InfiniBand In-Network Computing for HPC and AI

Unleash The Power Of Data

Oct 2019
Data Center Is Moving To Data Centric

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

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

Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale
Data Growth in HPC and AI Computing

- The large amount of image data being generated every day from state-of-the-art direct electron detectors
- Intensive computation needed for 3D reconstruction software, such as the RELION package

**Massive Datasets**

Huge volumes of data generated

**Compute Intensive**

Scaling from Workstation to Cluster

**High Speed Interconnect**

High speed and low latency for linear scaling efficiency
Leading Supplier of InfiniBand and Ethernet End-to-End Interconnect Solutions

The Smart Choice for Intelligent Compute and Storage Platforms
Highest-Performance 200Gb/s Interconnect Solutions

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

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

**Switch**
- **Spectrum-2**
  - 16 400GbE, 32 200GbE, 128 25/50GbE Ports
    (10 / 25 / 40 / 50 / 100 / 200 GbE)
  - Throughput of 6.4Tb/s

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

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

**Software**
- **HPC-X**
  - MPI, SHMEM/PGAS, UPC
  - For Commercial and Open Source Applications
  - Leverages Hardware Accelerations
Building the HPC/AI Cluster for Enterprise

Mellanox InfiniBand
HDR/EDR 100/200Gb/s

HPE Apollo 6500

HPE StorageFabric M-series
Ethernet 1/10/25/40/50/100GbE

Parallel File System

Utility Nodes
- Login
- Mgmt
- Monitor
- ...

Internet
Deep Learning Demands Highest Performance

**TRAINING**
- 100/200/400G
- RDMA
- SHARP
- GPUDirect
- NVMe over Fabrics

**INFERENCING**
- 100G
- RDMA
- GPUDirect

Images
- Video
- Text
- Speech
- Tabular
- Time Series

**TRAINING DATASET**

**NEW DATA**

Billions of TFLOPS

Billions of FLOPS
Mellanox RDMA Technology

RDMA over InfiniBand or Ethernet

TCP/IP

RACK 1

RACK 2

Buffer 1

NIC

OS

Application

Buffer 1

Buffer 1

Buffer 1

Kernel

HARDWARE

USER
GPUDirect™

- GPUDirect
  - Accelerates communication between servers
  - Allows peer to peer transfers between GPUs
  - Requires RoCE or InfiniBand

ML Training Scenario
- Billions of TFLOPS per training run
- Years of compute-days on Xeon
- GPUDirect turns years to days

Without GPU Direct
- Same Data Copied 3x

With GPUDirect
(Requires RoCE or InfiniBand)
10X Higher Performance with GPUDirect™ RDMA

- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs
**Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)**

**In Network Data Reduction**

- **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
Scalability – too long on the Interconnect
Scalability – Simplifying one to many
Scalability – Compute in the interconnect

- Collective tree created to nodes
- Nodes send data up the tree to leaf switches, where collectives operation is run
- Leaf switches send data up the tree to spine, where collectives operation is run
- Result is sent to egress leaf
- Result arrives at Requester node where no operation is needed
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
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)
Adaptive Routing (AR) Performance – ORNL Summit

- 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.
Mellanox Accelerates TensorFlow with RDMA

Unmatched Linear Scalability at No Additional Cost

50% Better Performance

Higher is better
RDG: RDMA accelerated Horovod framework in Docker on HPE Apollo 6500 servers over InfiniBand fabric.

https://docs.mellanox.com/pages/releaseview.action?pageId=15049840
### Accelerating All Levels of HPC/AI Frameworks

| Application Framework | ▪ Data Analysis  
|▪ Configurable Logic |
|-----------------------|------------------|
| Communication Framework | ▪ SHARP – Data Aggregation  
|▪ MPI Tag Matching  
|▪ MPI Rendezvous  
|▪ SNAP - Software Defined Virtual Devices |
| Network Framework | ▪ Network Transport Offload  
|▪ RDMA and GPU-Direct  
|▪ SHIELD (Self-Healing Network)  
|▪ Adaptive Routing and Congestion Control |
| Connectivity Framework | ▪ Multi-Host  
|▪ Enhanced Topologies  
|▪ Dragonfly+ |
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