High Performance MPI over Slingshot

8/23/2022

Network-based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University

MUG 2022

Kawthar Shafie Khorassani
shafiekhorassani.1@osu.edu
Introduction

• Frontier at OLCF (#1 Supercomputer on Top500 System) - deployed with Slingshot-11 networking across nodes
• MPI-level communication and performance on upcoming networking for exascale systems (i.e. Frontier & El-Capitan)

#1 Supercomputers Top500 Over Time …

Sunway TaihuLight
(‘16-’17)  Summit
(‘18-’19)  Fugaku
(‘20-’21)  Frontier
(‘22)
Top500 Supercomputers Interconnect Statistics

Reference: https://www.top500.org
Background

• Many Supercomputers deployed with Mellanox Infiniband Interconnect technology
• MPI Libraries have been optimized over the years to expand on Mellanox Infiniband features and support
• Underlying interconnect technology critical for achieving low latency and high throughput at scale on next-generation exascale systems

Drive future research and innovations to provide scalable and competitive options in this Slingshot ecosystem.
Slingshot Interconnect

High-performance network designed by HPE Cray for upcoming exascale-era systems

- Based on Ethernet
- Adaptive Routing
- Congestion Control
- Isolated Workloads

**Empowering the #1 Supercomputer --- Frontier**

- Deployed as the interconnect for inter-node communication
- Expected to be deployed on upcoming supercomputers --> El-Capitan at LLNL, Aurora at Argonne
Limitations of State-of-the-art Approaches for Communication

Current accessibility and deployment on early access Slingshot systems:

- Ecosystem with Slingshot-10 interconnection amongst nodes
- Slingshot-10 running over a Slingshot Network with a Mellanox Infiniband adapter

Future accessibility and deployment on upcoming Slingshot systems (i.e. El-Capitan and Frontier):

- Slingshot-11
- Deployed over a slingshot fabric and adapter

This second-generation deployment introduces additional challenges for communication libraries to develop functionality over the underlying adapter and fabrics.
Experimental Setup

System & Software Details
Spock Compute Node

Reference: https://docs.olcf.ornl.gov/systems/spock_quick_start_guide.html
Frontier Compute Node

- 1 HPC and AI Optimized 3rd Gen AMD EPYC CPU
- 4 Purpose Built AMD Instinct 250X GPUs
- **CPU-GPU Interconnect**: AMD Infinity Fabric
- **System Interconnect**: Multiple Slingshot NICs providing 100 GB/s network bandwidth.
  - Slingshot network which provides adaptive routing, congestion management and quality of service.

Reference: [https://www.olcf.ornl.gov/frontier/](https://www.olcf.ornl.gov/frontier/)
Software Details

MPI & Communication Libraries

- CrayMPICH 8.1.14
  - [https://docs.nersc.gov/development/programming-models/mpi/cray-mpich/](https://docs.nersc.gov/development/programming-models/mpi/cray-mpich/)
- MVAPICH2-GDR 2.3.7 & MVAPICH2-X 2.3 & MVAPICH2-3.0a
  - [https://mvapich.cse.ohio-state.edu](https://mvapich.cse.ohio-state.edu)
- OpenMPI 4.1.4 + UCX 1.12.1
  - [https://www.open-mpi.org](https://www.open-mpi.org)
- RCCL 5.0.2
  - [https://github.com/ROCmSoftwarePlatform/rccl](https://github.com/ROCmSoftwarePlatform/rccl)

OSU Microbenchmarks 5.9

- [https://mvapich.cse.ohio-state.edu/benchmarks/](https://mvapich.cse.ohio-state.edu/benchmarks/)

ROCm version 5.0.2
Experiment Details

CrayMPICH 8.1.14
– Module load cray-mpich/8.1.14
– Module load craype-accel-amd-gfx908
– Run: MPICH_GPU_SUPPORT_ENABLED=1

MVAPICH2-3.0a
– Configure: --with-device=ch4:ofi --with-libfabric=<path-to-libfabric>

MVAPICH2-GDR 2.3.7
– Run: MV2_USE_ROCM=1

OpenMPI 4.1.4 + UCX 1.12.1
– Compile UCX: --with-rocm=<path-to-rocm> --without-knem --without-cuda --enable-optimizations
– Compile OpenMPI: --with-ucx=<path-to-ucx> --without-verbs
– Run: -x UCX_RNDV_THRESH=128

RCCL 5.0.2
– Compile: CXX=<path-to-rocm>/bin/hipcc
Performance Evaluation

CPU
Point-to-Point Performance - Intra-Node CPU

**Peak Bandwidth:**
- MVAPICH2-X 39.2 GB/s
- OpenMPI+UCX 38.2 GB/s
- CrayMPICH 42 GB/s

**Latency at 4 Bytes:**
- MVAPICH2-X 0.22 us
- OpenMPI+UCX 0.31 us
- CrayMPICH 0.27 us

**AMD Epyc Rome CPUs on Spock System**

Point-to-Point Performance - Inter-Node CPU

Latency (small messages)

Latency (large messages)

Bandwidth

Bi-Directional Bandwidth

Slingshot-10 Interconnect for over network communication (12.5+12.5 GB/s)

Peak Bandwidth:
- MVAPICH2-X 122.4 MB/s
- OpenMPI+UCX 122.4 MB/s
- CrayMPICH 122.4 MB/s

Latency at 4 Bytes:
- MVAPICH2-X 2.55 us
- OpenMPI+UCX 2.27 us
- CrayMPICH 2.07 us

AMD Epyc Rome CPUs on Spock System

Collectives Performance - CPU

GATHER

256 CPUs - 4 Nodes & 64 PPN on Spock System

Collectives Performance - CPU

**REDUCE**

- **Reduce (small messages)**
  - MVAPICH2-X 2.3
  - OpenMPI 4.1.4 + UCX 1.12.1
  - Cray MPICH 8.1.14
  - Latency (μs) versus Message Size (Bytes)
  - 205 μs at 4K bytes

- **Reduce (large messages)**
  - MVAPICH2-X 2.3
  - OpenMPI 4.1.4 + UCX 1.12.1
  - Cray MPICH 8.1.14
  - Latency (ms) versus Message Size (Bytes)
  - 1.60 ms at 1M bytes

**ALLREDUCE**

- **Allreduce (small messages)**
  - MVAPICH2-X 2.3
  - OpenMPI 4.1.4 + UCX 1.12.1
  - Cray MPICH 8.1.14
  - Latency (μs) versus Message Size (Bytes)
  - 23.37 μs at 4K bytes

- **Allreduce (large messages)**
  - MVAPICH2-X 2.3
  - OpenMPI 4.1.4 + UCX 1.12.1
  - Cray MPICH 8.1.14
  - Latency (ms) versus Message Size (Bytes)
  - 4.46 ms at 1M bytes

**256 CPUs - 4 Nodes & 64 PPN on Spock System**

Performance Evaluation

GPU
Point-to-Point Performance - Intra-Node GPU

Latency (small messages)

- MVAPICH2-GDR 2.3.7
- OpenMPI 4.1.4 + UCX 1.12.1
- Cray MPICH 8.1.14

Latency (large messages)

- MVAPICH2-GDR 2.3.7
- OpenMPI 4.1.4 + UCX 1.12.1
- Cray MPICH 8.1.14

Bandwidth (small messages)

- MVAPICH2-GDR 2.3.7
- OpenMPI 4.1.4 + UCX 1.12.1
- Cray MPICH 8.1.14

Bandwidth (large messages)

- MVAPICH2-GDR 2.3.7
- OpenMPI 4.1.4 + UCX 1.12.1
- Cray MPICH 8.1.14

All GPUs connected by Infinity Fabric (46+46GB/s)

- PCI Bar Mapped Memory for small message sizes.
- ROCm Inter-Process Communication (IPC) used in med-large message range.

Peak Bandwidth:
- MVAPICH2-GDR 52.5 GB/s
- OpenMPI+UCX 30.2 GB/s
- CrayMPICH 88 GB/s

Latency at 4 Bytes:
- MVAPICH2-GDR 2.01 us
- OpenMPI+UCX 3.79 us
- CrayMPICH 2.44 us

MI100 GPUs on Spock System

Point-to-Point Performance - Inter-Node GPU

Latency (small messages)

Latency (large messages)

Bandwidth

Bi-Directional Bandwidth

Slingshot-10 Interconnect for over network communication (12.5+12.5 GB/s)

Peak Bandwidth:
- MVAPICH2-GDR 9.9 GB/s
- OpenMPI+UCX 9.8 GB/s
- CrayMPICH 9.2 GB/s

Latency at 4 Bytes:
- MVAPICH2-GDR 3.73 us
- OpenMPI+UCX 4.23 us
- CrayMPICH 3.8 us

MI100 GPUs on Spock System

Collectives Performance - GPU

**BROADCAST**

<table>
<thead>
<tr>
<th>Latency (us)</th>
<th>Message Size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH2-GDR 2.3.7</td>
<td>150</td>
</tr>
<tr>
<td>OpenMPI 4.1.4 + UCX 1.12.1</td>
<td>100</td>
</tr>
<tr>
<td>Cray MPICH 8.1.14</td>
<td>70</td>
</tr>
<tr>
<td>RCCL 5.0.2</td>
<td>50</td>
</tr>
</tbody>
</table>

Broadcast (small messages)

<table>
<thead>
<tr>
<th>Latency (us)</th>
<th>Message Size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH2-GDR 2.3.7</td>
<td>200</td>
</tr>
<tr>
<td>OpenMPI 4.1.4 + UCX 1.12.1</td>
<td>150</td>
</tr>
<tr>
<td>Cray MPICH 8.1.14</td>
<td>100</td>
</tr>
<tr>
<td>RCCL 5.0.2</td>
<td>70</td>
</tr>
</tbody>
</table>

Broadcast (large messages)

**ALLTOALL**

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Message Size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH2-GDR 2.3.7</td>
<td>8.1</td>
</tr>
<tr>
<td>OpenMPI 4.1.4 + UCX 1.12.1</td>
<td>5</td>
</tr>
<tr>
<td>Cray MPICH 8.1.14</td>
<td>3</td>
</tr>
<tr>
<td>RCCL 5.0.2</td>
<td>1</td>
</tr>
</tbody>
</table>

Alltoall (small messages)

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Message Size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH2-GDR 2.3.7</td>
<td>764.28</td>
</tr>
<tr>
<td>OpenMPI 4.1.4 + UCX 1.12.1</td>
<td>642.52</td>
</tr>
<tr>
<td>Cray MPICH 8.1.14</td>
<td>518.75</td>
</tr>
<tr>
<td>RCCL 5.0.2</td>
<td>415.2</td>
</tr>
</tbody>
</table>

Alltoall (large messages)

**64 GPUs - 16 Nodes & 4 GPUs Per Node on Spock System**

Collectives Performance - GPU

GATHER

ALLGATHER

64 GPUs - 16 Nodes & 4 GPUs Per Node on Spock System

Collectives Performance - GPU

**REduce**

```
<table>
<thead>
<tr>
<th>Message Size (Bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3.31</td>
</tr>
<tr>
<td>8</td>
<td>14.35</td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td></td>
</tr>
<tr>
<td>1K</td>
<td></td>
</tr>
<tr>
<td>2K</td>
<td></td>
</tr>
<tr>
<td>4K</td>
<td></td>
</tr>
</tbody>
</table>
```

**AllReduce**

```
<table>
<thead>
<tr>
<th>Message Size (Bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8K</td>
<td></td>
</tr>
<tr>
<td>16K</td>
<td></td>
</tr>
<tr>
<td>32K</td>
<td></td>
</tr>
<tr>
<td>64K</td>
<td></td>
</tr>
<tr>
<td>128K</td>
<td></td>
</tr>
<tr>
<td>256K</td>
<td></td>
</tr>
<tr>
<td>512K</td>
<td></td>
</tr>
<tr>
<td>1M</td>
<td>918 us</td>
</tr>
</tbody>
</table>
```

**64 GPUs - 16 Nodes & 4 GPUs Per Node on Spock System**

Performance Evaluation
Slingshot-11

CPU
Point-to-Point Performance - Intra-Node CPU

**LATENCY**

![Latency Graph](image)

**BANDWIDTH**

![Bandwidth Graph](image)

**BI-BANDWIDTH**

![Bi-Bandwidth Graph](image)

*System with Slingshot-11 Networking*
Point-to-Point Performance - Inter-Node CPU

LATENCY

![Latency Graph]

BANDWIDTH

![Bandwidth Graph]

BI-BANDWIDTH

![BI-Bandwidth Graph]

System with Slingshot-11 Networking
THANK YOU!

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/