ACCELERATING SCIENTIFIC COMPUTING

NVIDIA InfiniBand Networking, MUG August 2021
EXPANDING UNIVERSE OF SCIENTIFIC COMPUTING

REAL-TIME PROCESSING

SIMULATIONS

TRAINING AI

DIGITAL TWIN

VISUALIZATION

SUPERCOMPUTING

EDGE

NETWORK

EXTREME IO

CLOUD
NVIDIA 200G HDR INFINIBAND INFRASTRUCTURE

In-Network Computing Accelerated Network for Supercomputing

- Metrox-2 Long-haul
- Skyway Gateway
- UFM Cyber-AI
- ConnectX Adapter
- BlueField DPU
- Quantum Switch
- Linkx
IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING
Software-Defined, Hardware-Accelerated, InfiniBand Network

Most Advanced Networking

<table>
<thead>
<tr>
<th>End-to-End</th>
<th>High Throughput</th>
<th>Extremely Low Latency</th>
<th>High Message Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDMA</td>
<td>GPUDirect RDMA</td>
<td>GPUDirect Storage</td>
<td></td>
</tr>
<tr>
<td>Adaptive Routing</td>
<td>Congestion Control</td>
<td>Smart Topologies</td>
<td></td>
</tr>
</tbody>
</table>

In-Network Computing

- All-to-All
- MPI Tag Matching
- Data Reductions (SHARP)
- Data processing units (Arm cores)
- Self Healing Network
- Data security / tenant isolation

Adapter/DPU

Switch

End-to-End

End-to-End
10X HIGHER PERFORMANCE WITH GPUDIRECT™ RDMA

No GPUDirect
Network Handled by CPU and CPU-Memory

2 Full Copy Operations
0

2 PCIe Transactions
1

↓ GPU Utilization
↓

↑ CPU Usage
↓

↑ Latency
↓

GPUDirect
Network Goes Directly to GPU Memory

1 Receives incoming packets

(1) Receives incoming packets

(3) DMA Packet from Host Memory to GPU memory

(2) DMA Packet to Host Memory

(4) Trigger the GPU
**INFINIBAND RDMA ADVANTAGES**

**RDMA Latency Node to Node No Switch**
- InfiniBand is 7% Faster vs Ethernet RoCE

**RDMA Latency Node to Node Single Switch**
- InfiniBand is 30% Faster vs Ethernet RoCE

**InfiniBand RDMA (No SHARP) Performance Advantage Over RoCE**
- InfiniBand RDMA Enables 16% Higher BERT AI Performance (4 Nodes)
INFINIBAND ADAPTIVE ROUTING

mpiGraph: Static Routing versus Adaptive Routing
(Summit)

VASP
(GaAsB1512)

Network Efficiency

Network Efficiency

BSMBench
(Balance)

Jobs / day

GFlops

Number of Nodes

Number of Nodes

Static Routing

Adaptive Routing

Static Routing

Adaptive Routing

Static Routing

Adaptive Routing
 HDR INFINIBAND CONGESTION CONTROL

Without InfiniBand Congestion Control

With InfiniBand Congestion Control
INFINIBAND CONGESTION CONTROL
Eliminates Congestion Hot Spots
SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree Based Aggregation Mechanism

Multiple Simultaneous Outstanding Operations

Small Message and Large Message Reduction

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
SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree Based Aggregation Mechanism

Multiple Simultaneous Outstanding Operations

Small Message and Large Message Reduction

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 Aggregation Node: Switch Resident

Host: Data source and Destination
SHARP ALLREDUCE PERFORMANCE ADVANTAGES

7x Higher MPI Performance, 2.5x Higher AI Performance

**Allreduce Latency**

- Lower is Better

**NCLL AllReduce Bandwidth with SHARP**

- 128 NVIDIA DGX A100 (1024 GPUs, 1024 InfiniBand Adapters)

- Higher is Better
INFINIBAND SHARP PERFORMANCE ADVANTAGE OVER ROCE

4X Higher Performance

**MPI AllReduce Latency Performance**

<table>
<thead>
<tr>
<th>Message Size (Bytes)</th>
<th>InfiniBand SHARP</th>
<th>InfiniBand</th>
<th>RoCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>65536</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>131072</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>262144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>524288</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1048576</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MPI AllReduce Latency Performance**

<table>
<thead>
<tr>
<th>Message Size (Bytes)</th>
<th>InfiniBand SHARP</th>
<th>InfiniBand</th>
<th>RoCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
INFINIBAND MPI TAG MATCHING HARDWARE ENGINE

Wait for Software Tag Matching

Posting Buffers with Tags (Expected Messages)

Software

Hardware

Matching List

Unexpected Message

Expected Message

Arriving New Messages

Gather Remote Data

Scatter to Local Buffer

Rendezvous?

YES

NO
# In-Network Computing Accelerated Supercomputing

Software-Defined, Hardware-Accelerated, InfiniBand Network

## In-Network Computing Acceleration Engines

- New generations Introduce and Enhance Acceleration Technologies
- NDR InfiniBand Includes SHARP v3 and All-to-all Engines

## Faster Data Communications

<table>
<thead>
<tr>
<th>Technology</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Data Reduction</td>
<td>7x Faster All-Reduce</td>
</tr>
<tr>
<td>SHARP v2</td>
<td>~15% higher Performance for OpenFOAM, DL-POLY-4, Relion</td>
</tr>
<tr>
<td>Large Data Reduction</td>
<td>2.5x Faster All-Reduce</td>
</tr>
<tr>
<td>SHARP v2</td>
<td>15% Faster Deep Learning Recommend. 17% Faster Natural Language Processing</td>
</tr>
<tr>
<td>MPI Tag Matching</td>
<td>1.8x Faster MPI Iscatterv</td>
</tr>
<tr>
<td>100% Overlapping</td>
<td>Up to 40% Higher Performance for LAMMPS, Nekbone, 3D Stencil</td>
</tr>
<tr>
<td>All-to-All</td>
<td>4x Higher Throughput</td>
</tr>
<tr>
<td>(Introduced with NDR 400G)</td>
<td>Coming Soon with NVIDIA NDR InfiniBand!</td>
</tr>
</tbody>
</table>

## Higher Application Performance

- Large Data Reduction
- Small Data Reduction
- MPI Tag Matching
- All-to-All (Introduced with NDR 400G)
CLOUD-NATIVE SUPERCOMPUTING

Bare-metal Secured Infrastructure
Higher Application Performance
From the Edge to the Main Data Center
CLOUD-NATIVE SUPERCOMPUTING INFRASTRUCTURE

Traditional Supercomputing

- **Host**
  - Applications
  - HPC / AI Communication Frameworks
  - HPC / AI Storage: File System Client
  - Management Isolation Monitoring

- **InfiniBand Adapter**
  - Acceleration Engines

- **InfiniBand Switch**
  - Acceleration Engines

Cloud-Native Supercomputing

- **Host**
  - Applications
  - InfiniBand BlueField DPU
  - HPC / AI Communication Frameworks
  - HPC / AI Storage: File System Client
  - Management Isolation Monitoring
  - Magnus IO
  - DOCA
  - Acceleration Engines

- **InfiniBand Switch**
  - Acceleration Engines
MULTI-TENANT ISOLATION
Zero-Trust Architecture

Secured Network Infrastructure and Configuration
Storage Virtualization
Tenant Service Level Agreement (SLA)
32K Concurrent Isolated Users on Single Subnet

Secure Partitioning with Bare-Metal Performance

NVIDIA BlueField DPU
MORPHEUS AI CYBERSECURITY FRAMEWORK

UFM Cyber-AI with Morpheus for Higher Security

<table>
<thead>
<tr>
<th>Action</th>
<th>Post-Process</th>
<th>Inference</th>
<th>Pre-Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAPIDS</td>
<td>Cyber Log Accelerator (CLX)</td>
<td>Triton</td>
<td>TensorRT</td>
</tr>
<tr>
<td>EGX Stack</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real-Time Telemetry → Updated Policies

DOCA Telemetry

DOCA Flow Policy Engine
HIGHER APPLICATION PERFORMANCE
DPU-Accelerated HPC Communications

Collective Offloads
Active Messages
Smart MPI Progression
Data Compression
User-defined Algorithms
NON-BLOCKING MPI PERFORMANCE

44% Performance Increase for MPI iAlltoall, 36% Performance Increase for MPI iAllgather

Non-Blocking MPI Alltoall Overall Time

Not Offloaded  DPU Accelerated

0  400000  800000  1200000  1600000  2000000
16K  32K  64K  128K  256K

Message Size (Bytes)

Time (usec)

Non-Blocking MPI Allgather Overall Time

Not Offloaded  DPU Accelerated

0  10000  20000  30000  40000  50000
128K  256K  512K  1M  2M

Message Size (Bytes)

Time (usec)

Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions
HIGHER APPLICATION PERFORMANCE
100% Communication - Computation Overlap

Overlap of Communication and Computation with Nonblocking Alltoall

Not Offloaded  DPU Accelerated

<table>
<thead>
<tr>
<th>Message Size</th>
<th>Overlap (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1K</td>
<td>100</td>
</tr>
<tr>
<td>2K</td>
<td>98</td>
</tr>
<tr>
<td>4K</td>
<td>96</td>
</tr>
<tr>
<td>8K</td>
<td>94</td>
</tr>
<tr>
<td>16K</td>
<td>92</td>
</tr>
<tr>
<td>32K</td>
<td>90</td>
</tr>
<tr>
<td>64K</td>
<td>88</td>
</tr>
<tr>
<td>128K</td>
<td>86</td>
</tr>
<tr>
<td>256K</td>
<td>84</td>
</tr>
<tr>
<td>512K</td>
<td>82</td>
</tr>
</tbody>
</table>

Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions

32 servers, Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz (32 processes per node), NVIDIA BlueField-2 HDR 100 DPUs and ConnectX-6 HDR100 adapters, NVIDIA HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand, 256GB DDR4 2400MHz RDIMMs memory and 1TB 7.2K RPM SATA 2.5” hard drive per node.
HIGHER APPLICATION PERFORMANCE

Higher App Performance, MPI Collectives Offload

P3DFFT Performance
- Not Offloaded
- DPU Accelerated

![Graph showing P3DFFT performance with grid size and time in microseconds.](image)

**32 servers, Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz (32 processes per node), NVIDIA BlueField-2 HDR100 DPUs and ConnectX-6 HDR100 adapters, NVIDIA HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand, 256GB DDR4 2400MHz RDIMMs memory and 1TB 7.2K RPM SATA 2.5" hard drive per node.**

Courtesy of: Ohio State University MVAPICH Team and X-ScaleSolutions
FROM THE EDGE TO THE DATA CENTER
Pre-Processing at the Edge, Native InfiniBand RDMA to the Main Supercomputer

EDGE COMPUTING

BlueField DPU

Metrox-2 Long-haul

SUPERCOMPUTING

NETWORK

CLOUD

EXEMPLARY IO
NVIDIA DOCA
Enabling Broad DPU Partner Ecosystem

Software Application Framework for BlueField DPUs
DOCA is for DPUs What CUDA is for GPUs
Protects Developer Investment for Future DPUs
Certified Reference Applications, APIs & Partner Solutions
Rich Partner Ecosystem Across Industries and Workloads

https://developer.nvidia.com/networking/doca
MELLANOX NDR 400G INFINIBAND: NEXT-GENERATION INFINIBAND ARCHITECTURE

ConnectX-7 Adapter
- NDR 400G InfiniBand
- PCIe Gen5
- Programmable Datapath
- In-Network Computing

BlueField-3 DPU
- NDR 400G InfiniBand with Arm Cores
- PCIe Gen5, DDR5
- AI Application Accelerators
- Programmable Datapath
- In-Network Computing

Quantum-2 Switch
- 64-ports NDR 400G InfiniBand
- 128-ports 200G NDR200
- In-Network Computing

Cable
- Copper Cables
- Active Copper Cables
- Optical Transceivers
ANNOUNCING NVIDIA NDR 400G INFINIBAND SYSTEMS
In-Network Computing Accelerated Network for Cloud-Native Supercomputing at Any Scale

- 2x Data Throughput 400 Gigabits per Second
- 32x More AI Acceleration SHARP In-Network Computing
- 6.5x Higher Scalability >1M nodes with DF+ 3 hops
- 5x Switch System Capacity >1.6 Petabit per Second
ANNOUNCING NVIDIA NDR 400G INFINIBAND SYSTEMS

In-Network Computing Accelerated Network for Cloud-Native Supercomputing at Any Scale

15%
Faster Deep Learning Recommendations

17%
Faster Natural Language Processing

15%
Faster Computational Fluid Dynamics Simulations

36%
Lower Power Consumption
ANNOUNCING NVIDIA NDR 400G INFINIBAND SYSTEMS
In-Network Computing Accelerates Cloud-Native Supercomputing at Any Scale

64 NDR Ports
128 NDR200 Ports

512 NDR Ports
1024 NDR200 Ports

1024 NDR Ports
2048 NDR200 Ports

2048 NDR Ports
4096 NDR200 Ports
FROM SUPERCOMPUTERS TO SUPERCLOUDS: CLOUD-NATIVE SUPERCOMPUTERS

DOCA Enabling Growing Partner Ecosystem

Bluefield-3 Next Generation 400G Data Center Infra Processor

NVIDIA Quantum NDR 400G InfiniBand In-network Computing Interconnect