In-Network Computing

Paving the Road to Exascale

5th Annual MVAPICH User Group (MUG) Meeting, August 2017
Exponential Data Growth – The Need for Intelligent and Faster Interconnect

CPU-Centric (Onload)

Data-Centric (Offload)

Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!

Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale
Data Centric Architecture to Overcome Latency Bottlenecks

CPU-Centric (Onload)

HPC / Machine Learning Communications Latencies of 30-40us

Data-Centric (Offload)

HPC / Machine Learning Communications Latencies of 3-4us

Intelligent Interconnect Paves the Road to Exascale Performance
In-Network Computing to Enable Data-Centric Data Centers

- Adapters
- Switches
- Co-Processor SmartNIC
- FPGA
- Fat Tree
- Mesh
- Torus
- Hypercube
- Dragonfly+

In-Network Computing Key for Highest Return on Investment
In-Network Computing to Enable Data-Centric Data Centers

In-Network Computing Key for Highest Return on Investment

- RDMA
- GPDirect
- Tag-Matching
- NVMe over Fabrics

- SHARP
- SHIELD

- Security
- NVMe Storage
- Programmable (ARM)

- Programmable (FPGA)

- BlueField

- Inova Next

- ConnectX

- Fat Tree
- Mesh
- Torus
- Hypercube
- Dragonfly+
In-Network Computing Advantages with SHARP Technology

InfiniBand Just Got Smarter

In-Network Computing

10X Performance Acceleration

Critical for High Performance Computing and Machine Learning Applications
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, 16/32/64/128 bits
SHARP Allreduce Performance Advantages

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency
SHIELD – Self Healing Interconnect Technology

InfiniBand Just Got Smarter

Self-Healing Technology

5000X Faster Network Recovery

Enable Unbreakable Data Centers
MPI Tag-Matching Hardware Engines

InfiniBand Just Got Smarter

In-Network Computing

35X Performance Acceleration

Mellanox In-Network Computing Technology Deliver Highest Performance
MPI Tag-Matching Hardware Engines

Application

Critical region

Unexpected Messages List

Post Receive List

Expected Message

Match

Network

Eager

requestor

responder

rendezvous

requestor

responder

Software

Post receive

Not a match

Software finds a match

Hardware

Not a match

Hardware finds a match

Post send

Another post send

Same tag
31% lower latency and 97% lower CPU utilization for MPI operations
Performance comparisons based on ConnectX-5
Multi-Host Socket Direct Technology

Innovative Solution to Dual-Socket Servers
Multi-Host Socket-Direct Adapters Increase Server Return on Investment

30%-60% Better CPU Utilization
50%-80% Lower Data Latency
15%-28% Better Data Throughout

Available for All Servers (x86, Power, ARM, etc.)
Highest Applications Performance, Scalability and Productivity
Multi-Host Socket-Direct Adapters Increase Server Return on Investment

- **30%-60%** Better CPU Utilization
- **50%-80%** Lower Data Latency
- **15%-28%** Better Data Throughput

Available for All Servers (x86, Power, ARM, etc.)

Highest Applications Performance, Scalability and Productivity
InfiniBand Router
InfiniBand Architecture

Native InfiniBand Connectivity Between Different InfiniBand Subnets (Each Subnet can Include 40K nodes)

Isolation Between Different InfiniBand Networks (Each Network can be Managed Separately)

Native InfiniBand Connectivity Between Different Network Topologies (Fat-Tree, Torus, Dragonfly, etc.)
InfiniBand Isolation Enables Great Flexibility

- Enable separation and fault resilience between multiple InfiniBand subnets
- Enable sharing a common storage network by multiple compute infrastructure
- Connect different topologies used by the different subnets
Application Performance

GROMACS Performance
(d.dppc)

NAMD Performance
(Apoa1)

Native InfiniBand Router – Native InfiniBand Performance
Mellanox Solutions

Future Proof Your Data Center
Mellanox to Connect Future #1 HPC Systems (Coral)

“Summit” System

“Sierra” System

Paving the Path to Exascale
Highest-Performance 100Gb/s Interconnect Solutions

Adapters

100Gb/s Adapter, 0.6us latency
175-200 million messages per second
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

Switch

36 EDR (100Gb/s) Ports, <90ns Latency
Throughput of 7.2Tb/s
7.02 Billion msg/sec (195M msg/sec/port)

Switch

32 100GbE Ports, 64 25/50GbE Ports
(10 / 25 / 40 / 50 / 100GbE)
Throughput of 3.2Tb/s

Interconnect

Transceivers

Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100Gb/s)
VCSELs, Silicon Photonics and Copper

Software

MPI, SHMEM/PGAS, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations
Highest-Performance 200Gb/s Interconnect Solutions

Adapters

ConnectX® 6
200Gb/s Adapter, 0.6us latency
200 million messages per second
(10 / 25 / 40 / 50 / 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

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

© 2017 Mellanox Technologies