InfiniBand In-Network Computing Technology

Paving the Road to Exascale

September 2019
HDR 200G InfiniBand Wins Next Generation Supercomputers

23.5 Petaflops
8K HDR InfiniBand Nodes
Fat-Tree Topology

KMA
Korea Meteorological Administration

50 Petaflops
7.2K HDR InfiniBand Nodes
Dragonfly+ Topology

Australian National University

3K HDR InfiniBand Nodes
Dragonfly+ Topology

3.1 Petaflops
1.8K HDR InfiniBand Nodes
Fat-Tree Topology

CSC
Finnish Meteorological Institute

1.7 Petaflops
2K HDR InfiniBand Nodes
Dragonfly+ Topology

Microsoft Azure

Highest Performance Cloud

SDSC
San Diego Supercomputer Center

1.6 Petaflops
Hybrid CPU-GPU-FPGA
Fat-Tree Topology

© 2019 Mellanox Technologies
ICON (ICOsaheedral Non-hydrostatic Model) Application

- New generation unified weather forecasting and climate model
- Developed by Max Planck Institute for Meteorology (MPI-M) and the German Meteorological Service (DWD)
- New data exchange module YAXT developed to replace traditional halo exchange mechanism
  - Main challenge lies on efficient handling of sparse data at scale
  - Improvement jointly developed by DKRZ, UTK, and Mellanox

![ICON Performance Chart](chart.png)

© 2019 Mellanox Technologies
The Need for Intelligent and Faster Interconnect

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

- CPU-Centric (Onload)
  - Must Wait for the Data
  - Creates Performance Bottlenecks

- Data-Centric (Offload)
  - Analyze Data as it Moves!
  - Higher Performance and Scale
Data Centric Architecture to Overcome Latency Bottlenecks

Intelligent Interconnect Paves the Road to Exascale Performance

CPU-Centric (Onload)

Communications Latencies of 30-40us

Data-Centric (Offload)

Communications Latencies of 3-4us
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

Scalable Hierarchical Aggregation and Reduction Protocol
SHARP Accelerates AI Performance

The CPU in a parameter server becomes the bottleneck

Performs the Gradient Averaging
Replaces all physical parameter servers
Accelerate AI Performance
NCCL-SHARP Delivers Highest Performance

Mellanox SHARP Plug-in for NCCL 2.4
(Bandwidth)

4 system nodes - (32) NVIDIA V100 16GB SXM2 with NVLINK
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
Quality of Service
InfiniBand Quality of Service

User / Workload | Category | Service Level | Virtual Lanes over Physical Link
--- | --- | --- | ---
User 1 | Other | SL 0-3 | VL-0 W 32
User 1 | Backup | SL 4 | VL-1 W 32
User 1 | Storage | SL 6 | VL-2 W 64
User 2 | MPI | SL 8 | VL-4 W 64
User 2 | MPI | SL 10 | VL-5 W 64
User 2 | Clock Sync | SL 12 | VL-6 W 32
User 3 | Other | SL 0-3 | VL-0 W 32
User 3 | Backup | SL 4 | VL-1 W 32
User 3 | Storage | SL 6 | VL-2 W 64
User 4 | MPI | SL 8 | VL-4 W 64
User 4 | MPI | SL 10 | VL-5 W 64
User 4 | Clock Sync | SL 12 | VL-6 W 32

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
SHIELD: Consider a Flow From A to B
SHIELD: The Simple Case: Local Fix
SHIELD: The Remote Case - Using Fault Recovery Notifications
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, 215 million message per second
- PCIe Gen3 / Gen4, 32 lanes
- Integrated PCIe switch
- Multi-Host

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

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)
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
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