MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

Enhancing MPI Communication using Accelerated Verbs and Tag Matching: The MVAPICH Approach

Talk at UCX BoF (ISC '19)

by

Dhabaleswar K. (DK) Panda

The Ohio State University

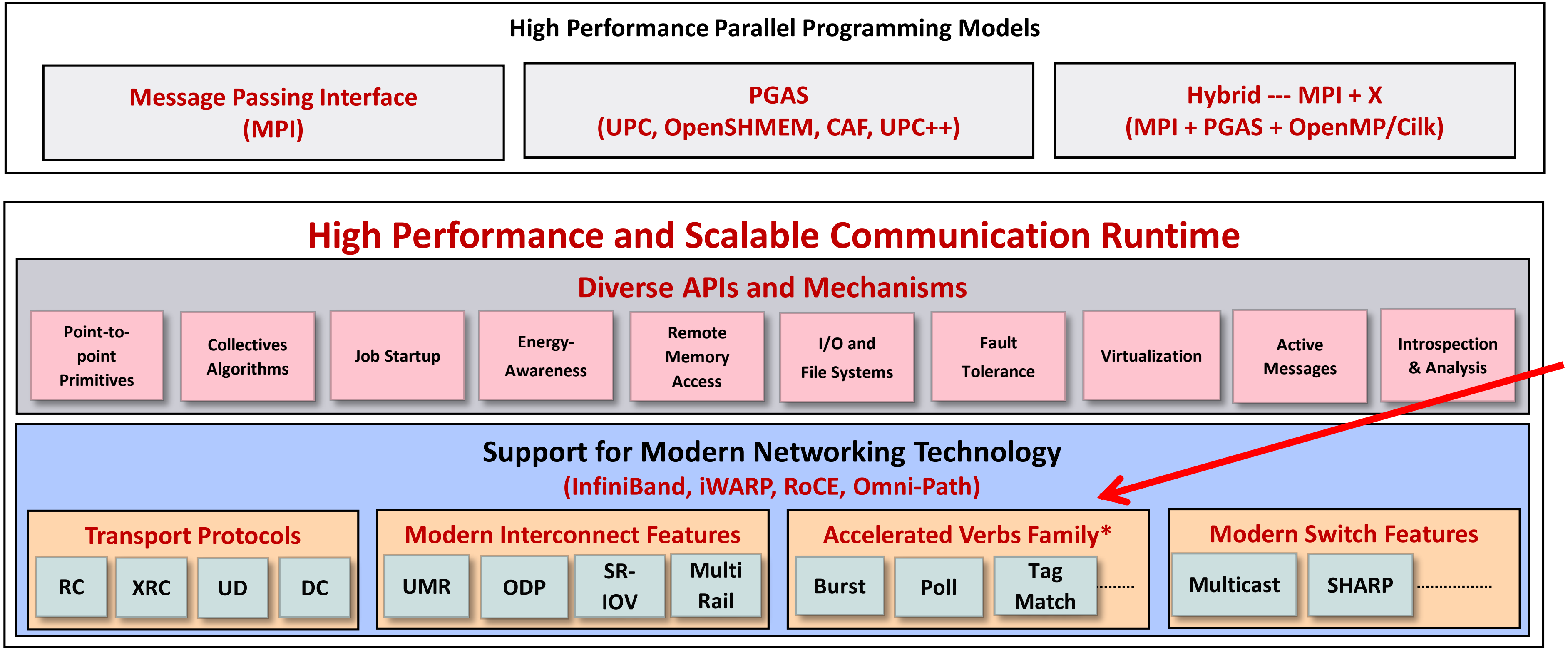
E-mail: panda@cse.ohio-state.edu

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
- Hardware Offloaded Tag Matching
- Different families of accelerated verbs available
  - Burst family
    - Accumulates packets to be sent into bursts of single SGE packets
  - Poll family
    - Optimizes send completion counts
    - Receive completions for which only the length is of interest
    - Completions that contain the payload in the CQE
- Can we integrate accelerated verbs and tag matching support in UCX into existing HPC middleware to extract peak performance and overlap?

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)

Transport Protocols: RC, XRC, UD, DC

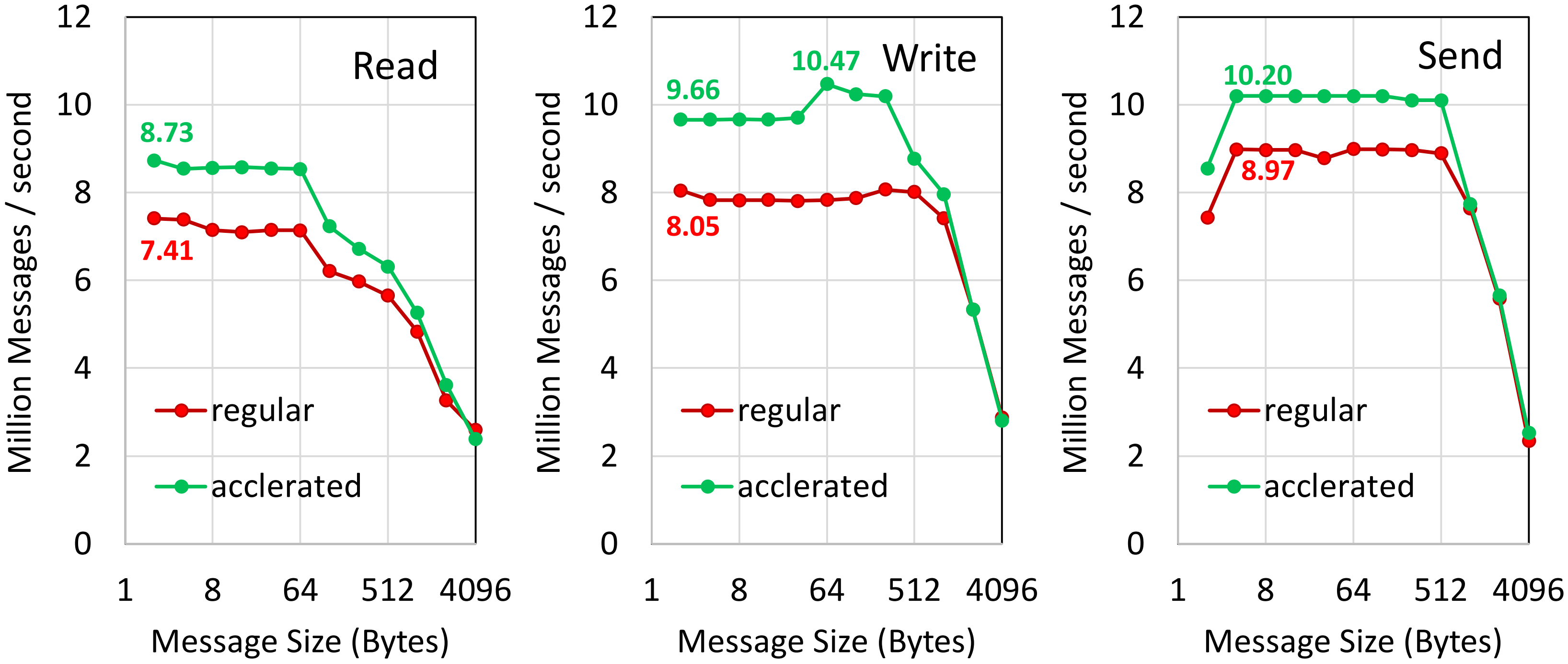
Modern Interconnect Features: UMR, ODP, SR-IOV, Multi Rail

Accelerated Verbs Family*: Burst, Poll, Tag Match ………

Modern Switch Features: Multicast, SHARP ………………

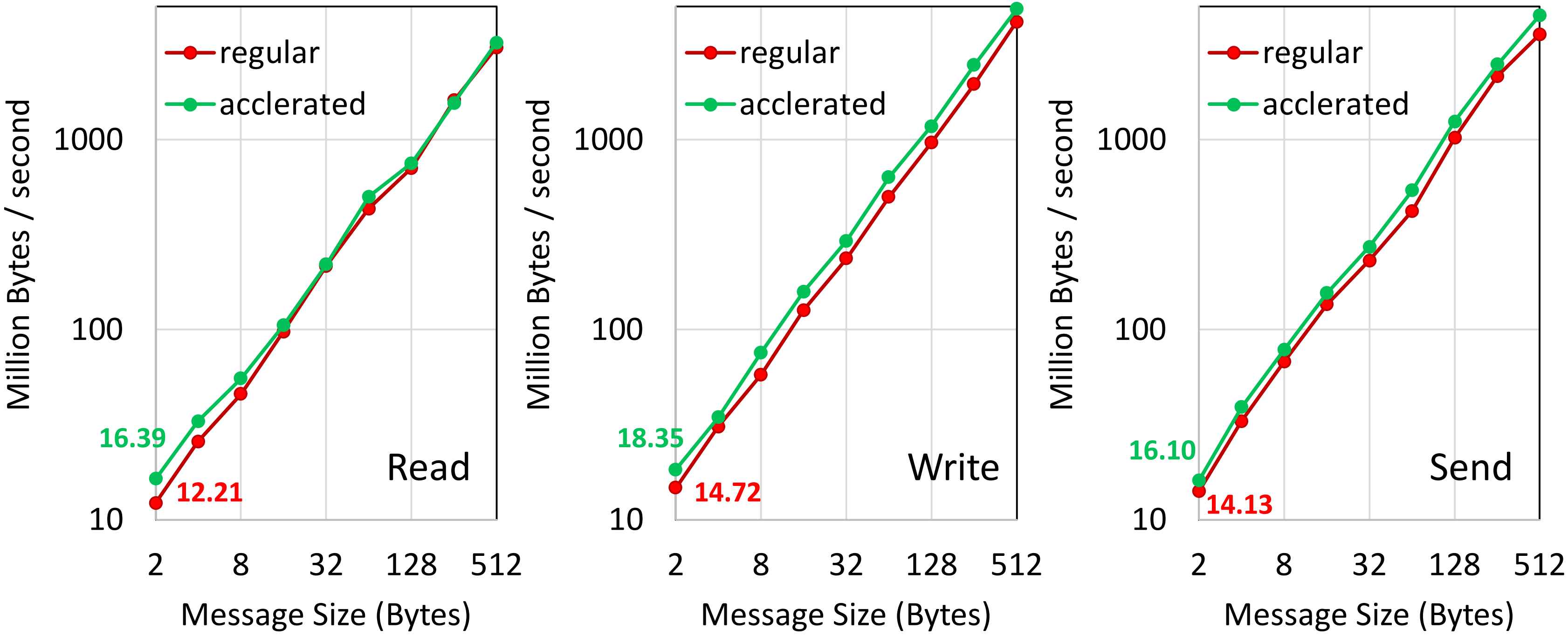
* Upcoming

Verbs-level Performance: Message Rate



ConnectX-5 EDR (100 Gbps), Intel Broadwell E5-2680 @ 2.4 GHz
MOFED 4.2-1, RHEL-7 3.10.0-693.17.1.el7.x86_64

Verbs-level Performance: Bandwidth

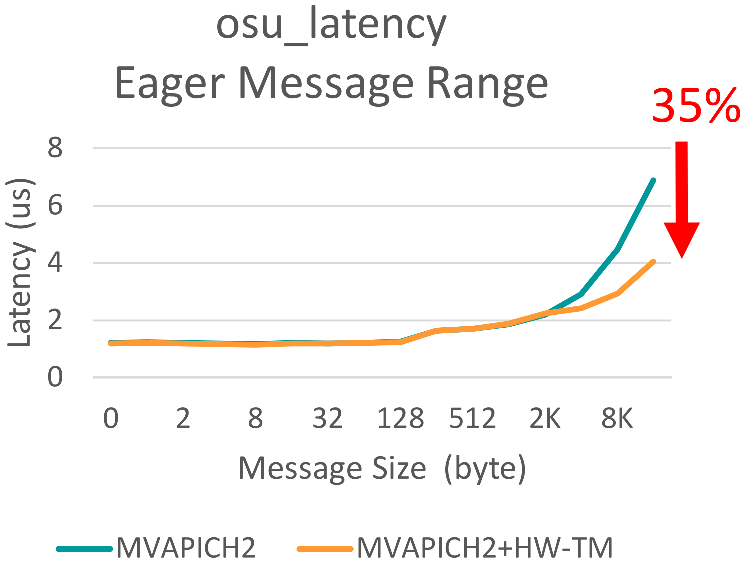


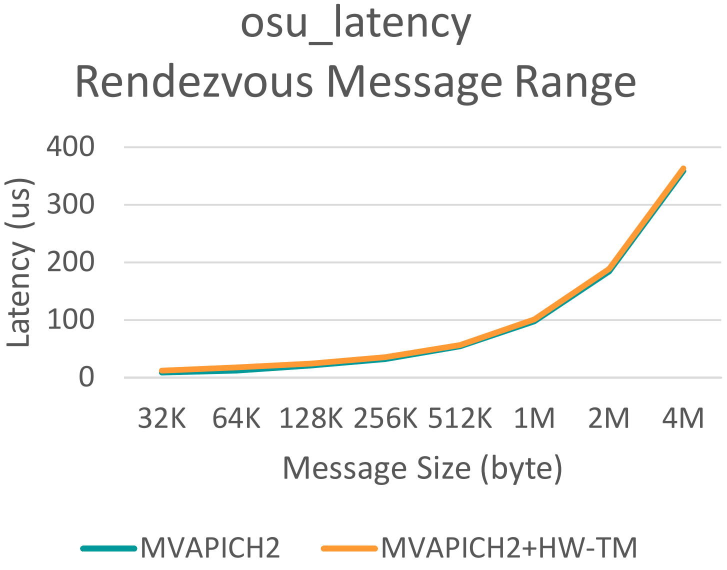
ConnectX-5 EDR (100 Gbps), Intel Broadwell E5-2680 @ 2.4 GHz
MOFED 4.2-1, RHEL-7 3.10.0-693.17.1.el7.x86_64

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

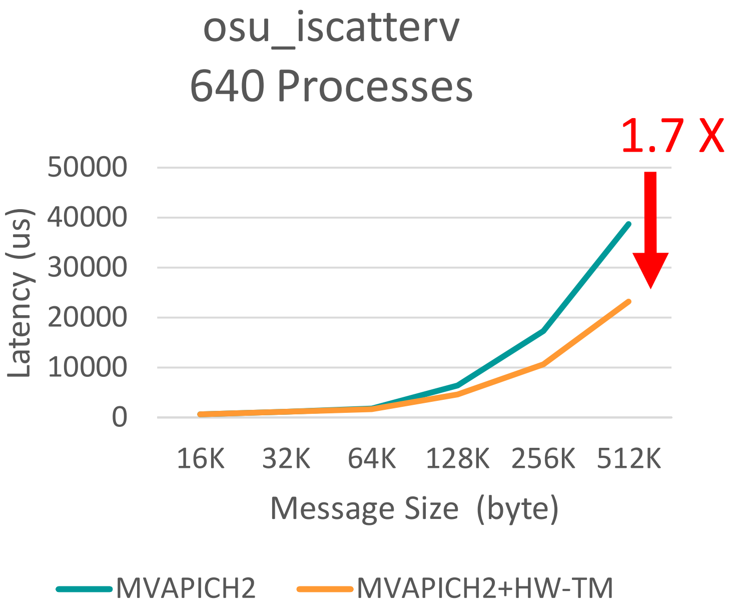
Impact of Zero Copy MPI Message Passing using HW Tag Matching

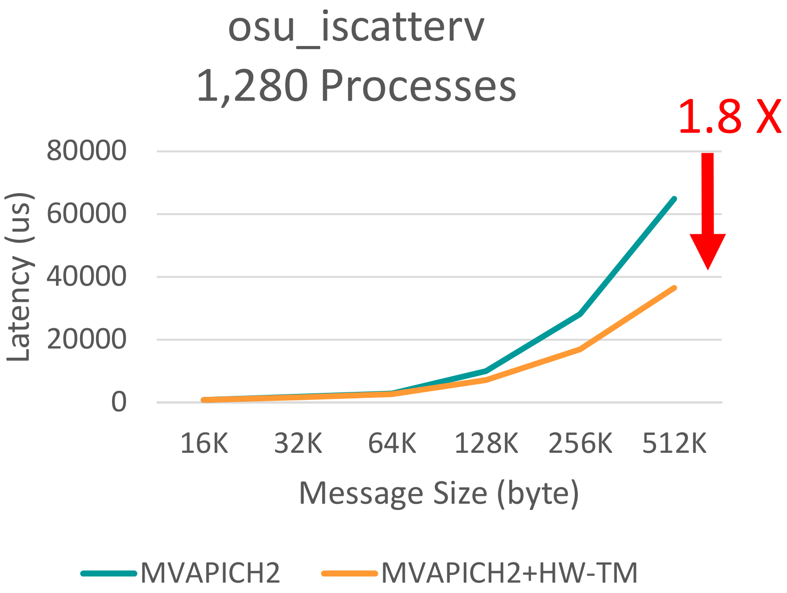




Removal of intermediate buffering/copies can lead up to 35% performance improvement in latency of medium messages

Impact of Rendezvous Offload using HW Tag Matching





The increased overlap can lead to 1.8X performance improvement in total latency of osu_iscatterv

Future Plans

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