
MPI Tag Matching
Hardware Engine
Tag Matching Hardware Engine Performance Advantage

MPI Latency (Eager)

- MVAPICH2
- MVAPICH2+HW-TM

35% performance advantage

MPI iscatterv (1,280 Processes)

- MVAPICH2
- MVAPICH2+HW-TM

1.8X performance advantage

Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University
GPUDirect
Mellanox PeerDirect™ Technology

- Purpose-built for acceleration of Deep Learning
- Provides significant decrease in communication latency for acceleration devices
- Peer-to-peer communications between Mellanox adapters and third-party devices
- Enables GPUDirect™ RDMA, GPUDirect™ ASYNC, ROCm and others

Designed for Deep Learning Acceleration
10X Higher Performance with GPUDirect™ RDMA

- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs

![Graph showing performance improvements](image)

GPU-InfiniBand-GPU Latency

- **10X** improvement in latency
- **1.88 usec** latency

GPU-InfiniBand-GPU Throughput

- **9X** improvement in throughput
- **11X** improvement in throughput (Bi-Dir)

Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

© 2019 Mellanox Technologies
Quality of Service
InfiniBand Quality of Service

<table>
<thead>
<tr>
<th>User / Workload</th>
<th>Category</th>
<th>Service Level</th>
<th>Virtual Lanes over Physical Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>Other</td>
<td>SL 0-3</td>
<td>VL-0 W 32</td>
</tr>
<tr>
<td>User 1</td>
<td>Backup</td>
<td>SL 4</td>
<td>VL-1 W 32</td>
</tr>
<tr>
<td>User 1</td>
<td>Storage</td>
<td>SL 6</td>
<td>VL-2 W 64</td>
</tr>
<tr>
<td>User 2</td>
<td>MPI</td>
<td>SL 8</td>
<td>VL-4 W 64</td>
</tr>
<tr>
<td>User 2</td>
<td>MPI</td>
<td>SL 10</td>
<td>VL-5 W 64</td>
</tr>
<tr>
<td>User 3</td>
<td>MPI</td>
<td>SL 12</td>
<td>VL-6 W 32</td>
</tr>
<tr>
<td>User 4</td>
<td>Clock Sync</td>
<td>SL 12</td>
<td></td>
</tr>
</tbody>
</table>

Low Priority VL Arbitrary

High Priority VL Arbitrary

Network
SHIELD
Self Healing Technology
SHIELD - Self Healing Technology

Enables Unbreakable Data Centers

- The ability to overcome network failures, locally, by the switches
- Software-based solutions suffer from long delays detecting network failures
  - 5-30 seconds for 1K to 10K nodes clusters
  - Accelerates network recovery time by 5000X
  - The higher the speed or scale the greater the recovery value
- Available with EDR and HDR switches and beyond
Adaptive Routing
InfiniBand Proven Adaptive Routing Performance

- Oak Ridge National Laboratory – Coral Summit supercomputer
- Bisection bandwidth benchmark, based on mpiGraph
  - Explores the bandwidth between possible MPI process pairs
- AR results demonstrate an average performance of 96% of the maximum bandwidth measured

mpiGraph explores the bandwidth between possible MPI process pairs. In the histograms, the single cluster with AR indicates that all pairs achieve nearly maximum bandwidth while single-path static routing has nine clusters as congestion limits bandwidth, negatively impacting overall application performance.

HDR InfiniBand
Highest-Performance 200Gb/s InfiniBand Solutions

**Adapters**
- ConnectX-6
- 200Gb/s Adapter, 0.6us latency
- 215 million messages per second
- (10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

**Switch**
- Mellanox Quantum
- 40 HDR (200Gb/s) InfiniBand Ports
- 80 HDR100 InfiniBand Ports
- Throughput of 16Tb/s, <90ns Latency

**SoC**
- BlueField
- System on Chip and SmartNIC
- Programmable adapter
- Smart Offloads

**Interconnect**
- LinkX
- Transceivers
- Active Optical and Copper Cables
- (10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

**Software**
- HPC-X
- MPI, SHMEM/PGAS, UPC
- For Commercial and Open Source Applications
- Leverages Hardware Accelerations
ConnectX-6 HDR InfiniBand Adapter

Leading Connectivity

- 200Gb/s InfiniBand and Ethernet
- HDR, HDR100, EDR (100Gb/s) and lower speeds
- 200GbE, 100GbE and lower speeds
- Single and dual ports

Leading Performance

- 200Gb/s throughput, 0.6usec latency, 215 million message per second
- PCIe Gen3 / Gen4, 32 lanes
- Integrated PCIe switch
- Multi-Host - up to 8 hosts, supporting 4 dual-socket servers

Leading Features

- In-network computing and memory for HPC collective offloads
- Security – Block-level encryption to storage, key management, FIPS
- Storage – NVMe Emulation, NVMe-oF target, Erasure coding, T10/DIF
HDR InfiniBand Switches

40 QSFP56 ports
- 40 ports of HDR, 200G
- 80 ports of HDR100, 100G

800 QSFP56 ports
- 800 ports of HDR, 200G
- 1600 ports of HDR100, 100G
Real Time Network Visibility

Built-in Hardware Sensors for Rich Traffic Telemetry and Data Collection

Advanced monitoring for troubleshooting

- 8 mirror agents triggered by congestion, buffer usage and latency
- Measure queue depth using histograms (64ns granularity)

Network status/health in real time

- Buffer snapshots
- Congestion notifications and buffers status
BlueField SoC
Advantages and Platforms
BlueField Block Diagram

- **Tile Architecture - 16 ARM® A72 CPUs subsystem**
  - SkyMesh™ fully coherent low-latency interconnect
  - 8MB L2 Cache, 8 Tiles

- **Dual Port 100g IO Controller, based on ConnectX-5**
  - Dual 100Gb/s Ethernet/InfiniBand, compatible with ConnectX-5
  - NVMe-oF hardware accelerator
  - High-end Networking Offloads: RDMA, Erasure Coding, T10-DIF

- **Fully Integrated PCIe switch**
  - 32 Bifurcated PCI Gen3/4 lanes (up to 200Gb/s)
  - Root Complex or Endpoint modes
  - 2x16, 4x8, 8x4 or 16x2 configurations

- **Memory Controllers**
  - 2x Channels DDR4 Memory Controllers w/ ECC
  - NVDIMM-N Support
BlueField for Smart Solutions

BlueField SoC (System on Chip)
- SoC: Compute, networking and PCIe connectivity
  - Dual port VPI EDR/100GbE
  - 16 Arm cores
  - 32 lanes of PCIe switch gen3/4

Storage Solutions
- NVMe-based storage platforms
  - RDMA, NVMe over Fabrics, RAID, Signature offload
  - Partner’s solutions based on BlueField storage controller

Smart Adapters
- In-network computing and collective offloads
- Co-processor running proprietary smart algorithms
- Security and privacy algorithms
BlueField Smart Adapter is a Computer

- CPU
- L2/3 Cache
- Hardware-based accelerators
- A fully functioning Operating System

Network Adapters

Memory
Highest Performance and Scalability for Exascale Platforms

96% Network Utilization

7X Higher Performance

Flat Latency

2X Higher Performance

5000X Higher Resiliency

HDR 200G

NDR 400G

XDR 1000G
Thank You