CSSI Frameworks: Designing Next-Generation MPI Libraries for Emerging Dense GPU Systems

PI: Dhabaleswar K. Panda, Co-PIs: Amitava Majumdar,, Bill Barth, Karen Tomko
Institutions: Ohio State University, San Diego Supercomputer Center, Texas Advanced Computing Center, Ohio Supercomputing Center

Award #: OAC-1931537

Research Challenge

How can existing production quality MPI middleware be enhanced to take advantage of emerging networking and storage technologies to deliver best possible scale-up and scale-out for HPC and DL applications on emerging dense GPU systems?

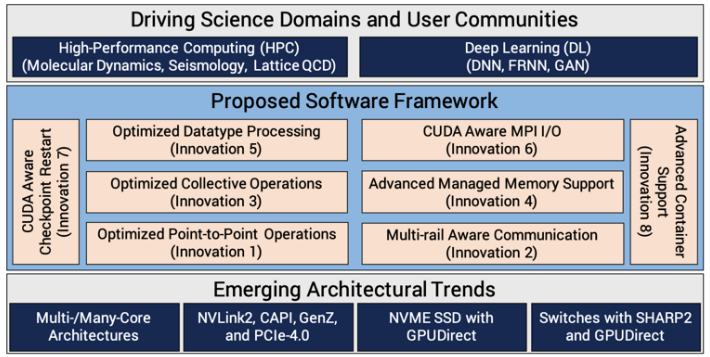
Development Strategy

- Develop high-performance and scalable Pt-to-pt. communication for GPU and CPU buffers
- Enhance and optimize communication to fully utilize multiple HCAs
- Accelerate GPU-based collective communication by utilizing in-network computing features like SHARP
- Employ novel solutions, like datatype processing and unified memory to improve application performance
- Develop a CUDA-aware I/O subsystem
- Add support for containerized environments to enable easy deployment
- Carry out integrated development and evaluation to ensure proper integration

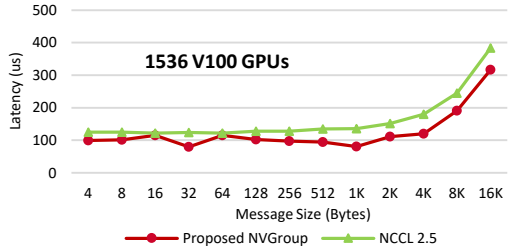
Mapping Applications to Proposed Innovations

	Innovations							
Application	1	2	3	4	5	6	7	8
Amber	✓	✓	✓	✓	✓		✓	
Gromacs	✓	✓	✓	✓				
Hoomd-Blue	✓	✓	✓	✓				
QUDA	✓	✓		✓	✓			
AWP-ODC	✓	✓		✓	✓	✓	✓	
SRGAN	✓	✓	✓			✓		✓
FRNN	✓	✓	✓			✓		✓
TensorFlow		✓	✓					✓

Proposed Framework

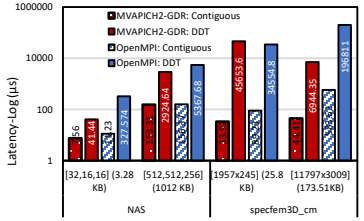


Innovation: Optimized Collective Operations



Innovation: Optimized Datatype Processing

- Architecture-aware designs for GPU-based non-contiguous datatype processing
- Zero-copy based *load-store* semantics for high-performance derived datatype communication on modern dense GPU systems

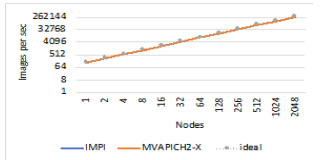


Software Release, Community Engagement and Metrics

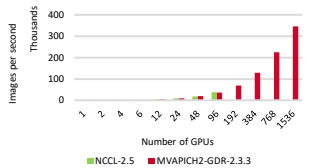
- MVAPICH2 2.3.3 GDR Release on 01/09/2020
- http://mvapich.cse.ohio-state.edu
- Installed on many GPU clusters worldwide (including LLNL, ORNL, SDSC, TACC, Juelich, ABCI, and Facebook)
- More than 1,200 downloads during the first three weeks of the release
- Tutorials at SC '19, HiPEAC '20, PPoPP '20*, GTC '20*, and ISCA '20* (* - To be presented)
- Community Engagement with LLNL, ORNL, Juelich, and Facebook

Deep Learning on CPUs and GPUs

2048 Nodes on Frontera using MVAPICH2-X



1536 Nodes on Summit using MVAPICH2-GDR



Research Publications

1. C-H Chu, J. Hashmi, K. S. Khorassani, H. Subramoni and D. K. Panda, "High-Performance Adaptive MPI Derived Datatype Communication for Modern Multi-GPU Systems," *26th IEEE International Conference on High Performance Computing, Data, Analytics and Data Science (HiPC '19)*, Hyderabad, India, Dec 17-20, 2019.
2. A. Jain, A. A. Awan, H. Subramoni and D. K. Panda, "Scaling TensorFlow, PyTorch, and MXNet using MVAPICH2 for High-Performance Deep Learning on Frontera," *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*, Denver, CO, USA, 2019, pp. 76-83. doi: 10.1109/DLS49591.2019.00015
3. A. A. Awan, A. Jain, Q. Anthony, H. Subramoni, and DK Panda, "HyPar-Flow: Exploiting MPI and Keras for Scalable Hybrid-Parallel DNN Training with TensorFlow", (Under Review).

Future Work

As part of future work we aim to:

- Enhance Pt-to-pt. and Collectives
- Design and Implement Sparse Reduction Collectives
- Optimize Unified-Memory based Communication
- Introduce Support for Containerized Environments

Supported by:
OAC-1931537, OAC-1931450, and OAC- 1931354