Evaluation of Hardware-Based MPI Acceleration on Astra

Michael Aguilar, Kevin Pedretti, Si Hammond, James Laros III, Andrew Younge, Matthew Curry

Sandia National Laboratories

[ March 19, 2019 ]

SAND Number: SAND2019-2774 C

Sandia National Laboratories is a multimission laboratory operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration. Sandia Labs has major research and development responsibilities in nuclear deterrence, global security, defense, energy technologies and economic competitiveness, with main facilities in Albuquerque
EVALUATION OF HARDWARE-BASED MPI ACCELERATION ON ASTRA

• Astra HPC System
• Astra InfiniBand Network
  • Overview
  • Implementation of MPI Hardware Collectives on Astra
    • Overview
    • Specifics
• Results
• References
• Questions?
ASTRA
HPE Apollo 70 Chassis: 4 nodes

HPE Apollo 70 Rack

18 chassis/rack
72 nodes/rack
3 IB switches/rack (one 36-port switch per 6 chassis)

36 compute racks (9 scalable units, each 4 racks)
2592 compute nodes (5184 TX2 processors)
3 IB spine switches (each 540-port)
• **2,592** HPE Apollo 70 compute nodes
  • Cavium Thunder-X2 **Arm** SoC, 28 core, 2.0 GHz
  • 5,184 CPUs, 145,152 cores, 2.3 PFLOPs system peak
  • 128GB DDR Memory per node (**8 memory channels per socket**)  
  • Aggregate capacity: 332 TB, Aggregate Bandwidth: 885 TB/s
  • Mellanox IB EDR, ConnectX-5
  • HPE Apollo 4520 All–flash storage, Lustre parallel file-system  
    • Capacity: 403 TB (usable)
    • Bandwidth 244 GB/s
ASTRA INFINIBAND NETWORK

- Mellanox mlx5-100Gb/s
- Socket Direct

- SHARP

36 Port Switch with an extra port for SHARP Hardware Collectives (Scalable Hierarchical Aggregation and Reduction Protocol)
OpenMPI-3.1.3 compiled to use the UCX API and Hardware Collective

UCS (Service)

Verbs

SHARP Hardware Collectives Tree

--with-hcoll=/opt/Mellanox/sharp
--with-ucx
fca hcoll monitoring portals4'
mca/coll/hcoll mca/coll/monitoring
mca/coll/portals4'
OPAL_CONFIGURE_CLI='""--with-hcoll=/opt/mellanox/hcoll" ""--
IMPLEMENTATION OF HARDWARE COLLECTIVES ON ASTA

- Enable SHARP in opensm.conf

- SHARP Aggregation Manager on the Subnet Manager node defaults to Fat-Tree Topology

- Daemons on compute nodes

```
sharp_enabled 2
routing_engine ftree,updn

[/opt/mellanox/sharp/bin/sharp_am -
/opt/mellanox/sharp/conf/sharp_am.cfg

Package: sharp-rc
- Version: 1.7.2
```
We chose to do our test runs with each MPI endpoint consisting of a complete node.
Runs were made with SLURM
Runs were toggled with Hardware Collectives On/Off
Our tests were done using IMB Benchmark, compiled for ARM64----AllReduce

```
hcoll=-x HCOLL_ENABLE_SHARP=1'       # Probe SHArP and use it (Barrier, Allreduce)
hcoll+="-x SHARP_COLL_LOG_LEVEL=2'   # verbose logging at 5
hcoll+="-x HCOLL_BCOL_P2P_ALLREDUCE_SHARP_MAX=4096'   # Allows larger messages with SHArP, 4096 apparently is the maximum
hcoll+="-x SHARP_COLLJOB_QUOTA_OSTS=256'   # The maximum number of Outstanding Messages
hcoll+="-x SHARP_COLLJOB_QUOTA_MAX_GROUPS=4'   # The number of Collective Groups that can be created
hcoll+="-x SHARP_COLLJOB_PAYLOAD_PER_OST=256'   # Fragment size for large messages.
hcoll+="-x SHARP_COLLJOB_MEMBER_LIST_TYPE=2'
hcoll+="-x HCOLL_BCOL_P2P_ALLREDUCE_SHARP_MAX=4096'   # Maximum Allreduce size run through SHArP
hcoll+="-x HCOLL_MAIN_IB=mlx5_0:1'    # The SHArP HCA enabled tree entry point

#echo $hcoll

mpirun -v -np 4 -mca pml ucx -x UCX_NET_DEVICES=mlx5_0:1 -mca coll_hcoll_enable 1 $hcoll //imb_benchmark/mpi-benchmarks-master/src_c/IMB-MPI1

#mpirun -v -np 4 -mca pml ucx -x UCX_NET_DEVICES=mlx5_0:1 //imb_benchmark/mpi-benchmarks-master/src_c/IMB-MPI1
```
Results

- IMB Benchmark

Allreduce 2 Processes 1 Process/Node

<table>
<thead>
<tr>
<th>t avg[sec]</th>
<th>Number of bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1e+07</td>
</tr>
<tr>
<td></td>
<td>1e+06</td>
</tr>
<tr>
<td></td>
<td>1e+05</td>
</tr>
<tr>
<td></td>
<td>1e+04</td>
</tr>
<tr>
<td></td>
<td>1e+03</td>
</tr>
<tr>
<td></td>
<td>1e+02</td>
</tr>
<tr>
<td></td>
<td>1e+01</td>
</tr>
<tr>
<td></td>
<td>1e+00</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Sharp Off
Sharp On
Results

- IMB Benchmark

![Graph showing Allreduce 1024 Processes 1 Process/Node](image)

- Sharp Off
- Sharp On

Number of bytes vs. Average time [usec]

- Y-axis: t_avg[usec]
- X-axis: Number of bytes

OpenFabrics Alliance Workshop 2019
Results

- IMB Benchmark

Allreduce 2048 Processes 1 Process/Node

\[ t_{avg}[\text{sec}] \]

Number of bytes

- Sharp Off
- Sharp On
Results

- IMB Benchmark

![Graph showing Allreduce 114688 Processes 56 processes/node with Sharp Off and Sharp On lines.](image)
References

- Brightwell, Barron, Hemmert—Challenges for High-Performance Networking for Exascale Computing
- Graham, Bloch, Burddy, Shainer, Smith---Towards a Data-Centric System Architecture SHARP
- Bureddy---SHARP: In-Network Scalable Hierarchical Aggregation and Reduction Protocol
Questions?
Exceptional Service in the National Interest