Enhancing MPI Communication using Hardware Tag Matching: The MVAPICH Approach

Talk at UCX BoF (SC '19)

by

Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: panda@cse.ohio-state.edu

http://www.cse.ohio-state.edu/~panda

Introduction, Motivation, and Challenge

- HPC applications require high-performance, low overhead data paths that provide
  - Low latency
  - High bandwidth
  - High message rate
  - Good overlap of computation with communication
- Hardware Offloaded Tag Matching
- Can we exploit tag matching support in UCX into existing HPC middleware to extract peak performance and overlap?

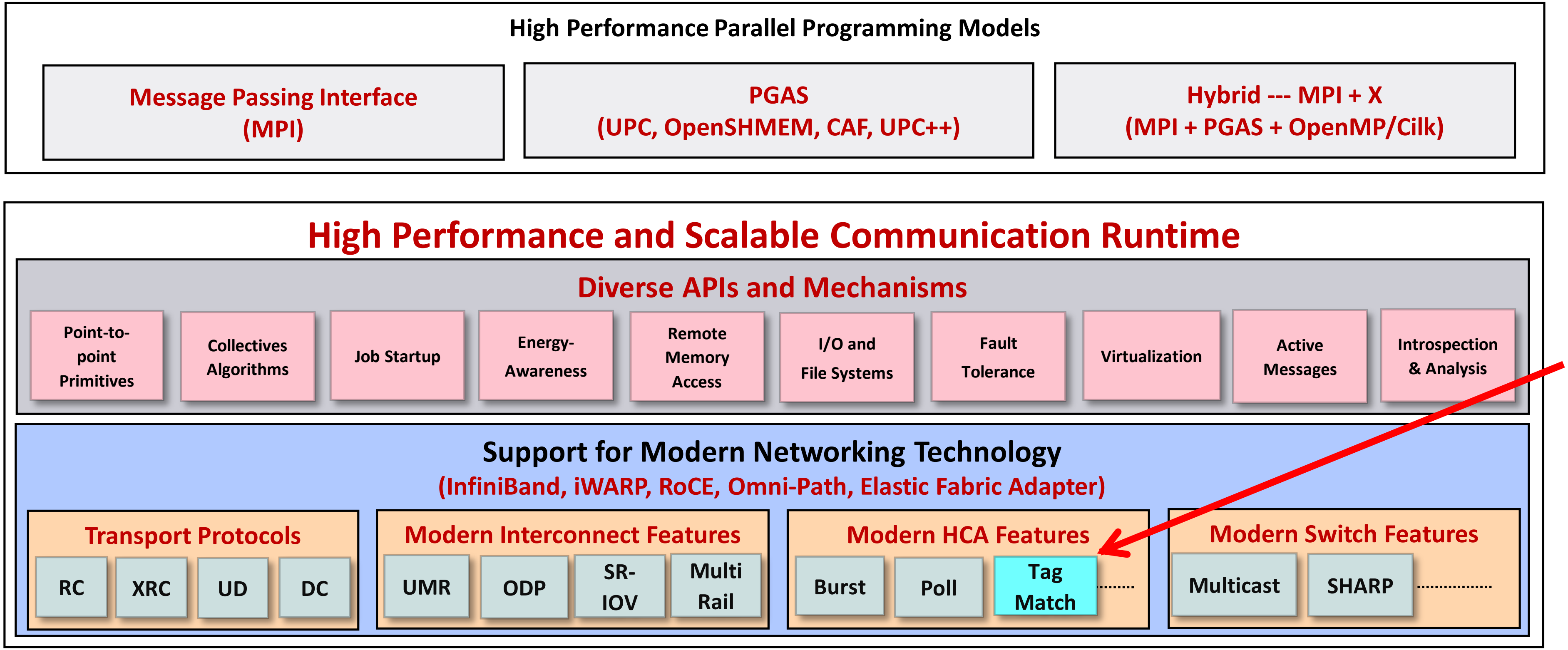
Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.1), Started in 2001, First version available in 2002 (SC '02)
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - **Used by more than 3,050 organizations in 89 countries**
  - **More than 615,000 (> 0.6 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '19 ranking)
    - 3rd, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center
    - 5th, 448, 448 cores (Frontera) at TACC
    - 8th, 391,680 cores (ABCI) in Japan
    - 14th, 570,020 cores (Neurion) in South Korea and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  - **http://mvapich.cse.ohio-state.edu**

Partner in the #5th TACC Frontera System

- Empowering Top500 systems for over a decade

The MVAPICH Approach



High Performance Parallel Programming Models

Message Passing Interface (MPI)

PGAS (UPC, OpenSHMEM, CAF, UPC++)

Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)

High Performance and Scalable Communication Runtime

Diverse APIs and Mechanisms

Point-to-point Primitives

Collectives Algorithms

Job Startup

Energy-Awareness

Remote Memory Access

I/O and File Systems

Fault Tolerance

Virtualization

Active Messages

Introspection & Analysis

Support for Modern Networking Technology

(InfiniBand, iWARP, RoCE, Omni-Path, Elastic Fabric Adapter)

Transport Protocols

RC

XRC

UD

DC

Modern Interconnect Features

UMR

ODP

SR-IOV

Multi Rail

Modern HCA Features

Burst

Poll

Tag Match

....

Modern Switch Features

Multicast

SHARP

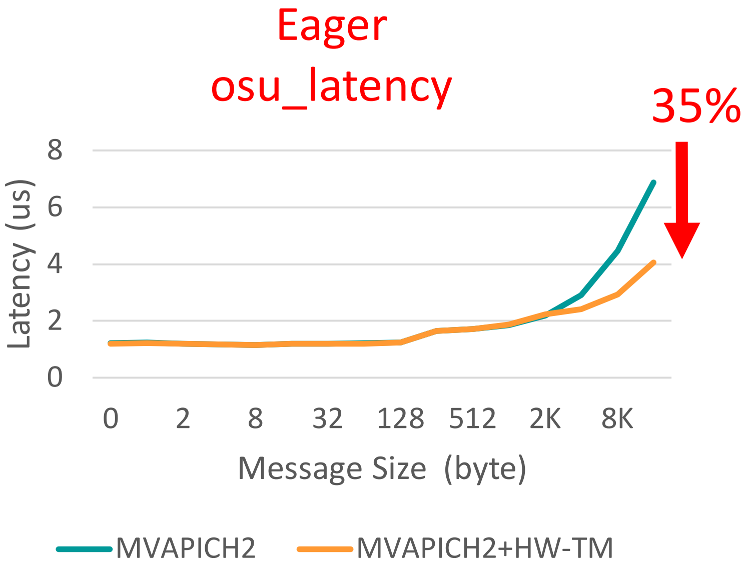
....

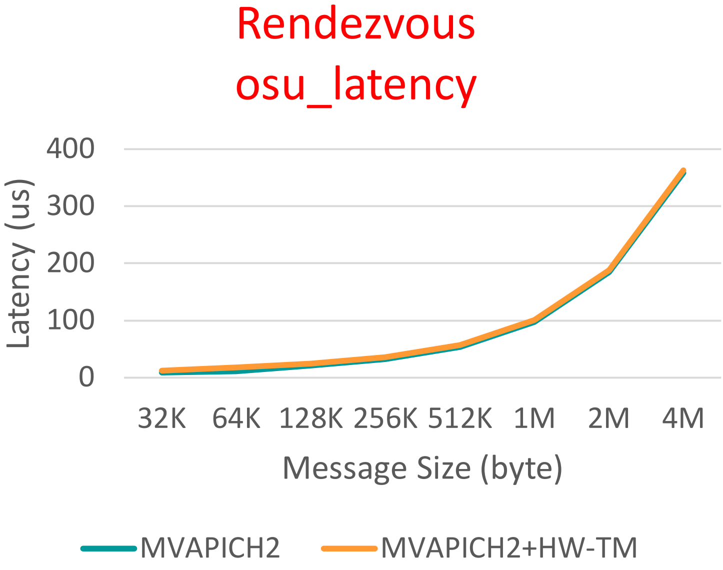
* Upcoming

Hardware Tag Matching Support

- Offloads the processing of point-to-point MPI messages from the host processor to HCA
- Enables zero copy of MPI message transfers
  - Messages are written directly to the user's buffer without extra buffering and copies
- Provides rendezvous progress offload to HCA
  - Increases the overlap of communication and computation

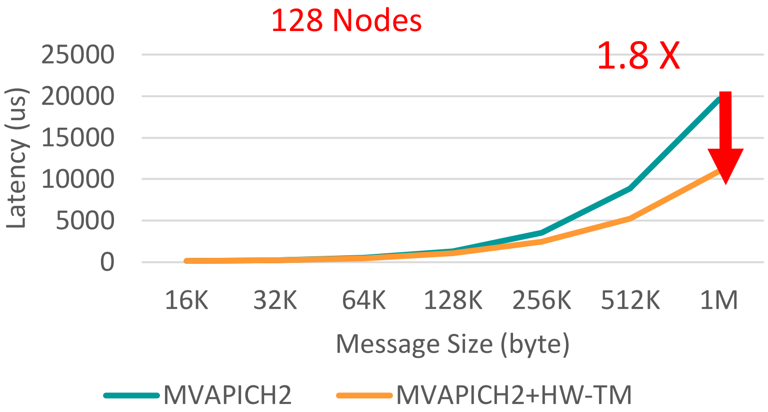
Impact of Zero Copy MPI Message Passing using HW Tag Matching (Point-to-point)

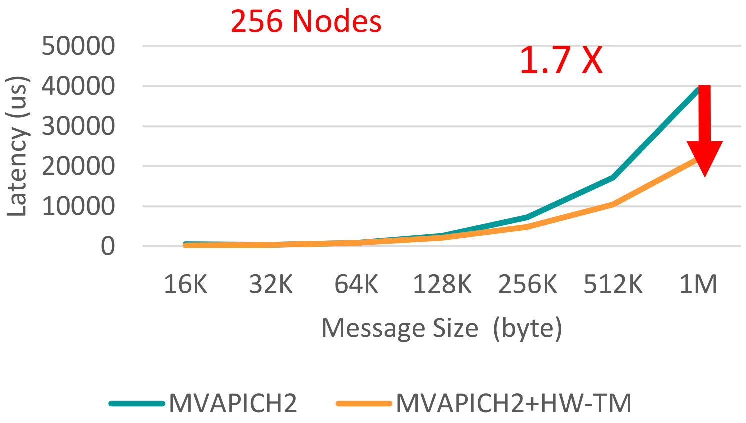


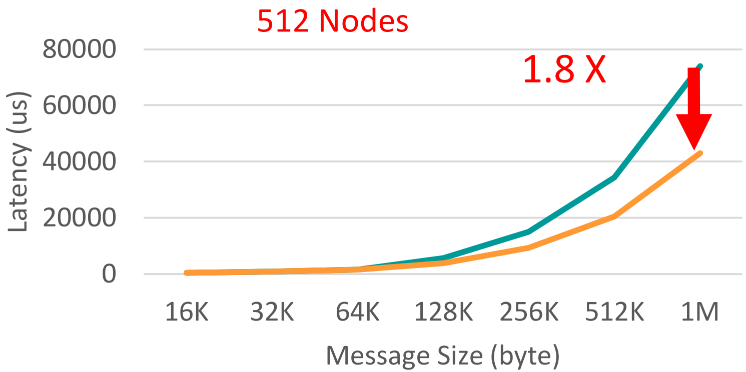


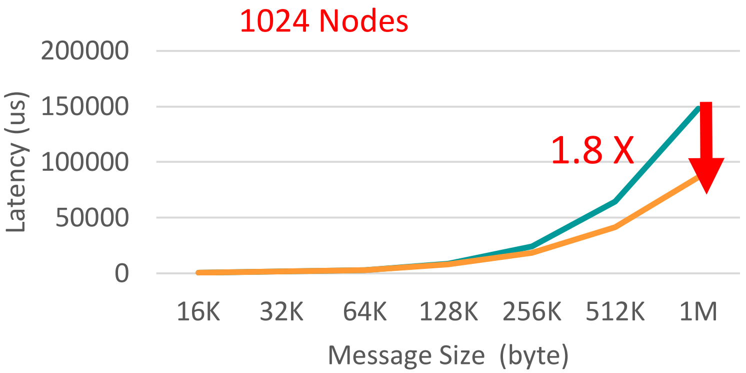
Removal of intermediate buffering/copies can lead up to 35% performance improvement in latency of medium messages on TACC Frontera

Performance of MPI_Iscatterv using HW Tag Matching on Frontera


128 Nodes

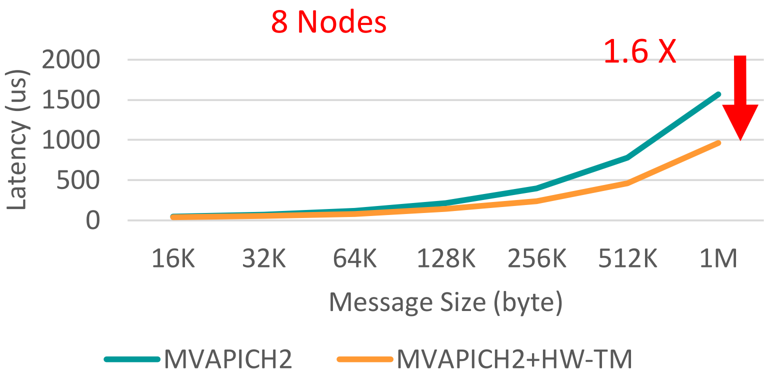

256 Nodes

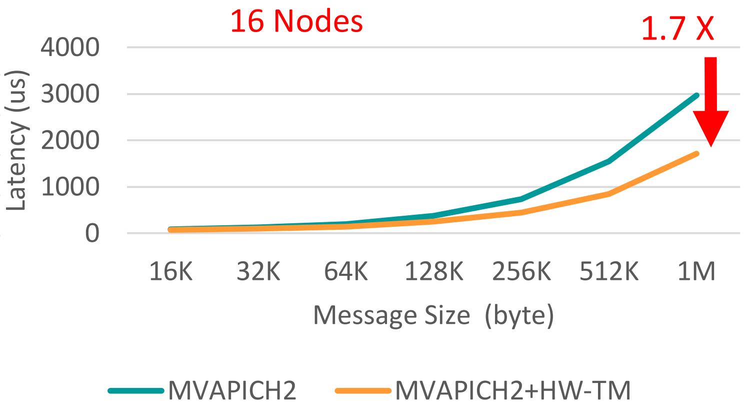

512 Nodes

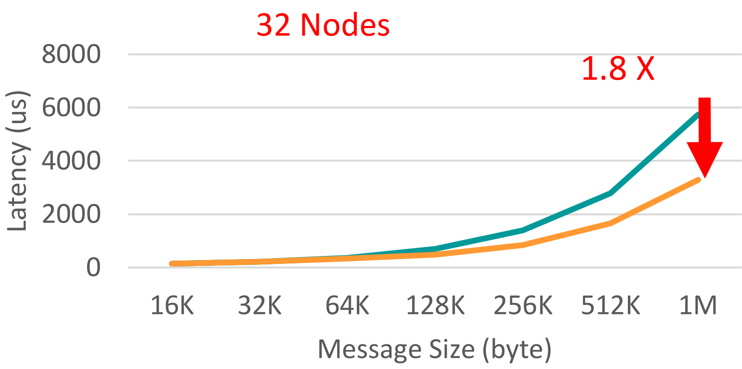

1024 Nodes

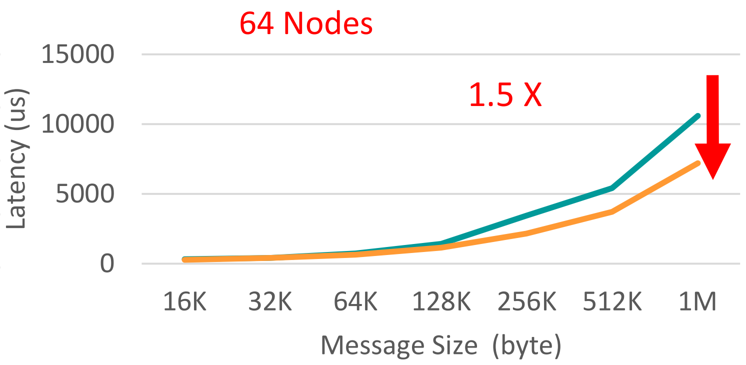
- Up to 1.8x Performance Improvement
- Sustained benefits as system size increases

Performance of MPI_Ialltoall using HW Tag Matching on Frontera



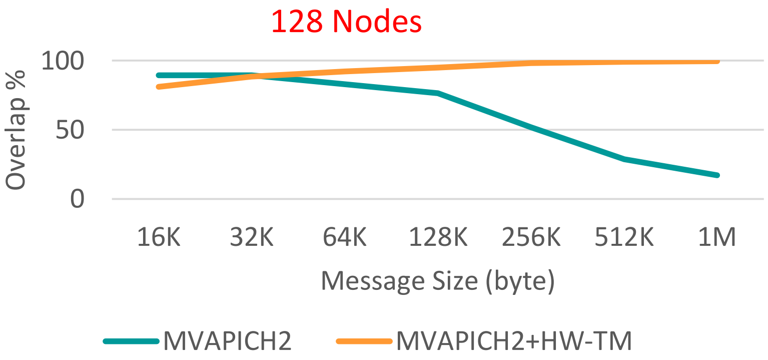


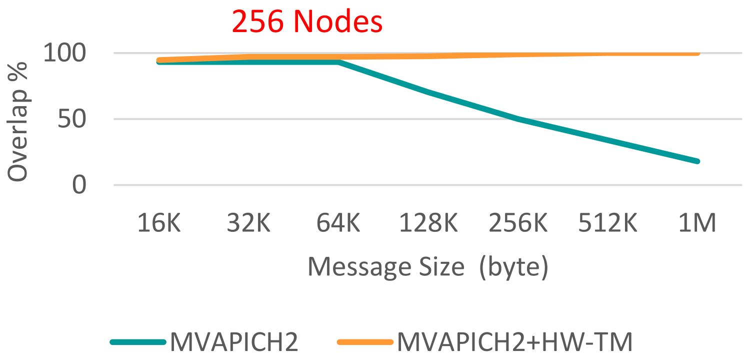


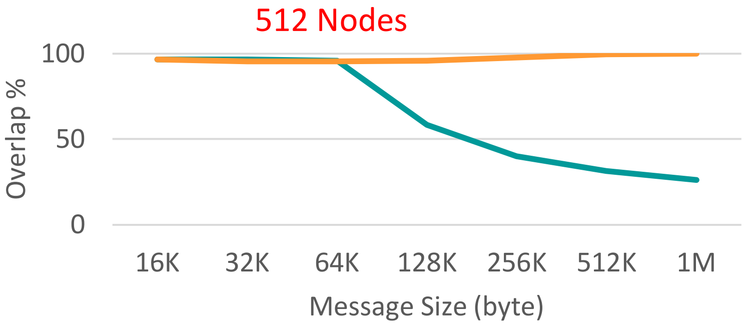


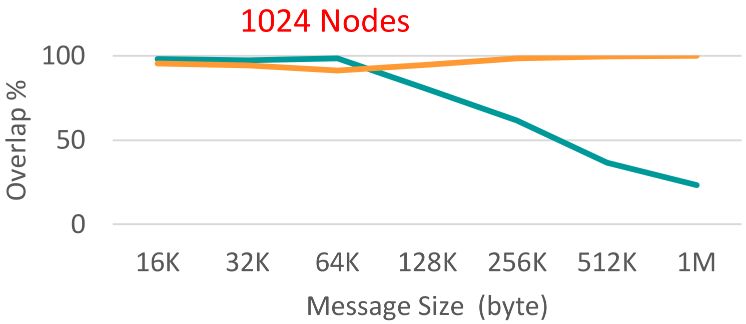
- Up to 1.8x Performance Improvement
- Sustained benefits as system size increases

Overlap with MPI_Iscatterv using HW Tag Matching on Frontera



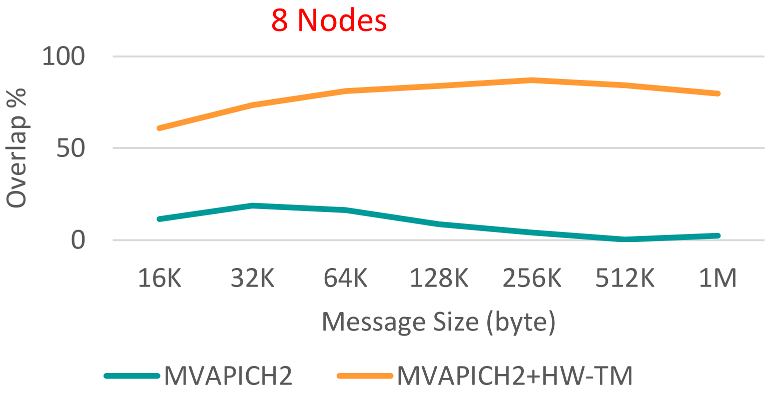


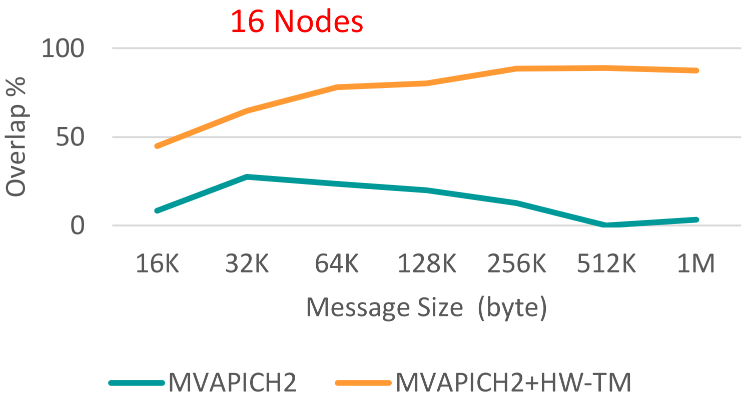


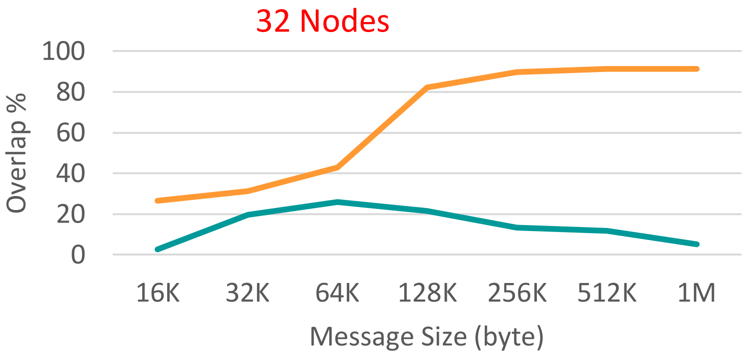


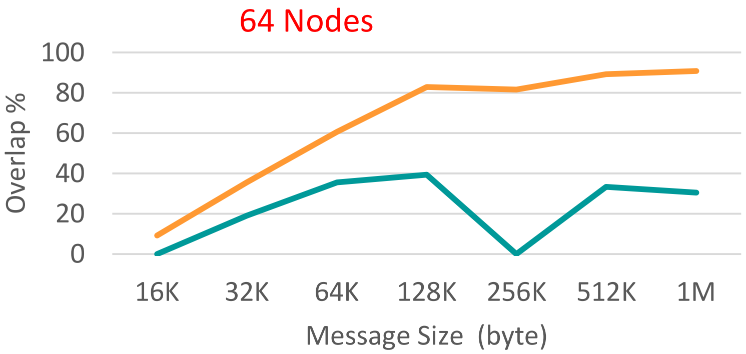
- Maximizing the overlap of communication and computation
- Sustained benefits as system size increases

Overlap with MPI_Ialltoall using HW Tag Matching on Frontera









- Maximizing the overlap of communication and computation
- Sustained benefits as system size increases

Future Plans

- Complete designs are being worked out
- Will be available in the future MVAPICH2 releases

Multiple Events at SC ‘19

- Presentations at OSU Booth (#2094)
  - Members of the MVAPICH, HiBD and HiDL members
  - External speakers
- Presentations at SC main program (Tutorials, Workshops, BoFs, Posters, and Doctoral Showcase)
- Presentation at many other booths (Mellanox, Intel, Microsoft, and AWS) and satellite events
- Complete details available at

  http://mvapich.cse.ohio-state.edu/conference/752/talks/