A Novel Framework for Efficient Offloading of Communication Operations to Bluefield SmartNICs

Presentation at the 11th Annual MVAPICH User Group (MUG) Conference (MUG ‘23)

Kaushik Kandadi Suresh
The Ohio State University
kandadisuresh.1@osu.edu

Introduction: HPC, MPI, Overlap

• MPI is the de-facto programming model in High Performance Computing (HPC)

• HPC applications have computation and communication

• HPC application performance can be improved by overlap

• MPI non-blocking primitives allows compute and communication overlap

• Progression of communication is needed to achieve overlap
Introduction: Overlap in Rendezvous Protocol

- Application processes schedule communication operation
- Application process free to perform useful compute in the foreground
- Little communication progress in the background
- All communication takes place at final synchronization
- Reduced buffer requirement
- Good communication performance if used for large message sizes and operations where communication library is progressed frequently
- Poor overlap of computation and communication => Poor Overall Application Performance
Introduction: Different ways of overlapping computation and communication

(a) MPI_Tester
(b) MPICH
(c) MVAPICH2
(d) Proposed
Background: BlueField DPU / Smart NIC Architecture

• BlueField includes the ConnectX6 network adapter and data processing cores

• System-on-chip containing 64-bit ARMv8 A72

• BlueField DPU has two modes of operation:
  • Separated Host mode
    • The ARM cores can appear on the network as any other host and the main CPU
Motivation: Problem with the existing Offload framework

- BluesMPI\cite{1} is a prior work that offloads certain MPI collectives to the DPU
  - Eg: Ring based broadcast in HPL

Staged offload by DPU requires 2 RDMA operations:
  - Local-Host-to-DPU Read, DPU-to-Remote-Host Write

\cite{1} Mohammadreza Bayatpour, Nick Sarkauskas, Hari Subramoni, Jahanzeb Maqbool Hashmi, and Dhabaleswar K. Panda. 2021. BluesMPI: Efficient MPI Non-blocking Alltoall Offloading Designs on Modern BlueField Smart NICs. In High Performance Computing: 36th International Conference, ISC High Performance 2021
Guest Virtual Machine ID (GVMI) is a capability provided by the Bluefield DPUs
- Allows DPU process to move data from one local to any remote host process without staging.

Introduces addition overheads:
- Host-level, DPU-level memory registrations and key-exchanges

We provide efficient designs by amortizing the GVMI overheads
Benchmark Results: MPI_Ialltoall

- OSU Microbenchmarks (OMB)
- 16 Nodes 32 PPN
- Proposed* Scheme at-least 25% better than BluesMPI
- Reason for improvements:
  - ~100% overlap
  - Absence of staging overhead

* Our Designs are available in the MVAPICH2-DPU library
Application Results: P3DFFT, HPL

- **P3DFFT**
  - Problem Size:
    - 16 Nodes 32 PPN
    - Proposed* Scheme at-least 20% better than BluesMPI

- **HPL**
  - Problem Size:
    - 16 Nodes 32 PPN
    - Proposed* Scheme at-least 8% better than HPL-1ring

* Our Designs are available in the MVAPICH2-DPU library and HPL-DPU
Conclusion & Future Work

• Conclusion
  – DPU based communication progression better than host-based progression
  – Offloading MPI non-blocking primitives using GVMI
  – Showed Application-level improvements
    • HPL, P3DFFT

• Future Work
  – Accelerate additional applications such as Octopus
  – Offload OpenSHMEM based applications
THANK YOU!

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

MVAPICH
MPI and PGAS Library
http://mvapich.cse.ohio-state.edu/

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

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