In Network Computing

Paving the Road to Exascale

May 2017
Exponential Data Growth – The Need for Intelligent and Faster Interconnect

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

CPU-Centric (Onload)

Data-Centric (Offload)

Must Wait for the Data Creates Performance Bottlenecks

Analyze Data as it Moves!
Data Centric Architecture to Overcome Latency Bottlenecks

CPU-Centric (Onload)

Data-Centric (Offload)

HPC / Machine Learning Communications Latencies of 30-40us

HPC / Machine Learning Communications Latencies of 3-4us

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

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

- CPU
- GPU
- CPU
- GPU
- CPU
- GPU

RDMA
CORE-Direct
Tag-Matching
GPUDirect
NVMe over Fabrics

SHARP
SHIELD

Security
Programmable (FPGA)

Security
NVMe Storage
Programmable (ARM)

In-Network Computing Key for Highest Return on Investment
InfiniBand Delivers Best Return on Investment

30%-250% Higher Return on Investment
Up to 50% Saving on Capital and Operation Expenses
Highest Applications Performance, Scalability and Productivity

Weather 1.3X Better
Automotive 2X Better
Chemistry 1.4X Better
Molecular Dynamics 2.5X Better
Genomics 1.3X Better
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

Compute in the Interior of the network
Recursive Doubling vs. SHARP Communication Patterns

Step 1

Recursive Doubling

Step 2

SHARP
SHARP Capabilities

- Reliable Scalable General Purpose Primitive, Applicable to Multiple Use-cases
  - In-network Tree based aggregation mechanism
  - Large number of groups
- Many simultaneous outstanding operations in flight
- Accelerating HPC applications with scalable high performance collective offload
  - Barrier, Reduce, All-Reduce and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16, 32, 64, 128 bit
  - Repeatable results
  - Up to 1024 bytes per reduction
- Significantly reduce MPI collective runtime
- Increase CPU availability and efficiency
- Enable communication and computation overlap
SHARP Advantage - MPI Allreduce Performance

![Graph showing MPI AllReduce Performance](image)

- **Latency (usec)**
- **Number of Hosts (1PPN)**

Legend:
- 8B SHARP
- 128B SHARP
- 8B No SHARP
- 128B No SHARP
SHARP Advantage - MPI Allreduce Performance

![MPI AllReduce Performance Graph](image-url)

- **Latency (usec)**
- **Number of Hosts (1PPN)**

Legend:
- 1024B SHARP
- 2048B SHARP
- 4096B SHARP
- 1024B No SHARP
- 2048B No SHARP
- 4096B No SHARP
OpenFOAM is a popular computational fluid dynamics application.

**OpenFOAM : Lid Driven Cavity Flow**
icoFoam solver, 2D 1 million cells

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Intel MPI</th>
<th>Open MPI</th>
<th>HPC-X (SHArP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>400</td>
<td>600</td>
</tr>
<tr>
<td>16</td>
<td>400</td>
<td>600</td>
<td>800</td>
</tr>
<tr>
<td>32</td>
<td>600</td>
<td>800</td>
<td>1000</td>
</tr>
<tr>
<td>64</td>
<td>800</td>
<td>1000</td>
<td>1200</td>
</tr>
</tbody>
</table>

**HPC-X with SHArP Delivers 2.2X Higher Performance over Intel MPI**
Tag Matching

Network Edge Computing
MPI Tag Matching Terminology

- **Sender** – exposes data to be sent to a specified target
- **Receiver** – exposes a data buffer that is available to accept data
- **Expected receive** – data arrives from sender to already available receive (implementation term)
- **Unexpected receive** – No receive posted to match data sent by a sender (implementation term)
- **Eager send** – small message that is sent with the message envelope (implementation term)
- **Rendezvous protocol** – protocol for keeping the amount of temporary buffers needed by the communication library small (implementation term)
Tag Matching – Common Implementation Protocols

Eager
requestor responder

rendezvous
requestor responder

Rendezvous completion

Eager control
RDMA_READ
Read Response
FIN
31% lower latency and 97% lower CPU utilization for MPI operations
- Performance comparisons based on ConnectX-5
The Generation of In-Network Computing – 10X Higher Performance

In-Network Computing Key for Highest Return on Investment
Thank You