InfiniBand In-Network Computing

Paving the Road to Exascale

June 2019
SUPERCONNECTING
the #1 Supercomputers

InfiniBand Accelerates 6 of Top 10 Supercomputers
SUPERCONNECTING the #1 Supercomputers

HDR 200G InfiniBand Accelerated Supercomputers
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
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**
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 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
GPUDirect
10X Higher Performance with GPUDirect™ RDMA

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

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

Network Topologies
Supporting Variety of Topologies

- **Fat Tree**
- **Torus**
- **Dragonfly**
- **Hypercube**
- **HyperX**
Dragonfly+ vs Traditional Dragonfly
Dragonfly+ Topology

- Several “groups”, connected using all to all links
- The topology inside each group can be any topology
- Reduce total cost of network (fewer long cables)
- Utilizes Adaptive Routing for efficient operations
- Simplifies future system expansion

1200-Nodes Dragonfly+ Systems Example

Full-Graph connecting every group to all other groups
BlueField SoC
Programmable Network
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
Thank You