Poster ID: **P61807** | Topic / Category: **Networking - Networking for Scientific Computing Workloads**

Benjamin Michalowicz: michalowicz.2@osu.edu
Steve Poole: swpoole@lanl.gov
Kaushik Kandadi Suresh: kandadisuresh.1@osu.edu
Hari Subramoni: subramoni.1@osu.edu
Dhabaleswar K (DK) Panda: panda.2@osu.edu

MVAPICH
http://mvapich.cse.ohio-state.edu

DPU-Bench: A New Microbenchmark Suite to Measure the Offload Efficiency of SmartNICs

Ben Michalowicz[1], Kaushik Kandadi Suresh[1], Hari Subramoni[1], Dhabaleswar K. Panda[1], Steve Poole[2]
[1]The Ohio State University, [2]Los Alamos National Laboratory
{michalowicz.2, kandadisuresh.1, subramoni.1, panda.2} @osu.edu
swpoole@lanl.gov

RESEARCH MOTIVATION

NVIDIA's BlueField DPU and others are becoming widespread in HPC clusters. Because of this, we need a DPU-aware micro-benchmark suite to determine how efficient they are in offloading communication operations. Previous research has designed ways to offload computation, communication, and deep learning to DPUs, but no efforts have been made along the micro-benchmark side.

CHALLENGES AND GOALS

Research Challenges

Given a collective communication pattern, message sizes, number of processes on a given server, and a number of worker-based processes placed on a DPU, can we accurately measure the offload potential from placing communication on them? Furthermore, can we empirically determine a sweet spot for work to be offloaded to demonstrate maximum efficiency?

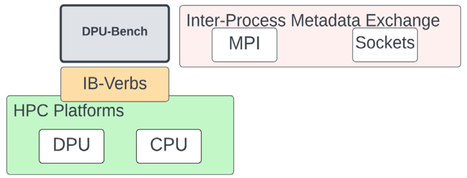
Research Goals

- Design a low-level benchmark suite to analyze the efficiency of offloading collective communication patterns to SmartNICs
- Examine simple algorithms for each pattern and empirically determine the number of DPU-based "workers" that would give optimal offload efficiency.
- Explore efficient/non-efficient algorithm designs to showcase what may happen if offload schemes are made inefficiently.

THE NEED FOR A NEW MICROBENCHMARK SUITE

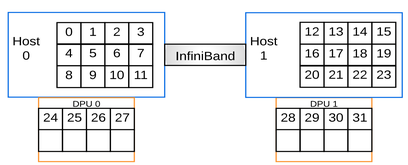
- All the micro/benchmark suites that exist today are NOT DPU-aware (OMB, IMB, OpenHPCA, SMB, etc.) That is, a standard MPI library will not know whether a process is placed on a CPU or a DPU and run operations naively. Previous works have offloaded communication, computation, and Deep Learning to DPUs from the context of applications and libraries.
- With SmartNICs becoming more widespread, we need more ways of measuring their efficacy in the context of HPC and Datacenter environments.

DPU-BENCH: DESIGN AND IMPLEMENTATION



- Why not MPI? Naively placing processes on a host server and a DPU will result in message progression being done on both pieces of hardware for nonblocking communication. Progress will become a bottleneck
- Why IB-Verbs? RDMA semantics → All operations can be issued from the DPU with no message progression being performed on the host server – making nonblocking communication through network primitives with the use of MPI for process management tool.

- Offload Efficiency: max(reference_time/max(pure_comm,compute)) * 100
- Several assumptions made about runtime:
  - Block style hostfile → Higher-numbered processes are on the DPU
  - Use of Multi-Program/Multi-Data mode in MPI libraries

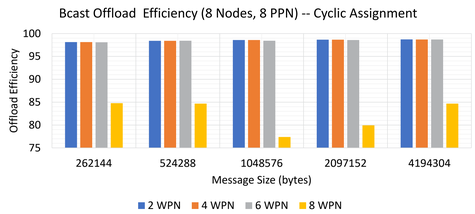


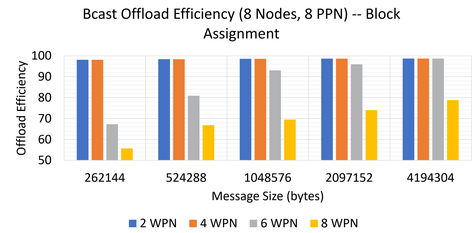
- MPMD-Mode + Block Hostfile: Helps organize config file passed in at runtime

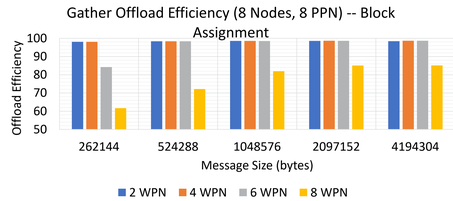
COLLECTIVES EXPLORED

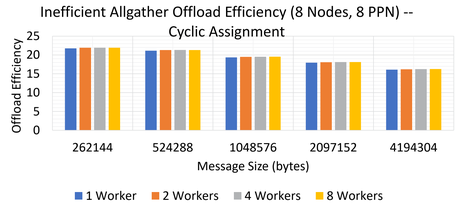
- Current Work: Linear algorithms for broadcast, reduce, and allgather
  - Allgather features an inefficient and an efficient design
- Explore cyclic and block distribution of work and impact on load balancing and work distribution

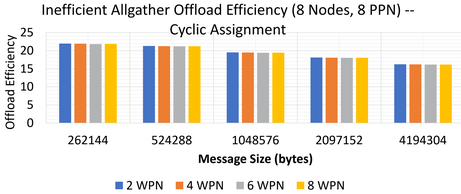
EXPERIMENTS AND SETUP

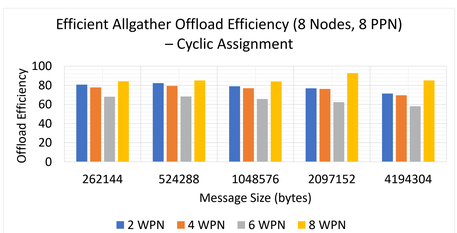












HPC-AI Advisory Council Cluster – "Thor" Partition

Running at 8-Nodes, 8-PPN on the host side, up to 64 workers total (8 WPN) on the DPU side, on messages ranging from 256KB to 4MB

Inefficient Allgather Design: Place gather and broadcast-like algorithms back-to-back and use one worker process as the leader among worker processes

Efficient Design: Perform buffer tagging so multiple workers can write to the same host process without overwriting each other

FUTURE WORK

- Advanced algorithms
- Generalize using UCX
- Generalize to other programming models
- Generalize to other SmartNICs

REFERENCES

1) DPU-Bench: A Micro-Benchmark Suite to Measure Offload Efficiency Of SmartNICs. B. Michalowicz, K. Suresh, H. Subramoni, DK Panda, and S. Poole, Practice and Experience in Advanced Research Computing 23, Jul 2023