2021 OFA Virtual Workshop

EFFICIENT MPI OFFLOADING DESIGNS ON MODERN BLUEFIELD SMART NICS

Mohammadreza Bayatpour, Nick Sarkauskas, Hari Subramoni, and Dhabaleswar K. (DK) Panda

The Ohio State University

Email:

{bayatpour.1, Sarkauskas.1}@osu.edu

{subramon, panda}@cse.ohio-state.edu

INTRODUCTION, MOTIVATION, AND CHALLENGE

- **HPC applications require high-performance, low overhead non-blocking collective communications supports that provide**
  - Low pure communication latency
  - High bandwidth
  - Minimum contention for host CPU resources to progress the collective
  - High overlap of computation with communication
- **CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation**

➔ **Network offload mechanisms are gaining attraction as they have the potential to completely offload the communication of MPI primitives into the network**

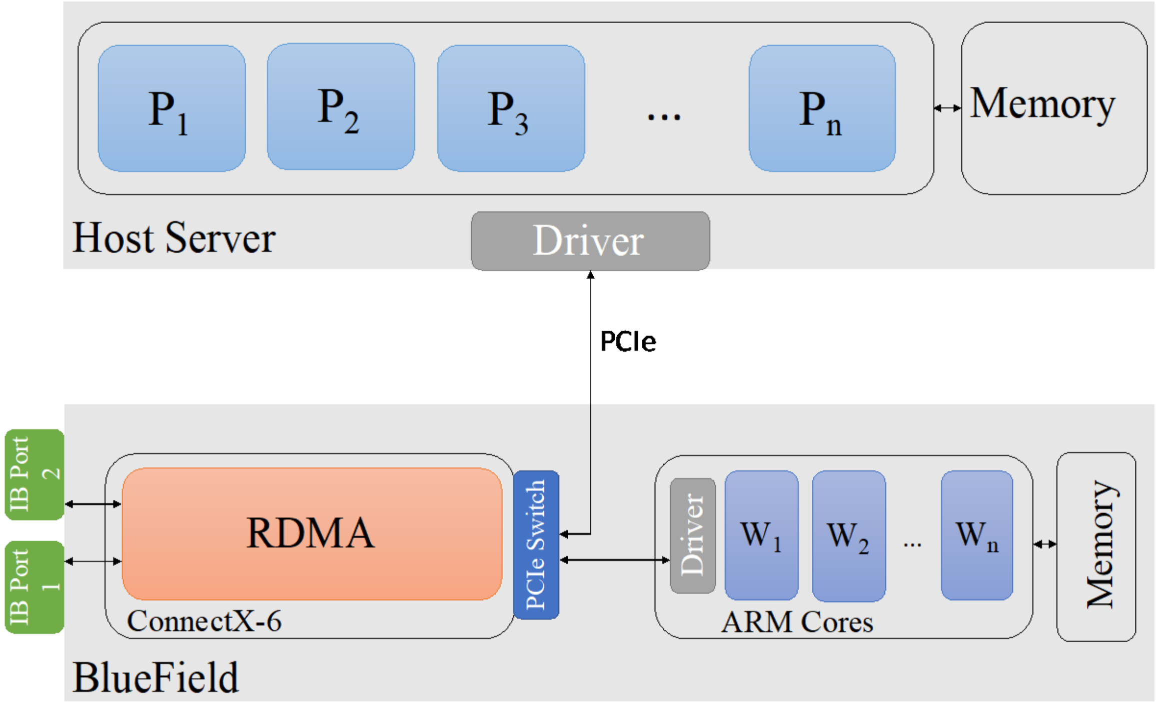
INTRODUCTION, MOTIVATION, AND CHALLENGE (CONT'D)

- The area of network offloading of MPI primitives is still nascent and cannot be used as a universal solution
- State-of-the-art BlueField Smart NICs bring more compute power into the network
- Can we exploit additional compute capabilities of modern BlueField Smart NICs into existing HPC middleware to extract
  - Peak pure communication performance
  - Overlap of communication and computation

  For dense non-blocking collective communications?

OVERVIEW OF BLUEFIELD SMART NIC/ DATA PROCESSING UNIT (DPU)

- 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
- Embedded CPU Function Ownership mode
  - Packet processing

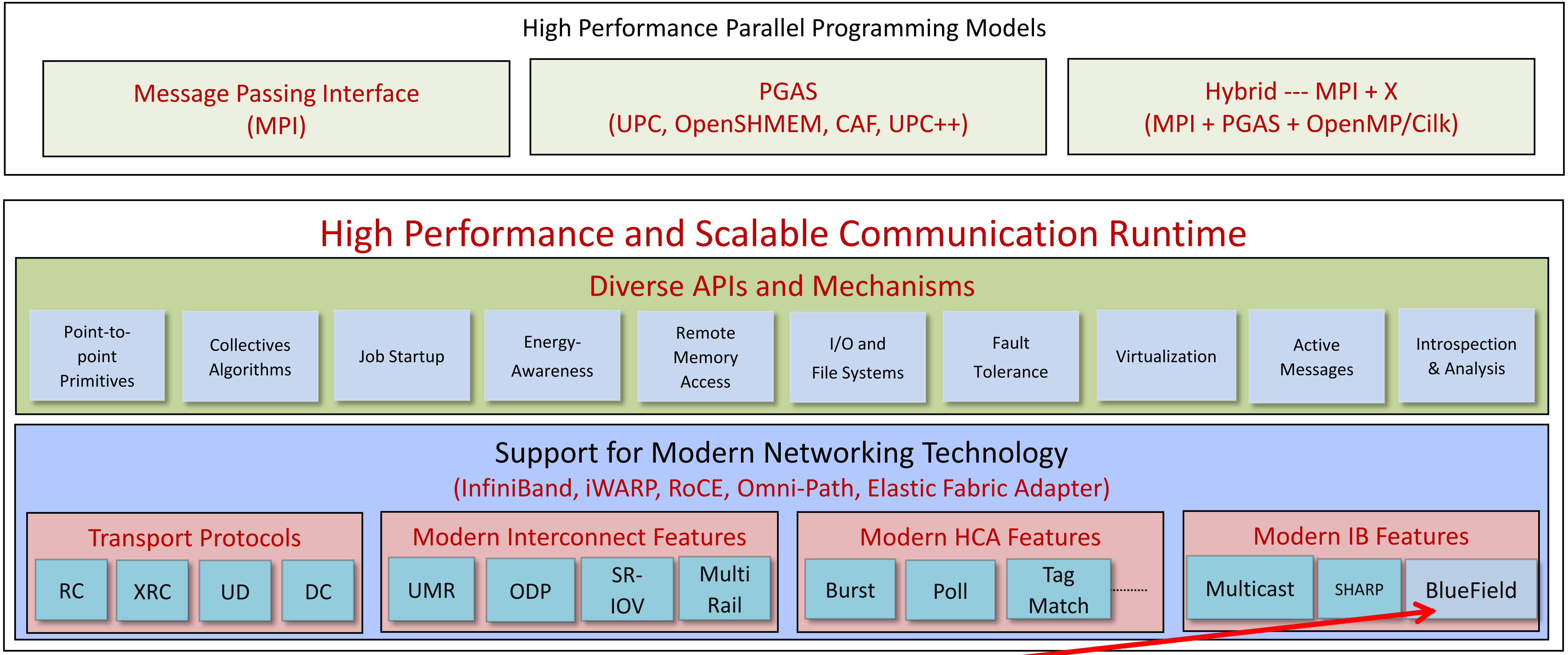


OVERVIEW OF THE MVAPICH2 PROJECT

- **High Performance open-source MPI Library**
- **Support for multiple interconnects**
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), and AWS EFA
- **Support for multiple platforms**
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD)
- **Started in 2001, first open-source version demonstrated at SC '02**
- **Supports the latest MPI-3.1 standard**
- **http://mvapich.cse.ohio-state.edu**
- **Additional optimized versions for different systems/environments:**
  - MVAPICH2-X (Advanced MPI + PGAS), since 2011
  - MVAPICH2-GDR with support for NVIDIA GPGPUs, since 2014
  - MVAPICH2-MIC with support for Intel Xeon-Phi, since 2014
  - MVAPICH2-Virt with virtualization support, since 2015
  - MVAPICH2-EA with support for Energy-Awareness, since 2015
  - MVAPICH2-Azure for Azure HPC IB instances, since 2019
  - MVAPICH2-X-AWS for AWS HPC+EFA instances, since 2019
- **Tools:**
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

- Used by more than 3,150 organizations in 89 countries
- More than 1.26 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '20 ranking)
  - 4^{th} , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 9^{th}, 448, 448 cores (Frontera) at TACC
  - 14^{th}, 391,680 cores (ABCI) in Japan
  - 21^{st}, 570,020 cores (Nurion) in South Korea and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 9^{th} ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

THE MVAPICH APPROACH



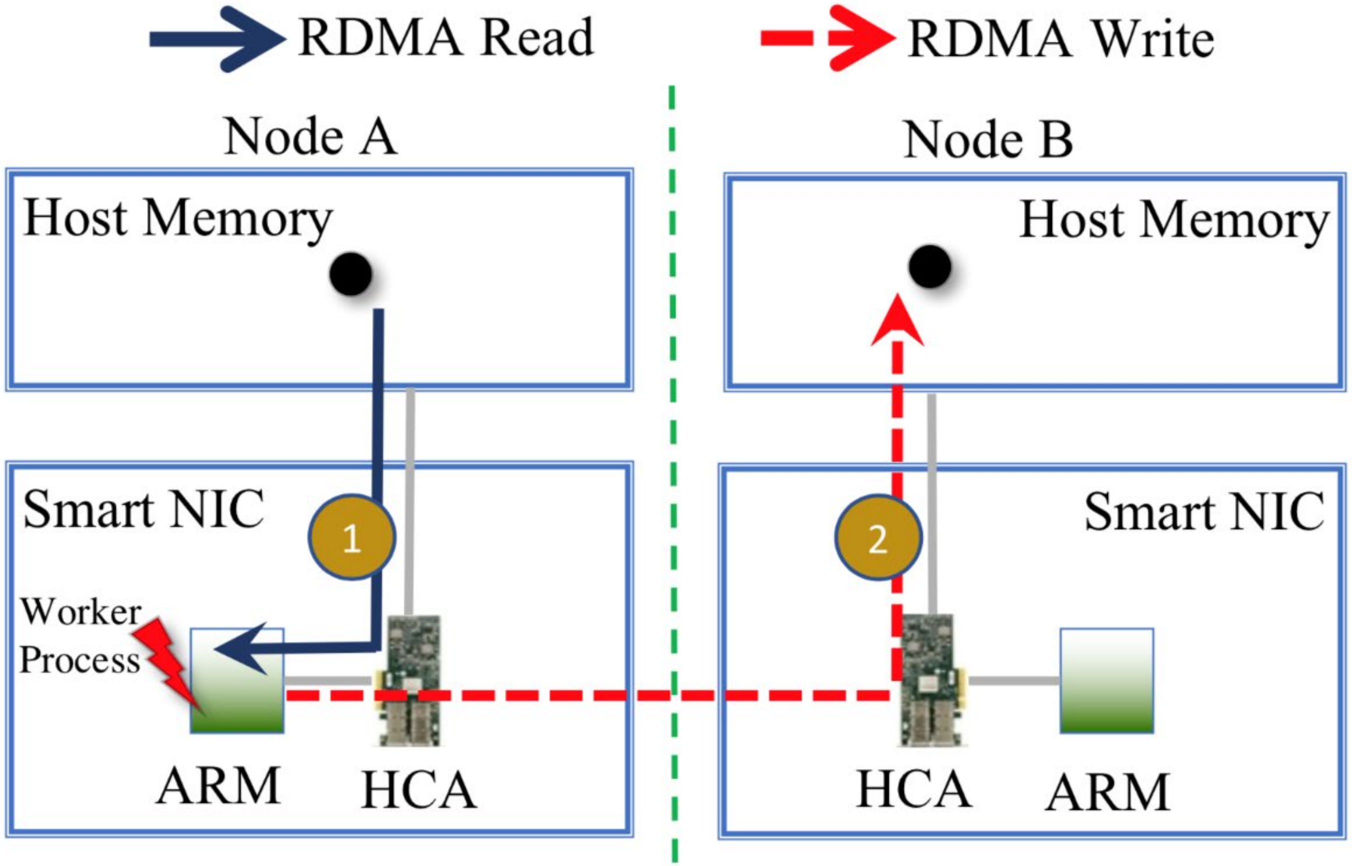
* Upcoming

PROPOSED OFFLOAD FRAMEWORK

- Non-blocking collective operations are offloaded to a set of Worker processes
- BlueField is set to separated host mode
- Worker processes are spawned to the ARM cores of BlueField
- Once the application calls a collective, host processes prepare a set of metadata and provide it to the Worker processes
- Using these metadata, worker processes can access host memory through RDMA
- Worker processes progress the collective on behalf of the host processes
- Once message exchanges are completed, worker processes notify the host processes about the completion of the non-blocking operation

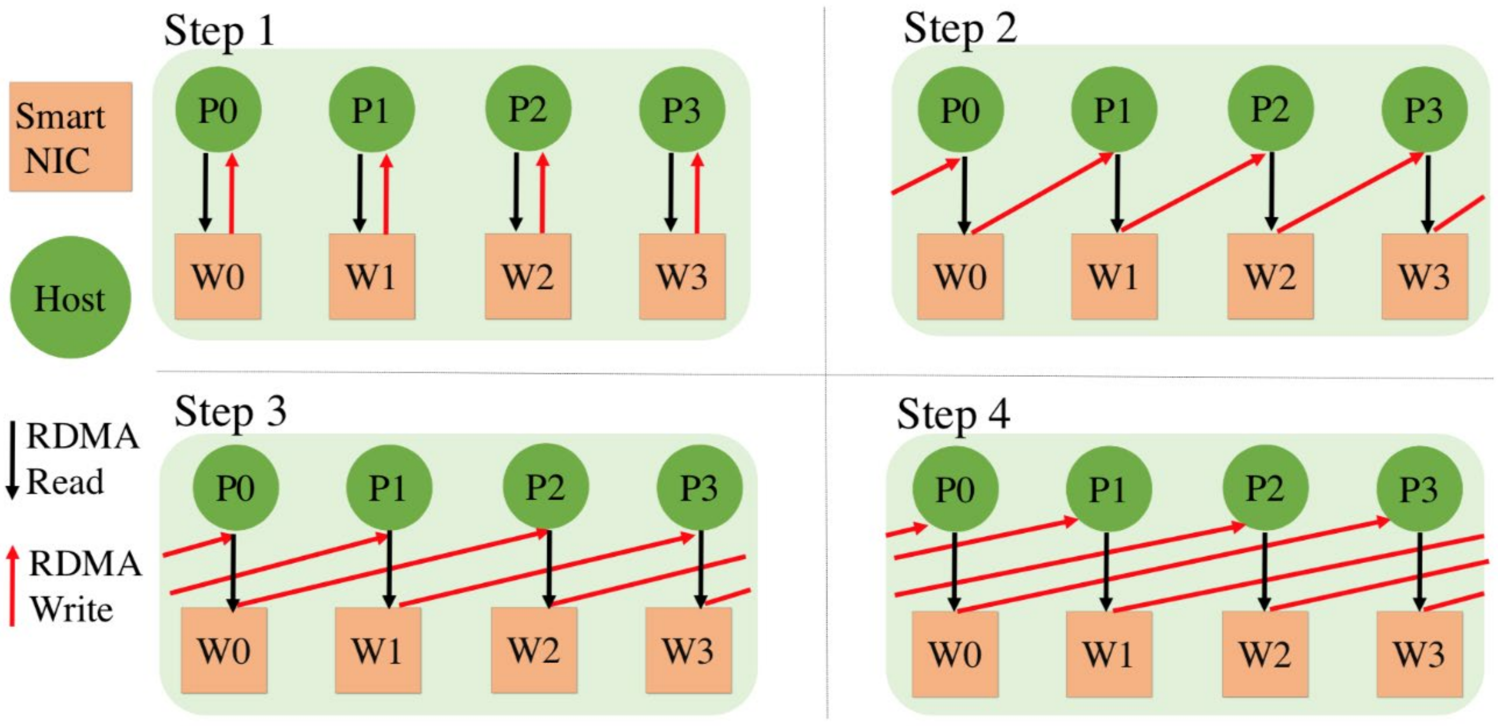
PROPOSED NON-BLOCKING ALLTOALL DESIGN

- Worker process performs RDMA Read to receive the data chunk from host main memory
- Once data is available in the ARM memory, worker process performs RDMA Write to the remote host memory



PROPOSED NON-BLOCKING ALLTOALL DESIGNS (CONT'D)

- Example: Scatter Destination Algorithm
- Focus is on medium and large messages
- Message chunking and pipelining is utilized to reduce the overheads of staging



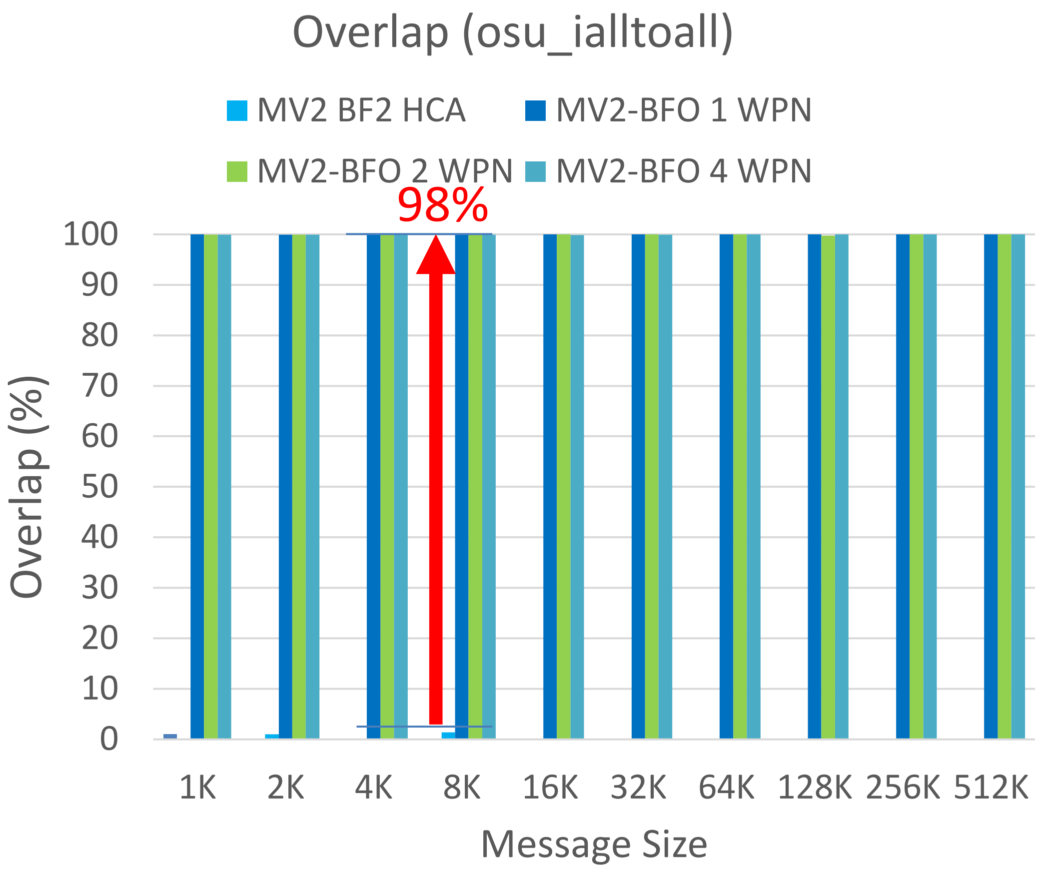
EXPERIMENTAL SETUP

- HPC Advisory Council High-Performance Computing Center
  - Cluster has 32 compute-node with Broadwell series of Xeon dual-socket, 16-core processors operating at 2.60 GHz with 128 GB RAM
  - Mellanox BlueField-2 HDR100 ConnectX-6 HCAs (100Gbps data rate) with OFED version 5.2-1.0.4
  - BlueField-2 adapters are equipped with 8 ARM cores operating at 1999 MHz with 16 GB RAM
- Based on the MVAPICH2 MPI library
- OSU Micro Benchmark for nonblocking Alltoall and P3DFFT Application
- PPN: Number of host processes per node
- WPN: Number of worker processes per smart NIC

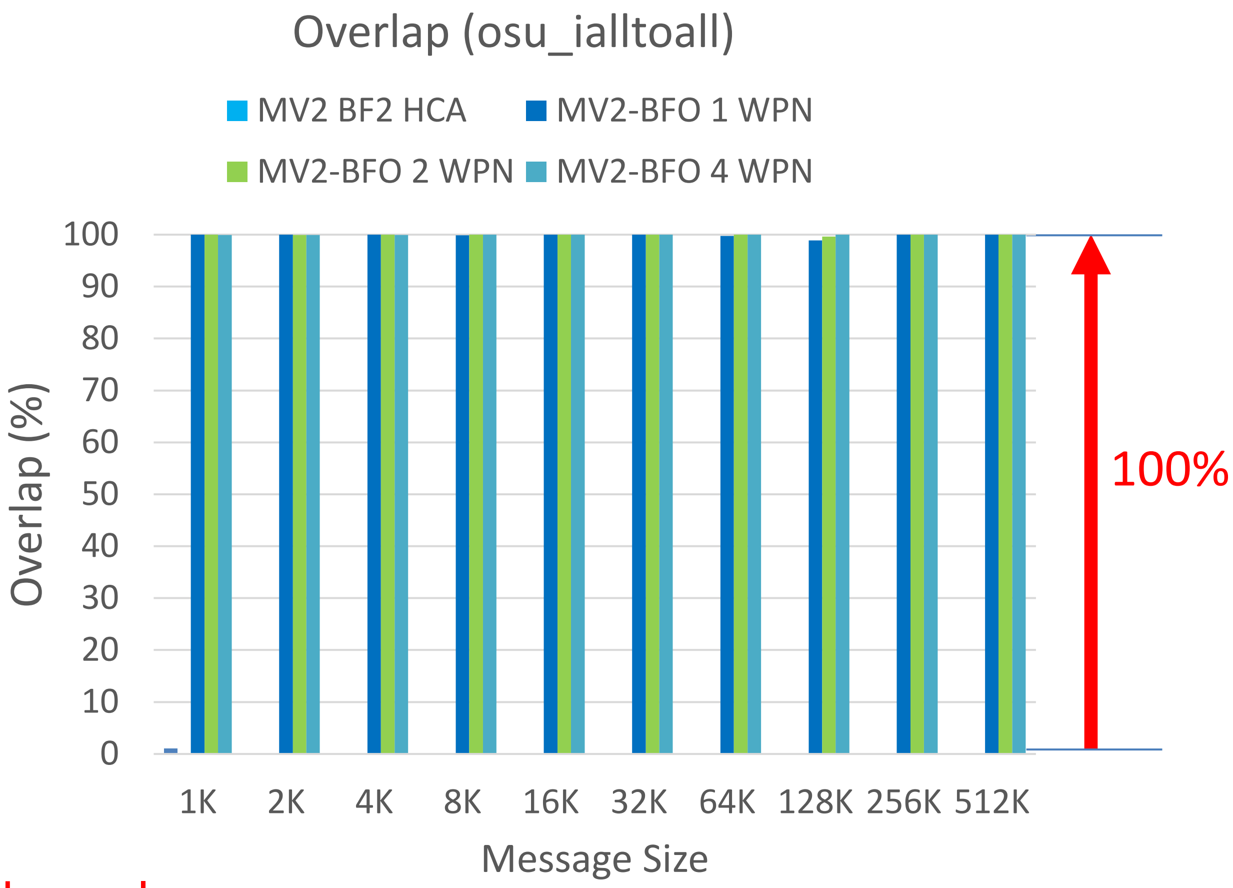
OSU MICRO BENCHMARK IALLTOALL

- **`osu_ialltoall`** benchmark metrics
  - `Pure communication time`
    - Latency t is measured by calling MPI_Ialltoall followed by MPI_Wait
  - `Total execution time`
    - Total T = MPI_Ialltoall + synthetic compute + MPI_Wait
  - `Overlap`
    - Benchmark creates a synthetic computation block that takes t microsecond to finish. Before starting compute, MPI_Ialltoall is called and after that MPI_Wait. Overlap is calculated based on total execution time and compute time.
  - Part of the standard OSU Micro-Benchmark

OVERLAP OF COMMUNICATION AND COMPUTATION WITH OSU_IALLTOALL



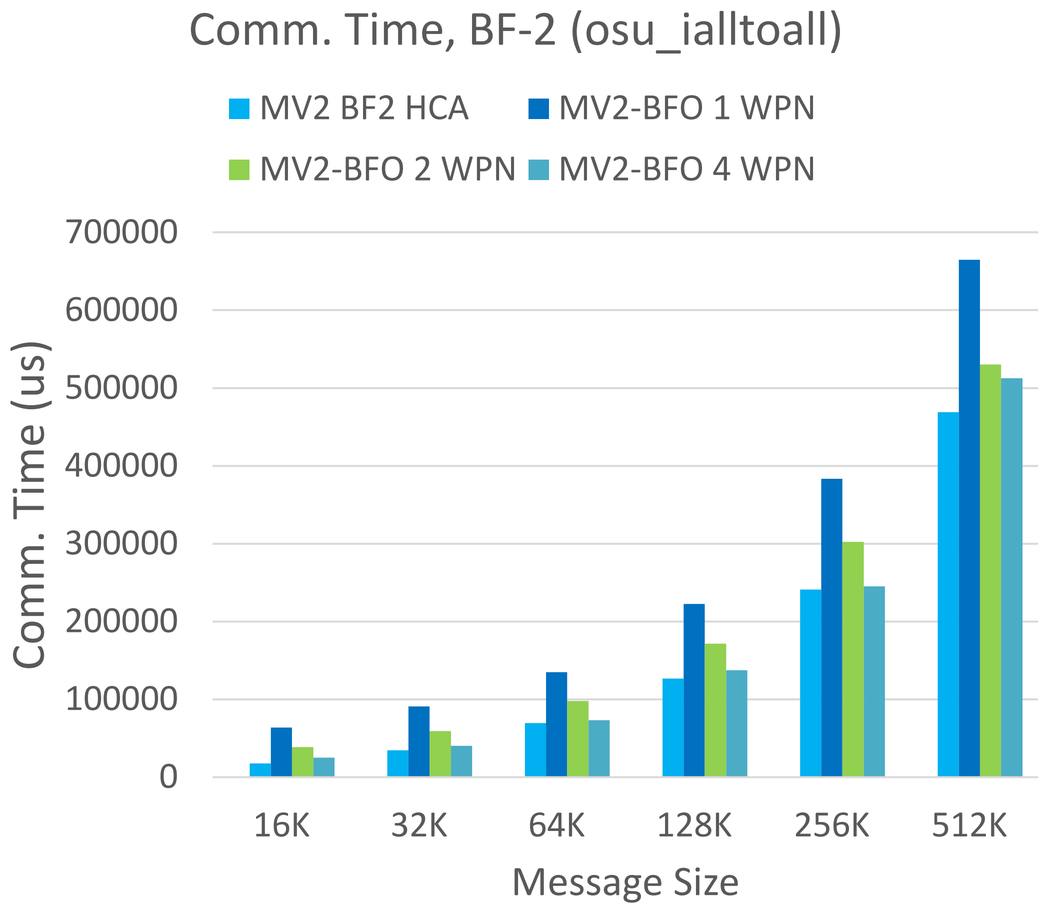
32 Nodes, 16 PPN



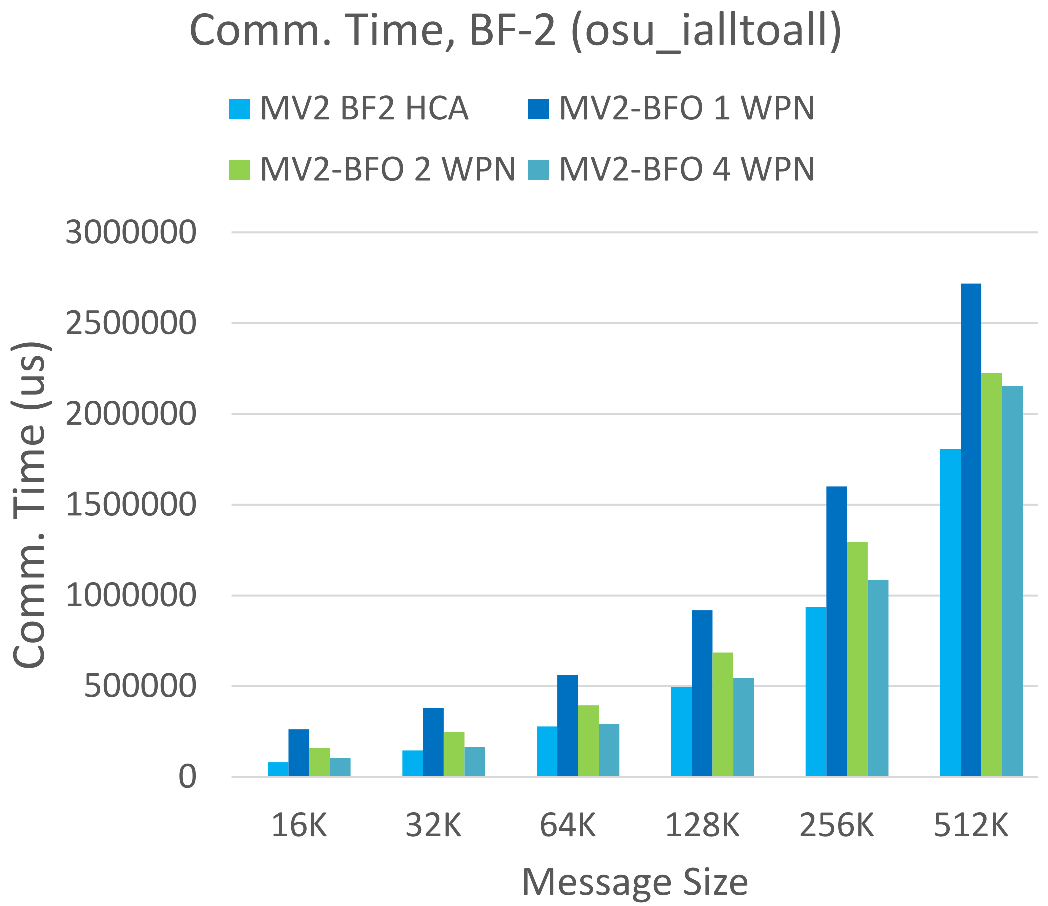
32 Nodes, 32 PPN

Delivers peak overlap

PURE COMMUNICATION LATENCY WITH OSU_IALLTOALL



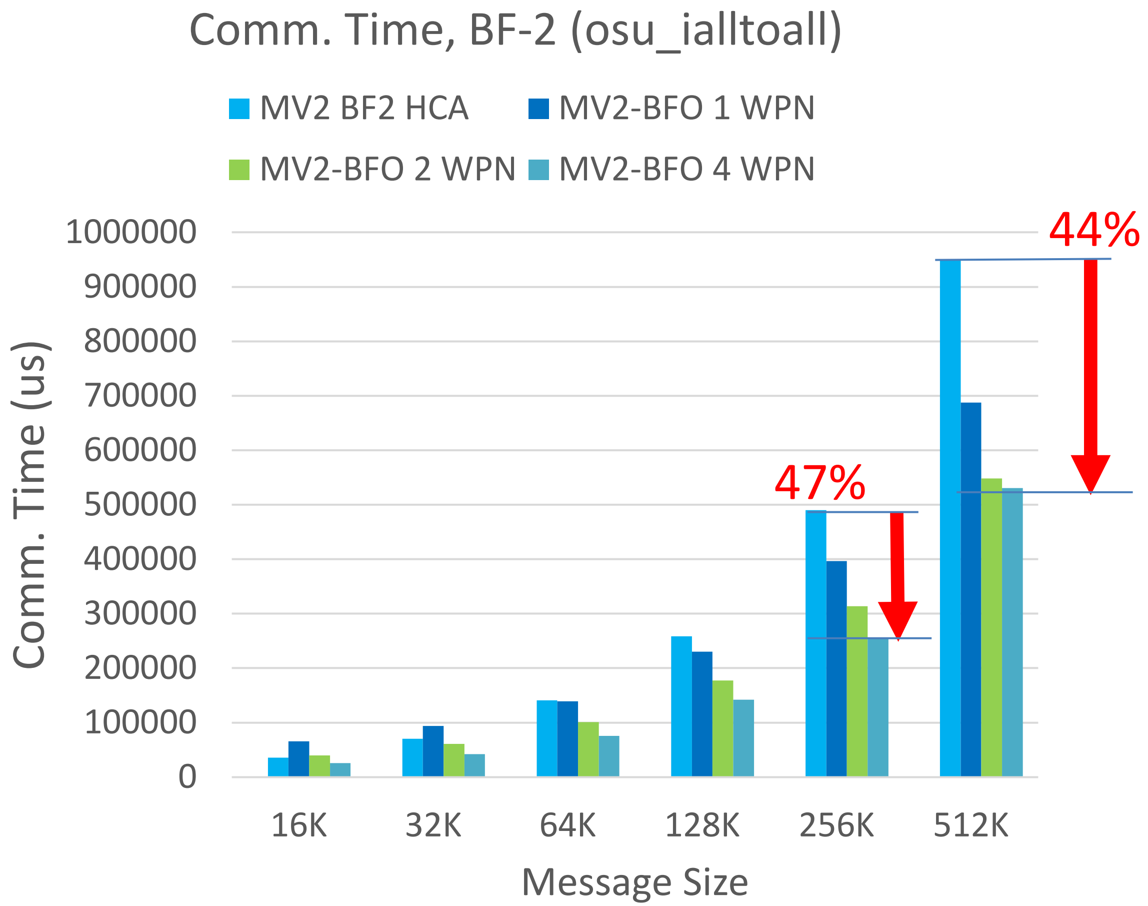
32 Nodes, 16 PPN



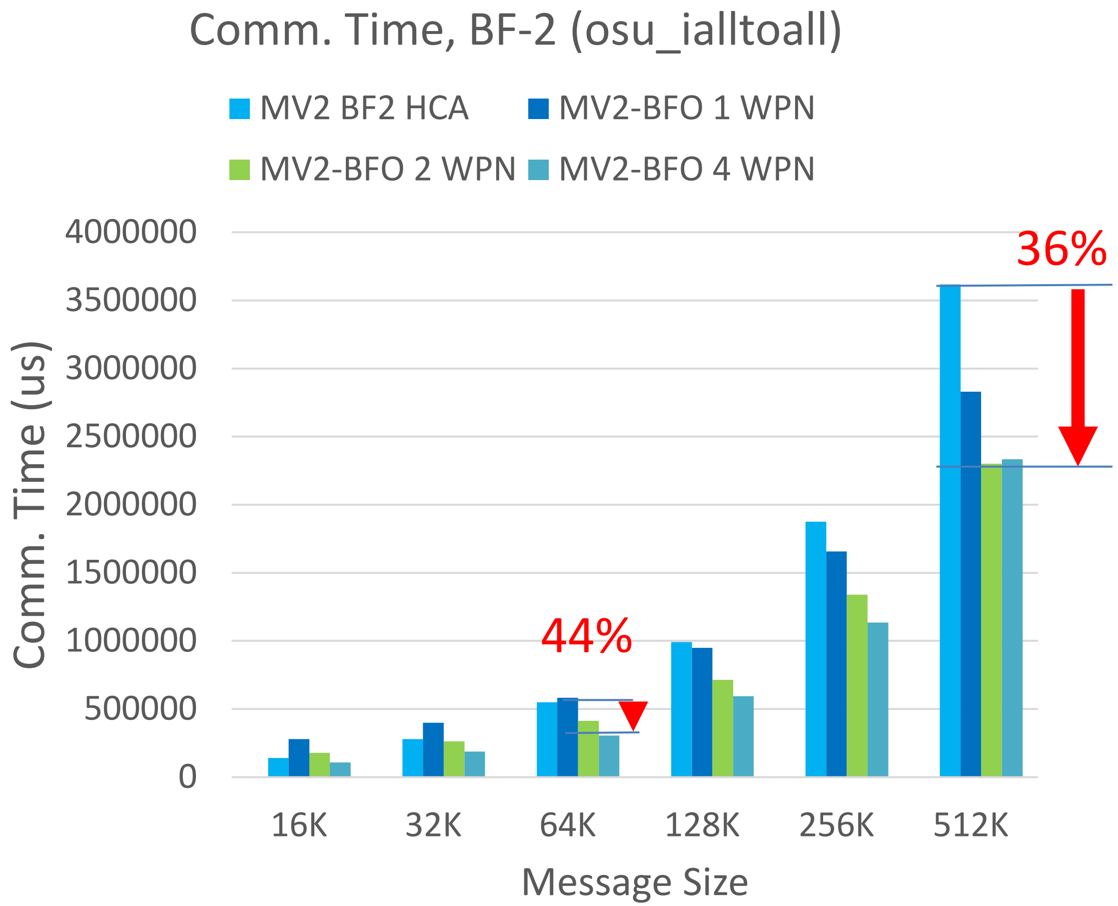
32 Nodes, 32 PPN

Delivers on-par host-based communication latency while exploiting overlap

TOTAL EXECUTION TIME WITH OSU_IALLTOALL



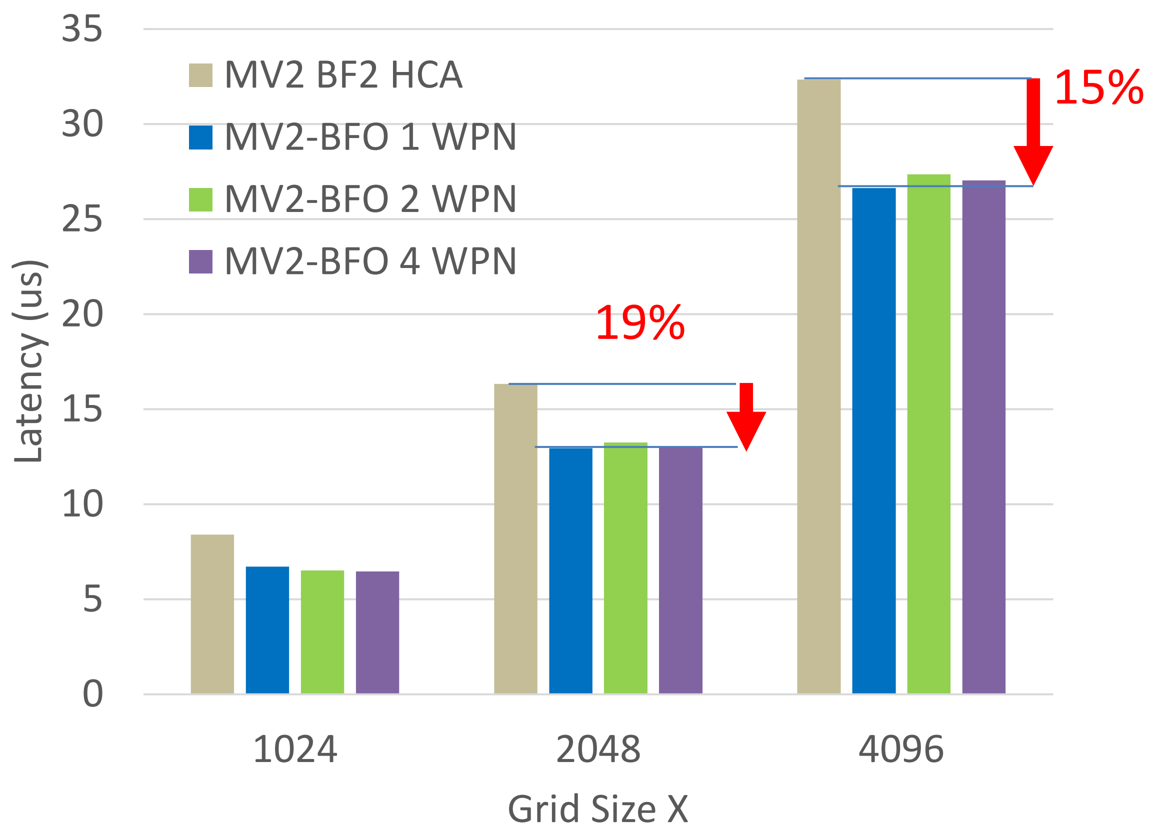
32 Nodes, 16 PPN



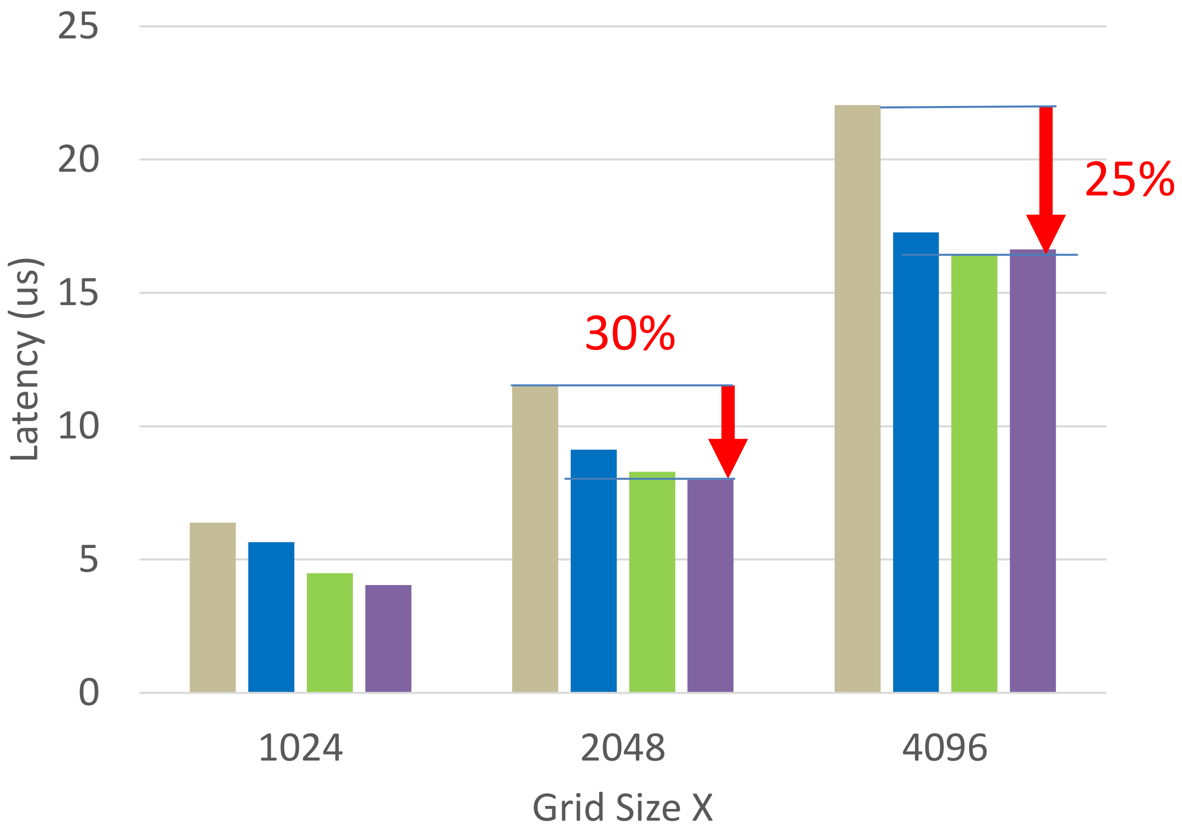
32 Nodes, 32 PPN

Total execution time of osu_ialltoall

P3DFFT APPLICATION EXECUTION TIME



32 Nodes, 16 PPN



32 Nodes, 32 PPN

BlueField-2 with MVAPICH2-BFO can deliver application-level performance benefits

PUBLICATION

All of the proposed designs and more thorough evaluations will be published in ISC High Performance 2021 conference proceedings

BluesMPI: Efficient MPI Non-blocking Alltoall offloading Designs on Modern BlueField Smart NICs

Mohammadreza Bayatpour, Nick Sarkauskas, Hari Subramoni, Jahanzeb Maqbool Hashmi, and Dhabaleswar K. (DK) Panda

CONCLUSION AND FUTURE PLANS

- Propose efficient designs for the MVAPICH2 MPI library that utilize the BlueField Smart NICs compute capability to progress MPI non-blocking collective operations
- Analyze the proposed designs from multiple aspects using benchmarks and HPC kernel applications to take advantage of the state-of-the-art features of modern BlueField Smart NICs
- Our proposed designs provide close to 100% overlap of communication and computation while having on-par pure communication latency for non-blocking Alltoall
- The design is to reduce the total execution time of P3DFFT application up to 30% on 1,024 processes
- Working on offloading designs for other non-blocking collective operations
- Will be available in the future MVAPICH2 releases

THANK YOU!

bayapour.1@osu.edu,sarkauskas.1@osu.edu,subramon@cse.ohio-state.edu, panda@cse.ohio-state.edu

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