THE NEW SCIENTIFIC COMPUTING WORLD

SUPERCOMPUTER

INFINIBAND NETWORK

EDGE

STORAGE

APPLIANCE
IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

INFINIBAND
Data Processing Unit

Software Defined
Hardware-Accelerated
In-network computing

Pre-configured DPUs
Programmable DPUs

INFINIBAND
DPU

GPU

SUPERCOMPUTER

CPU
IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

INFINIBAND
Data Processing Unit

Software Defined
Hardware-Accelerated
In-network computing

Pre-configured DPUs
Programmable DPUs

DPU
Programmable

DPU cores, MMU
Data pre-processing
User-defined algorithms

Pre-Configured

SHARP (data reductions)
MPI Tag-Matching
SHIELD (self healing resiliency)
NVMe over fabric
Data security and tenant isolations

Speed of Light

200G end-to-end, extremely low latency
RDMA, GPU Direct RDMA, GPU Direct storage
Enhanced Adaptive Routing and Congestion Control
Smart topologies

SUPERCOMPUTER
IN-NETWORK COMPUTING ACCELERATED SUPERCOMPUTING

**INFINIBAND**

Data Processing Unit (DPU)

- Software Defined
- Hardware-Accelerated
- In-network computing

- Pre-configured DPUs
- Programmable DPUs

**OpenSNAPI**

Applications / Services Acceleration

- Switching
- Routing
- Analytics, Telemetry
- Elastic Storage

Hardware Acceleration

- Communication
- Computation
- Storage

Applications and Services

- QoS, Isolation, Protection
- Security
INFINIBAND TECHNOLOGY FUNDAMENTALS

Smart End-Point

Architected to Scale

Centralized Management

Standard
HDR 200G INFINIBAND ACCELERATES NEXT GENERATION HPC AND AI SUPERCOMPUTERS (EXAMPLES)
SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)
SCALABLE HIERARCHICAL AGGREGATION AND REDUCTION PROTOCOL (SHARP)

In-network Tree based aggregation mechanism
Multiple simultaneous outstanding operations
For HPC (MPI / SHMEM) and 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

Providing Flat Latency, 7X Higher Performance
INFINIBAND SHARP AI PERFORMANCE ADVANTAGE

2.5X Higher Performance

128 NVIDIA DGX A100
(1024 GPUs, 1024 InfiniBand Adapters)
NCCL AllReduce Performance Advantage with SHARP
MPI TAG MATCHING
HARDWARE ENGINE
INFINIBAND MPI TAG MATCHING HARDWARE ENGINE

- **Matching List**
  -Arriving New Messages
  -Expected Message
  -Posting buffers with Tags (expected messages)

- **Wait for software Tag Matching**

- **Software**

- **Hardware**

- **Unexpected Message**

- **Expected Message**

- **Rendezvous?**
  -Yes
  -Gather remote data
  -No

- **Scatter to local buffer**
HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

1.8X Higher MPI_Iscatterv Performance on TACC Frontera

Graphs showing latency for 128 Nodes, 256 Nodes, 512 Nodes, and 1024 Nodes. The graphs compare MVAPICH2 and MVAPICH2+HW-TM for different message sizes (16K, 32K, 64K, 128K, 256K, 512K, 1M).
HARDWARE TAG MATCHING PERFORMANCE ADVANTAGES

Nearly 100% Compute - Communication Overlap

Overlap with MPI_Iscatterv (TACC Frontera 256 Nodes)

Overlap with MPI_Iscatterv (TACC Frontera 512 Nodes)
EFFICIENT COMMUNICATION FOR ACCELERATED TRAINING

10X Better Latency & Bandwidth, 3X Faster Deep Learning
QUALITY OF SERVICE
NETWORK CONGESTION TYPES

In-network Congestion

Solution: Adaptive Routing

In-cast Congestion

Solutions: Congestion Control
IN-NETWORK CONGESTION: ADAPTIVE ROUTING

The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems


mpiGraph: Static vs. Adaptive Routing

Static Routing

Adaptive Routing
INFINIBAND CONGESTION CONTROL

Without Congestion Control
- No congestion
- Highest throughput!

With Congestion Control
- Congestion
- Throughput loss

Congestion Factor

GPCNet Benchmark

Random Ring Latency
Random Ring Bandwidth
Allreduce

H1
H2
H3
H4
H5
H6
H7
INFINIBAND ACCELERATED SUPERCOMPUTING

**Speed of Light**
- 200Gb/s Data Throughput
- RDMA and GPUDirect RDMA
- 3X Better (Lower) Latency

**SHARP AI Technology**
- AI Acceleration Engines
- 2.5X Higher AI Performance

**SHIELD AI Technology**
- Self Healing Network
- 1000X Faster Recovery Time

**UFM Cyber AI**
- Data Center Cyber Intelligence and Analytics