2020 OFA Virtual Workshop

MVAPICH TOUCHES THE CLOUD: NEW FRONTIERS FOR MPI IN HIGH PERFORMANCE CLOUDS

Shulei Xu, Seyedeh Mahdieh Ghazimirsaeed, Hari Subramoni, Dhabaleswar K. (DK) Panda

The Ohio State University

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

AGENDA

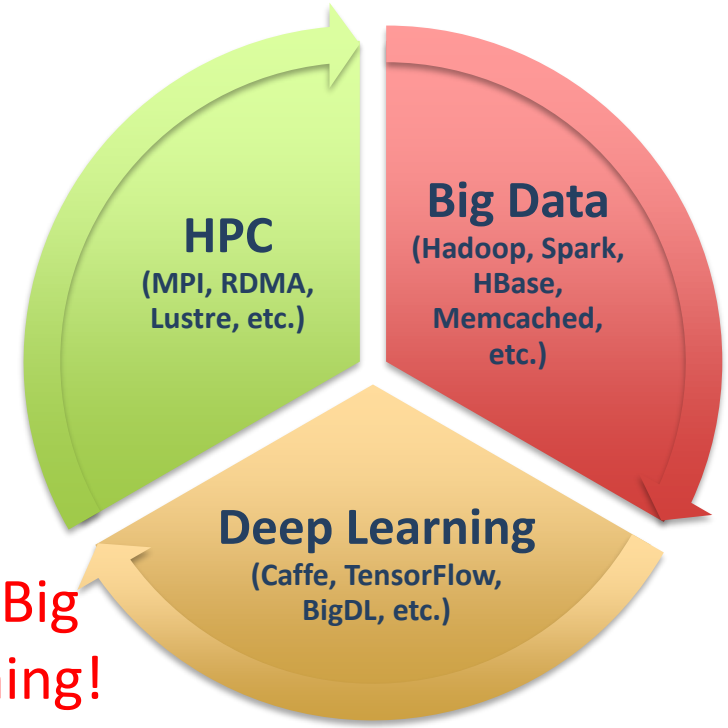
- Introduction & Motivation
- Support for AWS EFA
  - Overview
  - Challenges & Solutions for MPI Libraries Design on EFA
  - Experimental Evaluation
- Support for Azure
  - Performance Evaluation
  - Deployment
- Conclusions & Future Plans

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
- **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 3,090 organizations in 89 countries
- More than 765,000 (> 0.76 million) downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '19 ranking)
  - 3rd, 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 5th, 448, 448 cores (Frontera) at TACC
  - 8th, 391,680 cores (ABCI) in Japan
  - 14th, 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 5th ranked TACC Frontera system
- Empowering Top500 systems for more than 15 years

INCREASING USAGE OF HPC, BIG DATA AND DEEP LEARNING



Convergence of HPC, Big Data, and Deep Learning!

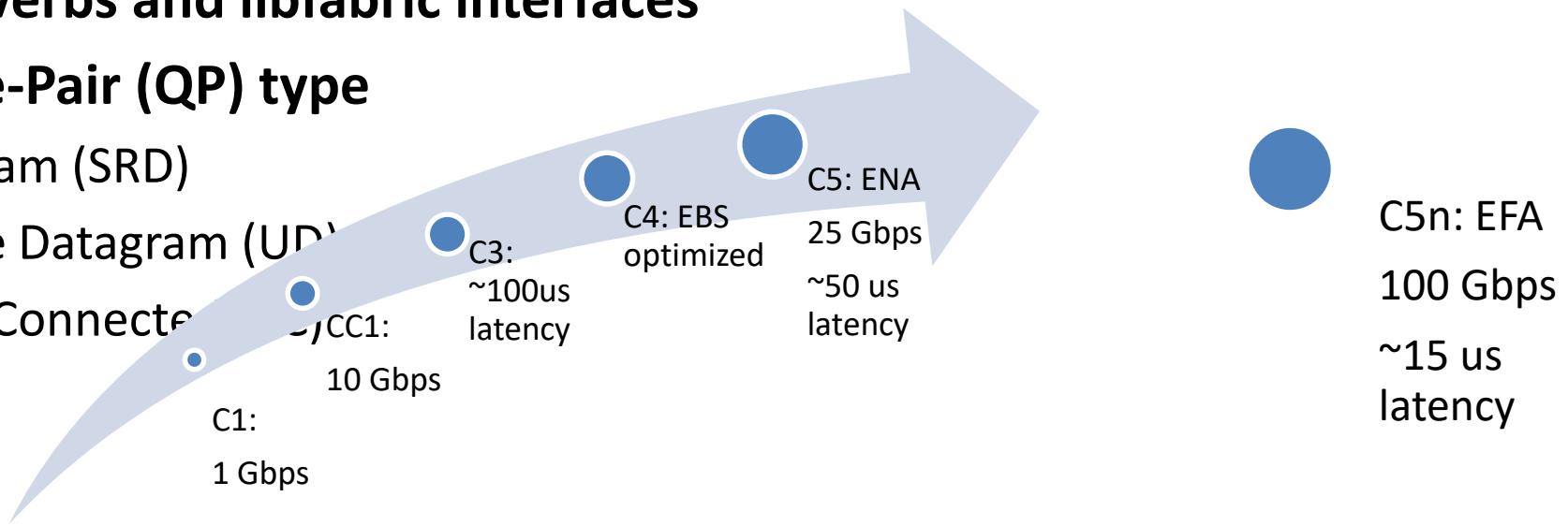
Increasing Need to Run these applications on the Cloud!!

AGENDA

- Introduction & Motivation
- Support for AWS EFA
  - Overview
  - Challenges & Solutions for MPI Libraries Design on EFA
  - Experimental Evaluation
- Support for Azure VM
  - Performance Evaluation
  - Deployment
- Conclusions & Future Plans

AMAZON ELASTIC FABRIC ADAPTER (EFA)

- **Enhanced version of Elastic Network Adapter (ENA)**
- **Allows OS bypass, up to 100 Gbps bandwidth**
- **Network aware multi-path routing**
- **Exposed through libibverbs and libfabric interfaces**
- **Introduces new Queue-Pair (QP) type**
  - Scalable Reliable Datagram (SRD)
  - Also supports Unreliable Datagram (UD)
  - No support for Reliable Connecte



Deep Dive on OpenMPI and Elastic Fabric Adapter (EFA) - AWS Online Tech Talks, Linda Hedges

Evolution of networking on AWS

SCALABLE RELIABLE DATAGRAMS (SRD): FEATURES & LIMITATIONS

Feature	UD	SRD
Send/Recv	✓	✓
Send w/ Immediate	✗	✗
RDMA Read/Write/Atomic	✗	✗
Scatter Gather Lists	✓	✓
Shared Receive Queue	✗	✗
Reliable Delivery	✗	✓
Ordering	✗	✗
Inline Sends	✗	✗
Global Routing Header	✓	✗
Max Message Size	4KB	8KB

- Similar to IB Reliable Datagram
  - No limit on number of outstanding messages per context
- Out of order delivery
  - No head-of-line blocking
  - Bad fit for MPI, can suit other workloads
- Packet spraying over multiple ECMP paths
  - No hotspots
  - Fast and transparent recovery from network failures
- Congestion control designed for large scale
  - Minimize jitter and tail latency

Amazon Elastic Fabric Adapter: Anatomy, Capabilities, and the Road Ahead, Raghu Raja, OpenFabrics Workshop 2019

CHALLENGE 1: RELIABLE AND IN-ORDER DELIVERY

- **Challenges**
  - MPI guarantees reliable and in-order message matching to applications
  - UD does not provide reliability or ordering
  - SRD provides reliability but not in-order delivery

- **Solution:**
  - Use acknowledgements and retransmissions for reliability
  - Piggy back acks on application messages for reducing overhead
  - Use sequence number and sliding window for re-ordering packets at the receiver process

M. J. Koop, S. Sur, Q. Gao, and D. K. Panda, "High Performance MPI Design using Unreliable Datagram for Ultra-scale InfiniBand Clusters," in *Proceedings of the 21st annual international conference on Supercomputing*

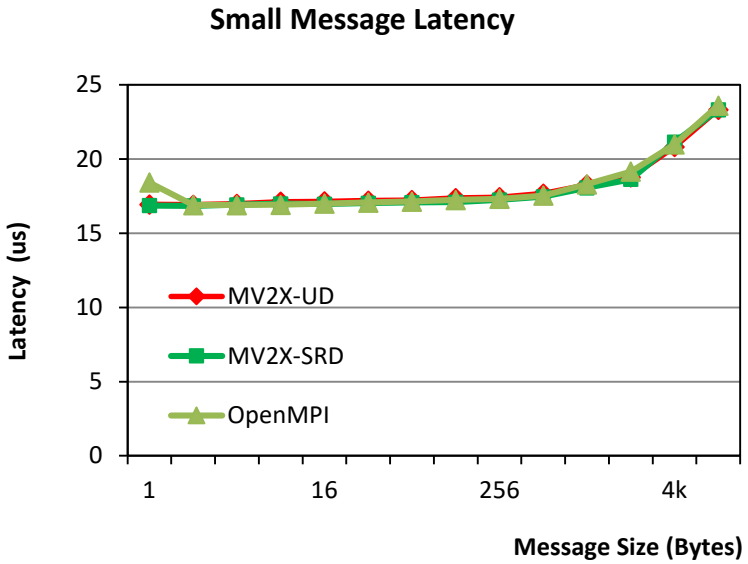
CHALLENGE 2: ZERO-COPY TRANSMISSION OF LARGE MESSAGES

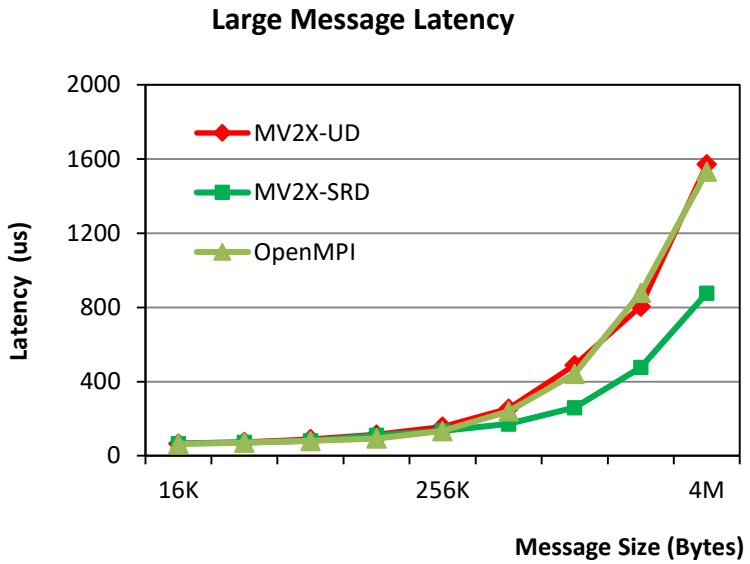
- **Challenges:**
  - MPI allows sending and receiving very large messages
  - Network message size bound by MTU size (4KB for UD, 8KB for SRD)
  - Need to handle segmentation and reassembly
  - Existing zero-copy designs* can not be used
    - Utilizes send-with-immediate for sequence numbers (not supported by EFA)
    - Retransmits entire message if out-of-order arrival is detected

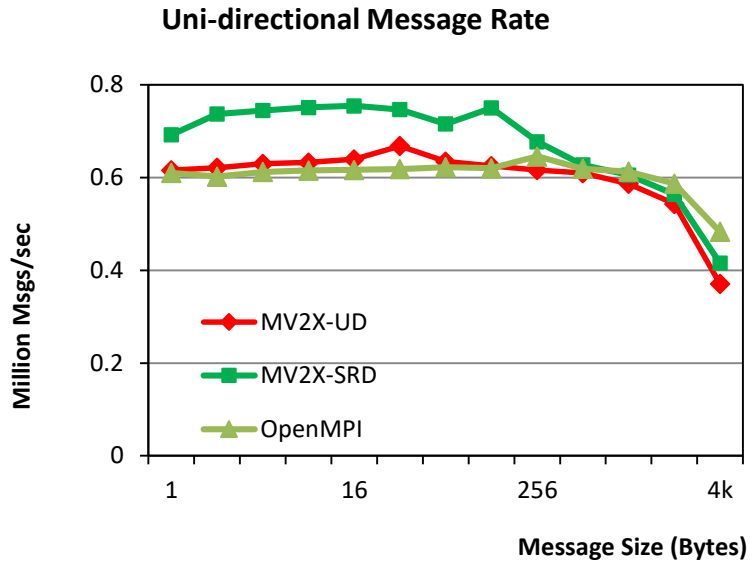
- **Solution: propose new design for zero-copy rendezvous transfers**
  - Maintain a pool of dedicated QPs for zero-copy transfers
  - Use scatter gather lists for sequence numbers
  - Reorder out-of-order packets at the receiver

* M. J. Koop, S. Sur, and D. K. Panda, "Zero-copy Protocol for MPI using InfiniBand Unreliable Datagram," in *2007 IEEE International Conference on Cluster Computing*

POINT-TO-POINT PERFORMANCE





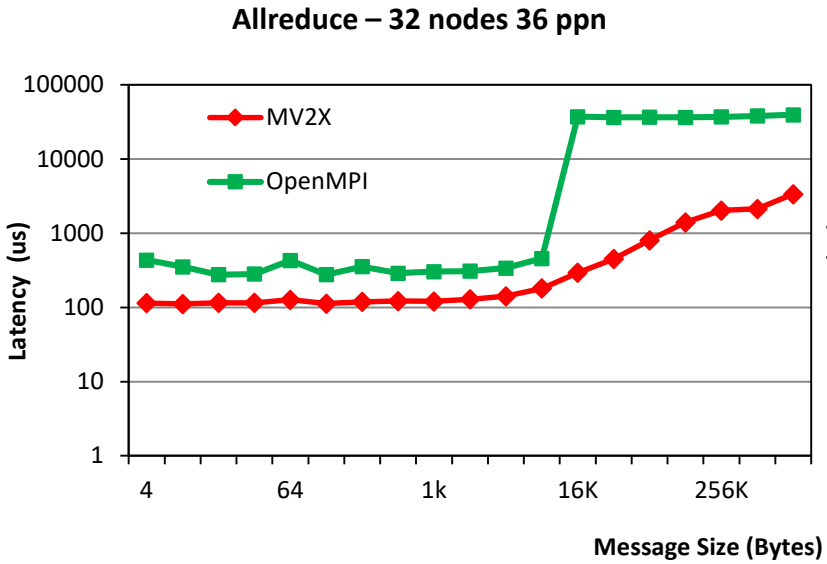


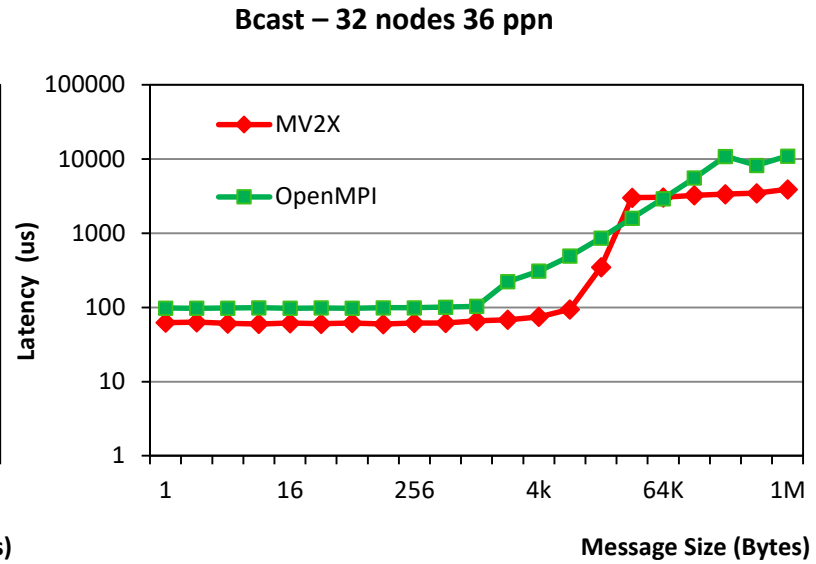
- Both UD and SRD shows similar latency for small messages
- SRD shows higher message rate due to lack of software reliability overhead
- SRD is faster for large messages due to larger MTU size

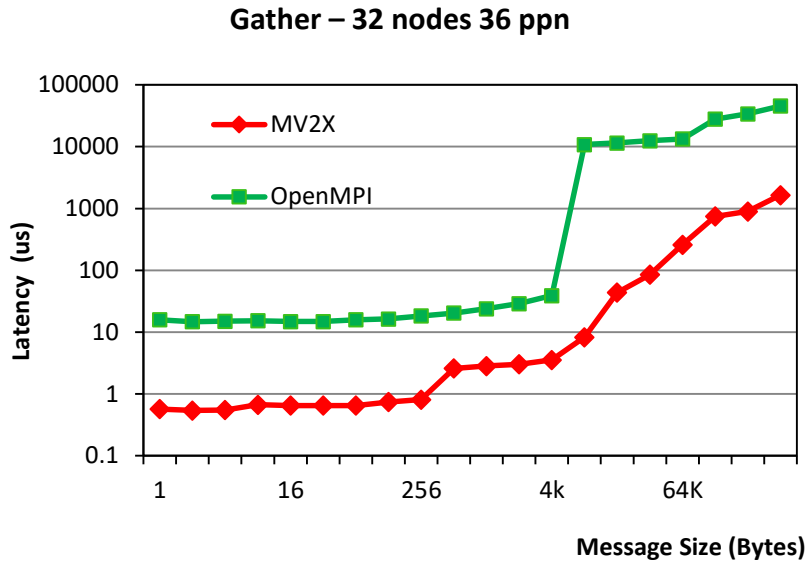
S. Chakraborty, S. Xu, H. Subramoni, and D. K. Panda, *Designing Scalable and High-performance MPI Libraries on Amazon Elastic Fabric Adapter,* 26th Symposium on High Performance Interconnects, (HOTI '19)

- **System Setup:**
  - Instance type: c5n.18xlarge
  - CPU: Intel Xeon Platinum 8124M @ 3.00GHz
  - Cores: 2 Sockets, 18 cores / socket
  - KVM Hypervisor, 192 GB RAM, One EFA adapter / node
  - MVAPICH2 version: MVAPICH2-X + SRD support
  - OpenMPI version: OpenMPI-4.0.3 with libfabric 1.9

COLLECTIVE PERFORMANCE



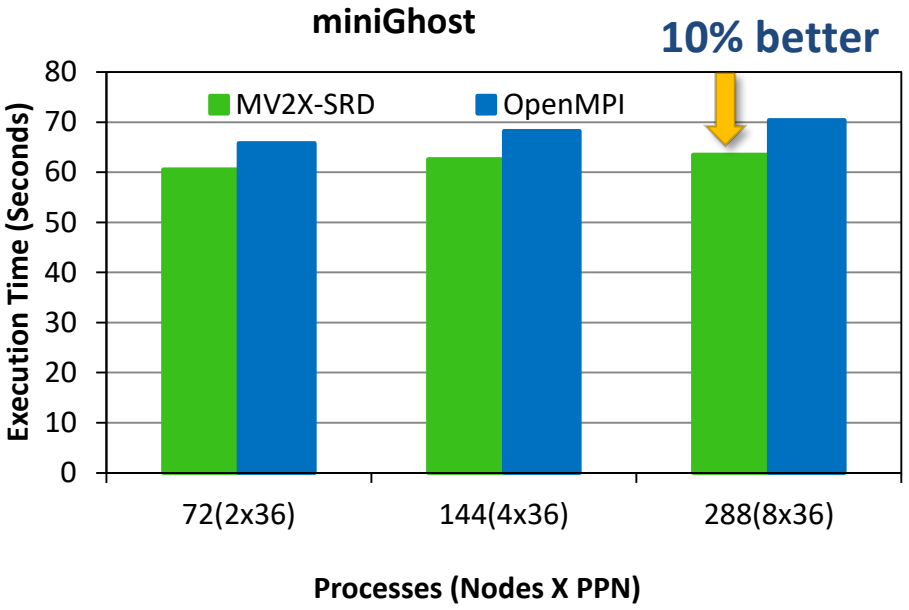


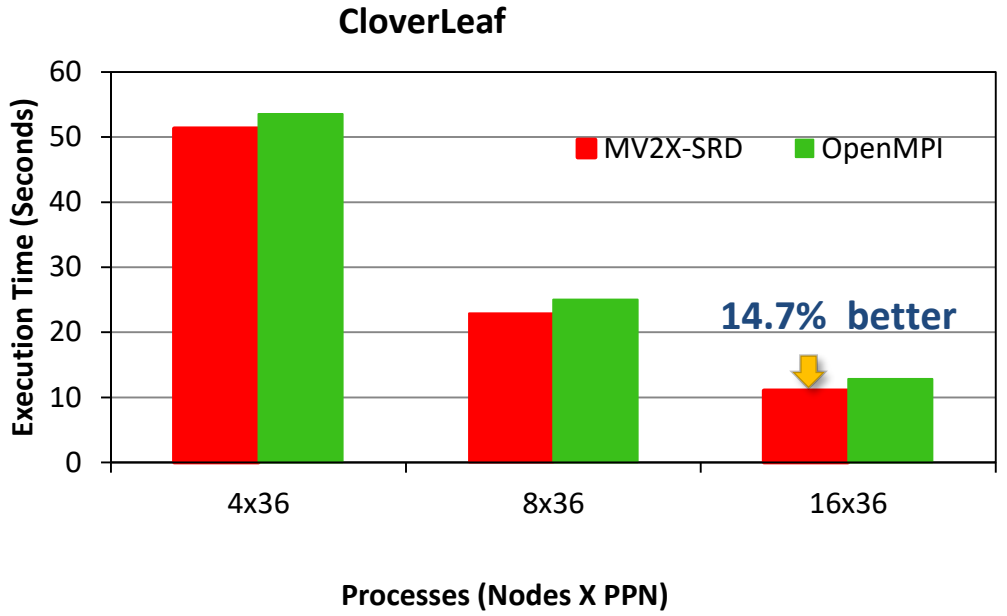


- Up to 25x better performance in Allreduce
- Up to 5x better performance in Bcast
- Up to 100x better performance in Gather

- **System Setup:**
  - Instance type: c5n.18xlarge
  - CPU: Intel Xeon Platinum 8124M @ 3.00GHz
  - Cores: 2 Sockets, 18 cores / socket
  - KVM Hypervisor, 192 GB RAM, One EFA adapter / node
  - MVAPICH2 version: MVAPICH2-X + SRD support
  - OpenMPI version: OpenMPI-4.0.3 with libfabric 1.9

APPLICATION PERFORMANCE





- Up to 10% performance improvement for MiniGhost on 8 nodes
- Up to 14.7% better performance with CloverLeaf on 16 nodes

Instance type: c5n.18xlarge
CPU: Intel Xeon Platinum 8124M @ 3.00GHz
MVAPICH2 version: Latest MVAPICH2-X + SRD support
OpenMPI version: Open MPI v4.0.3 with libfabric 1.9

AGENDA

- Introduction & Motivation
- Support for AWS EFA
  - Overview
  - Challenges & Solutions for MPI Libraries Design on EFA
  - Experimental Evaluation
- Support for Azure VM
  - Performance Evaluation
  - Deployment
- Conclusions & Future Plans

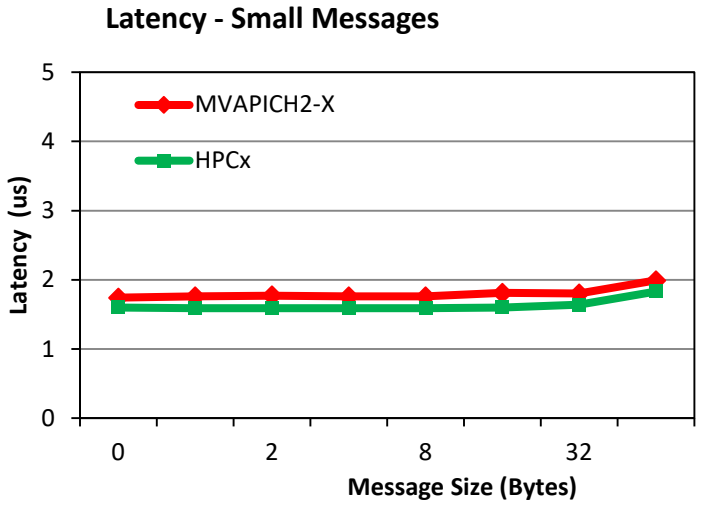
MVAPICH2 IN AZURE HPC ENVIRONMENTS

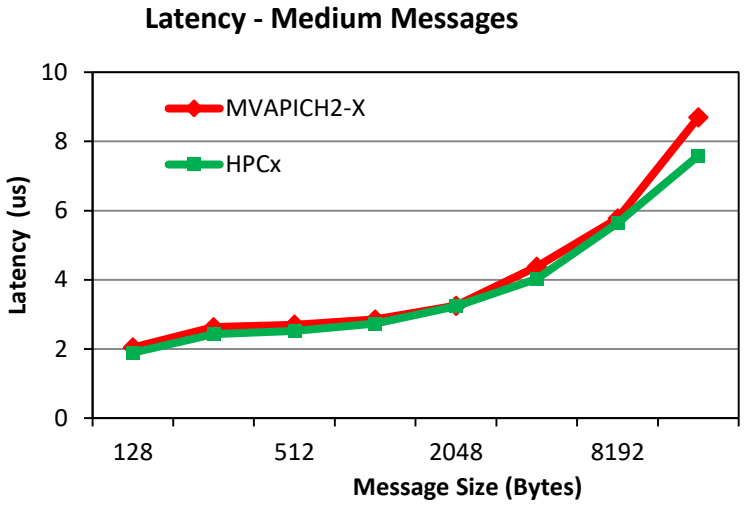
- Azure has been using RDMA-enabled network and software stacks for the last several years
- Moved to native InfiniBand support with Mellanox OFED for new instances (HB, HC, HB2, and upcoming ones)
- Uses SR-IOV support for virtualization
- MVAPICH2 libraries have been optimized and tuned fro Azure HB, HC and HB2 instances
- Uses one VM per node

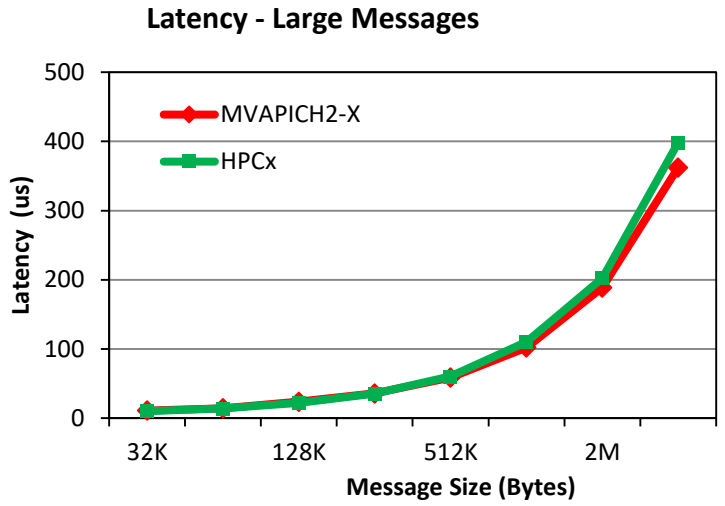
EVALUATION WITH AZURE HB AND HC VM TYPES

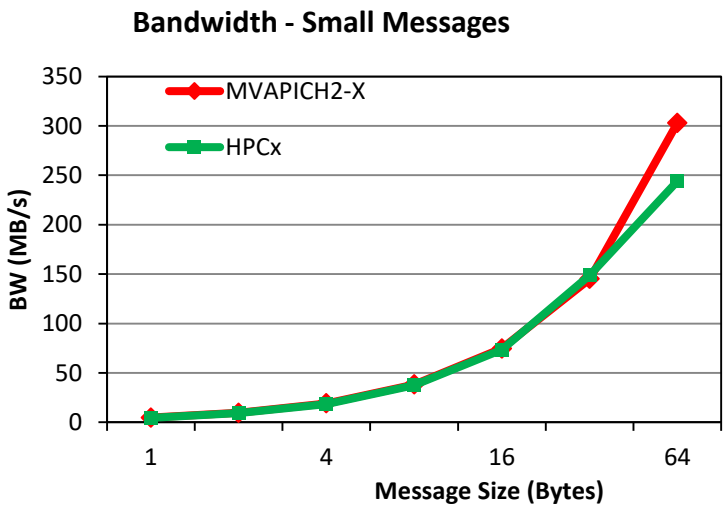
- System Configuration for Performance Evaluation
  - VM types:
    - Azure HC: CPU: Intel Xeon Platinum 8168 @ 2.70GHz, 44 cores
    - Azure HB: AMD EPYC 7551 @ 2 GHz, 60 cores
  - MVAPICH2 Version: Latest MVAPICH2-X w/ XPMEM support
  - HPCx Version: Built-in HPCx-v2.5.0-gcc-MLNX_OFED_LINUX-4.7-1.0.0.1-redhat7.6-x86_64
  - OMB Version: OSU-MicroBenchmars-5.6.2

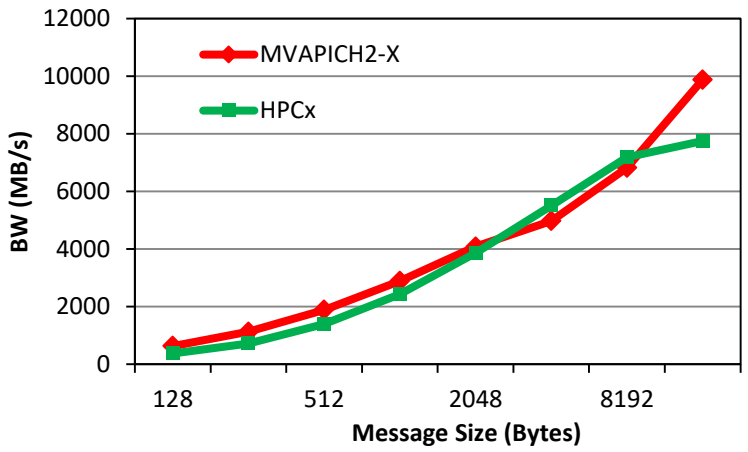
PERFORMANCE: INTER-NODE POINT-TO-POINT (HC)

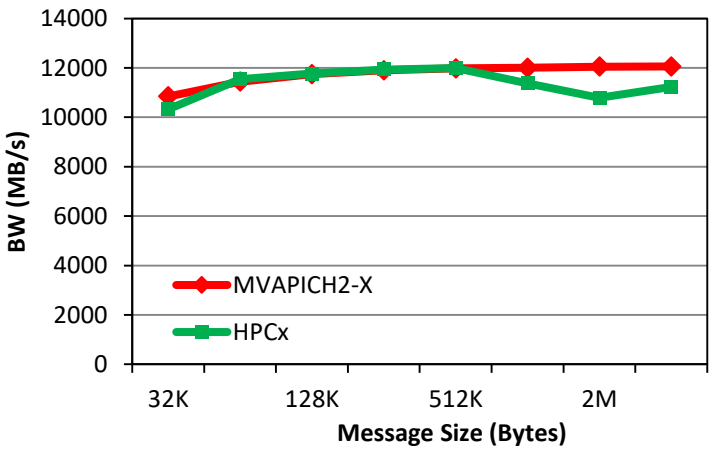




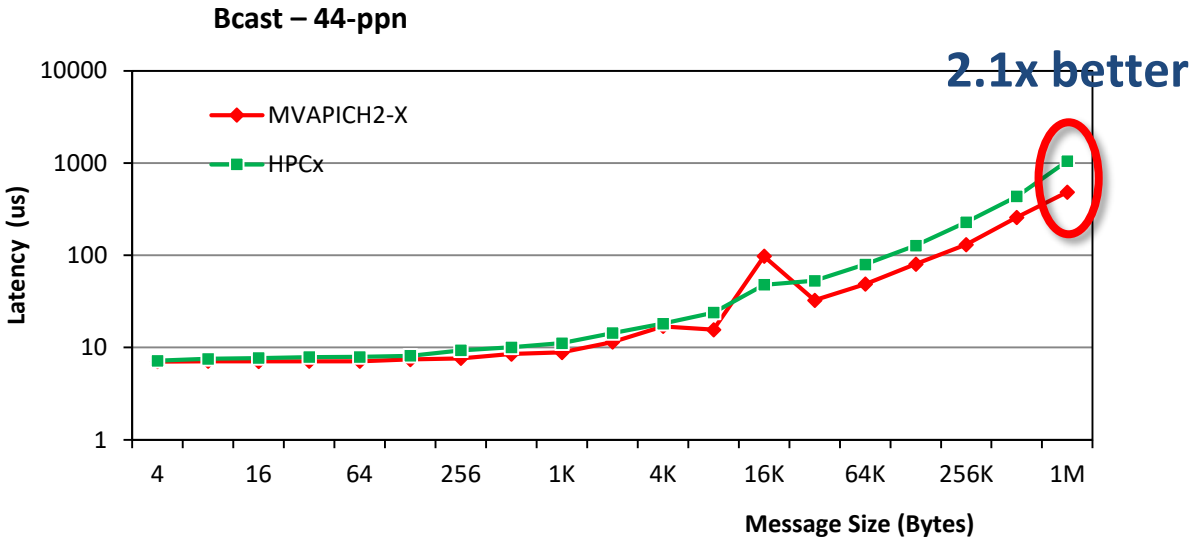


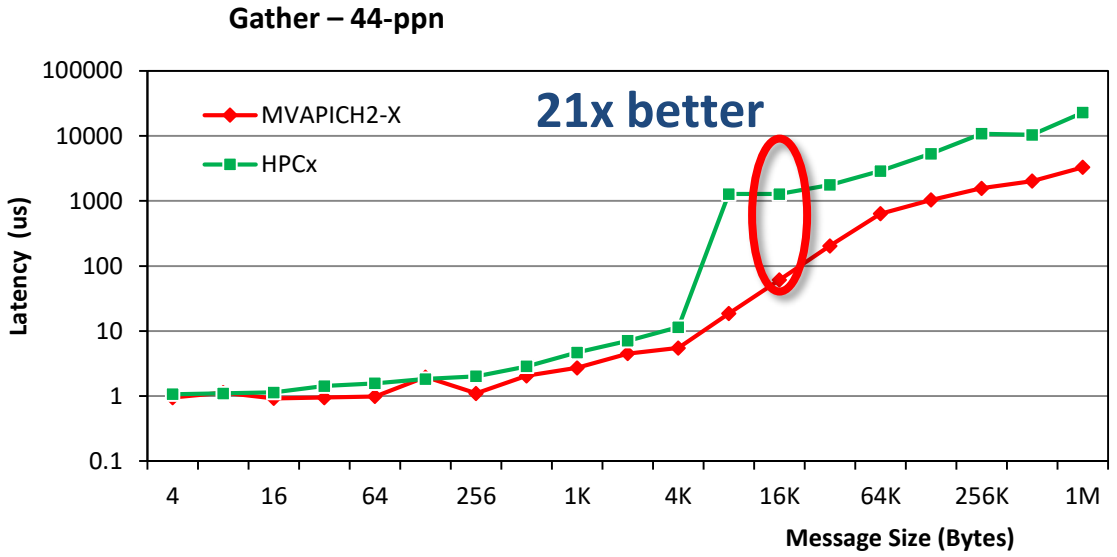


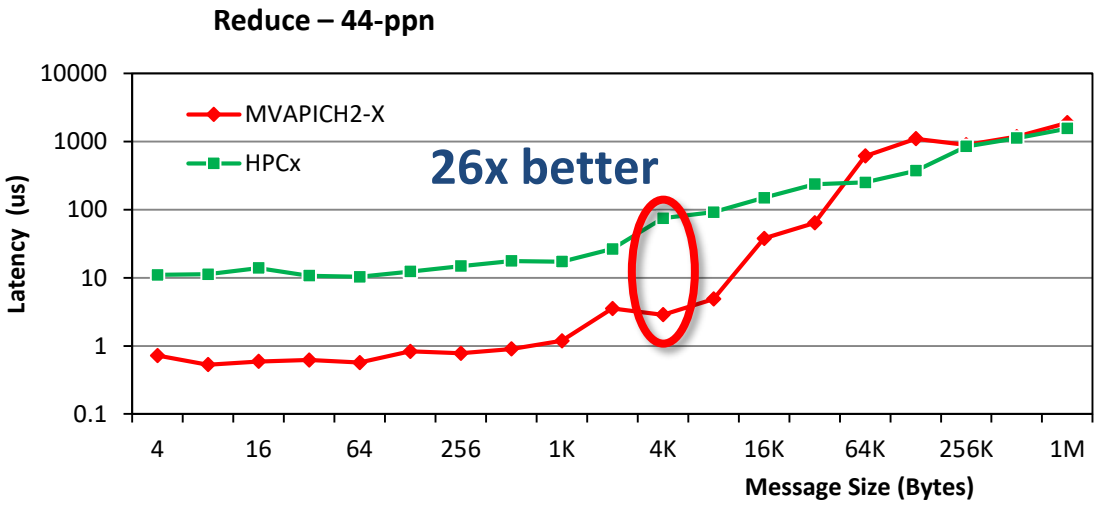


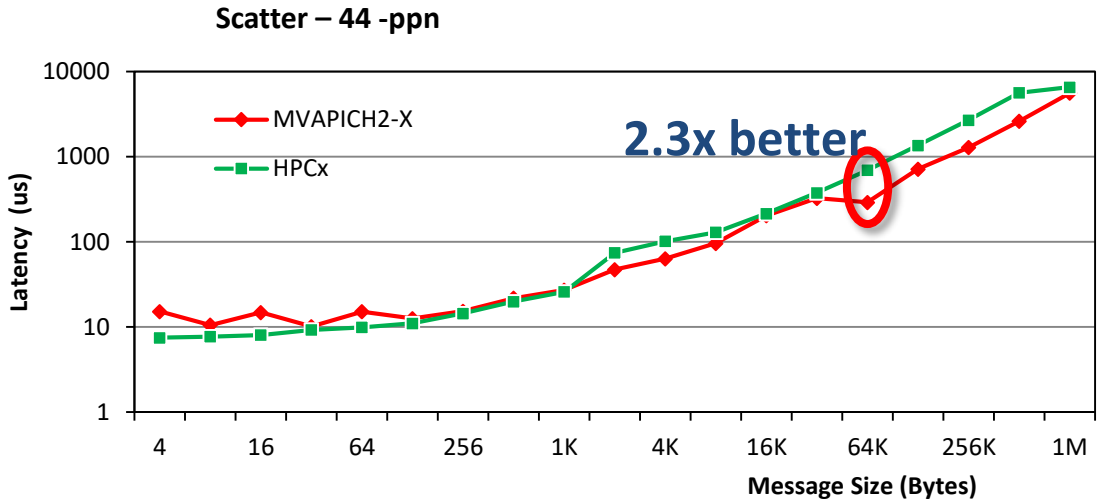


PERFORMANCE: 8-NODE COLLECTIVES (HC)

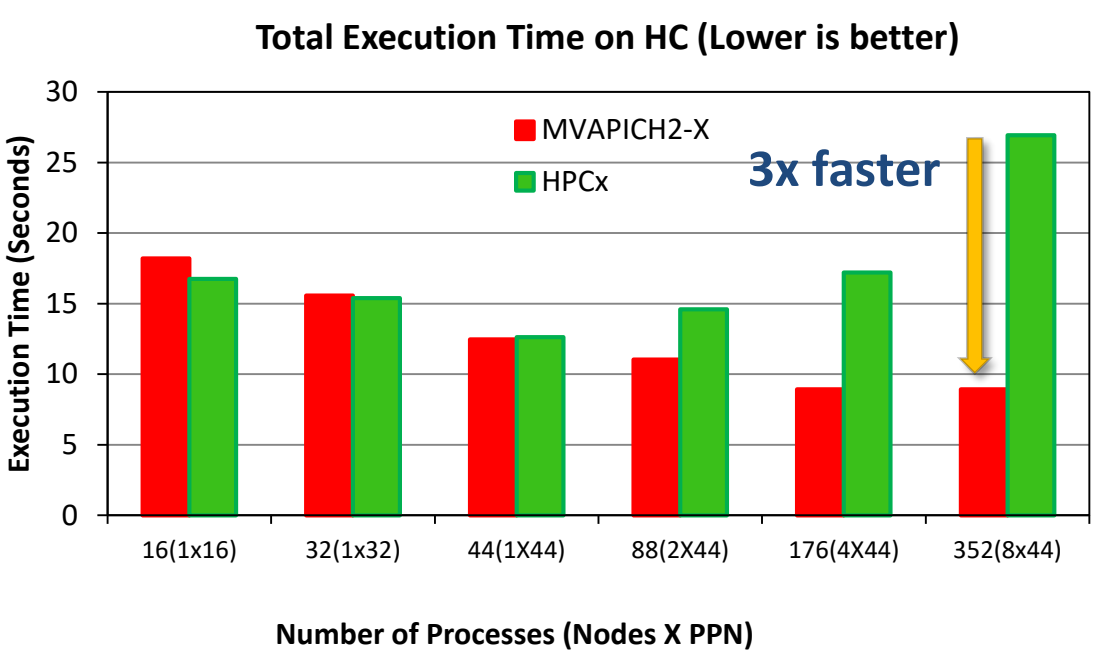


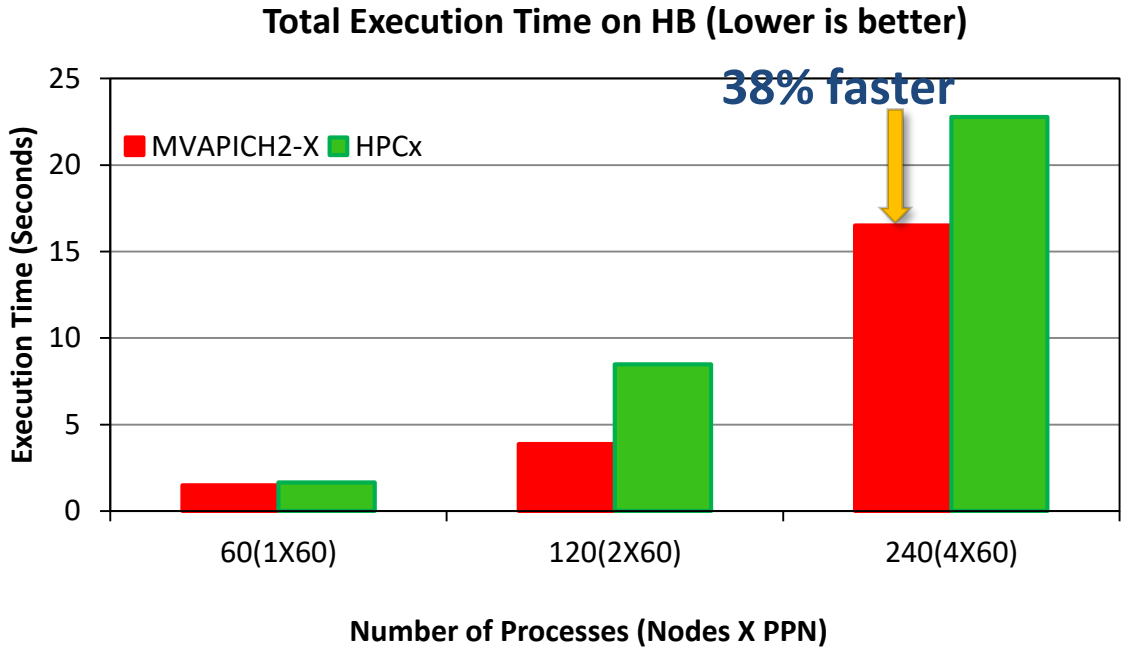






PERFORMANCE OF APPLICATION (RADIX)



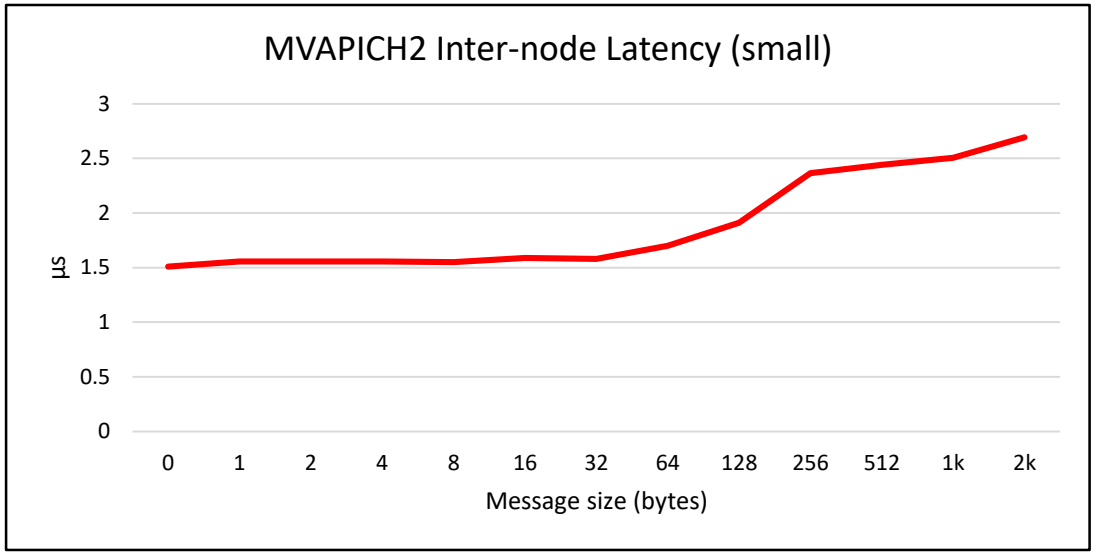


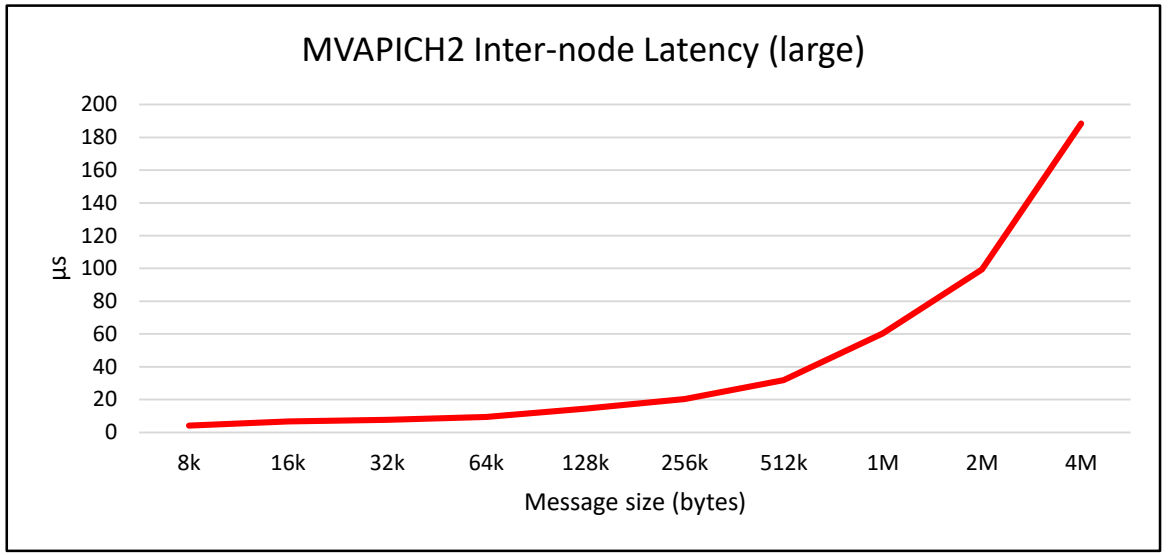
- MVAPICH2-X performs up to 3x better performance compared to HPCx on HC
- MVAPICH2-X performs up to 30% better performance compared to HPCx on HB

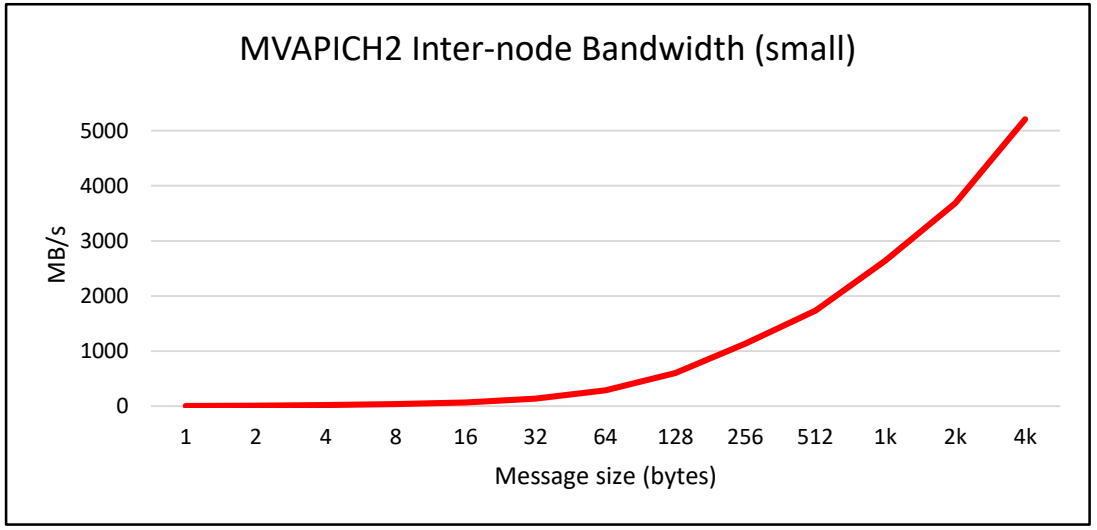
EVALUATION WITH AZURE HB2 VM TYPE

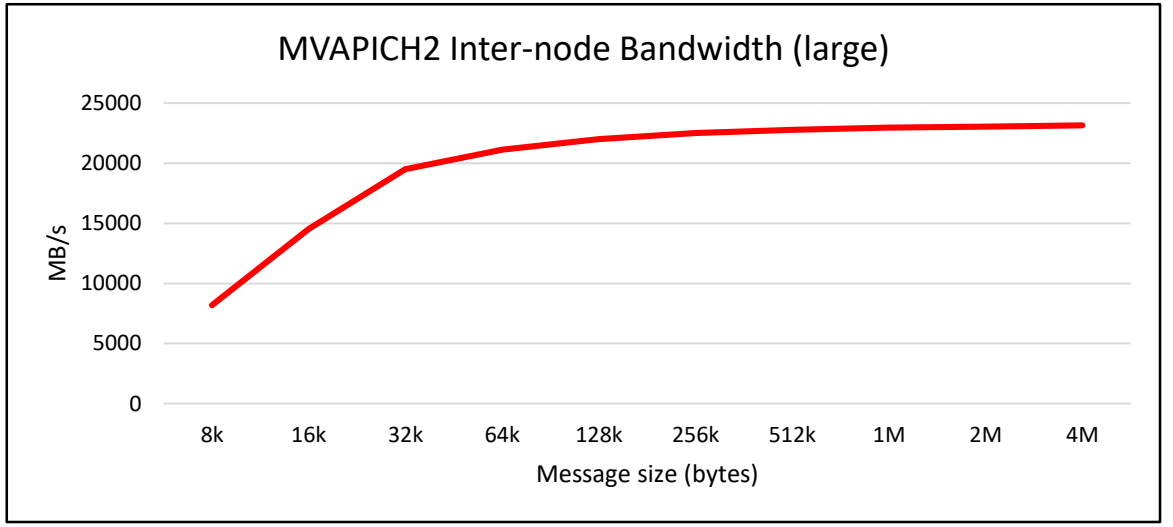
- System Configuration for Performance Evaluation
  - VM type: Azure HB2
  - CPU: AMD EPYC 7V12 CPU @ 2.45GHz
  - Cores: 120 cores
  - MVAPICH2 2.3.4
  - MVAPICH2-X 2.3.3rc3 w/ XPMEM support
  - OMB Version: OSU-MicroBenchmars-5.6.2

PERFORMANCE ON HBV2 INSTANCES – MVAPICH2 2.3.3

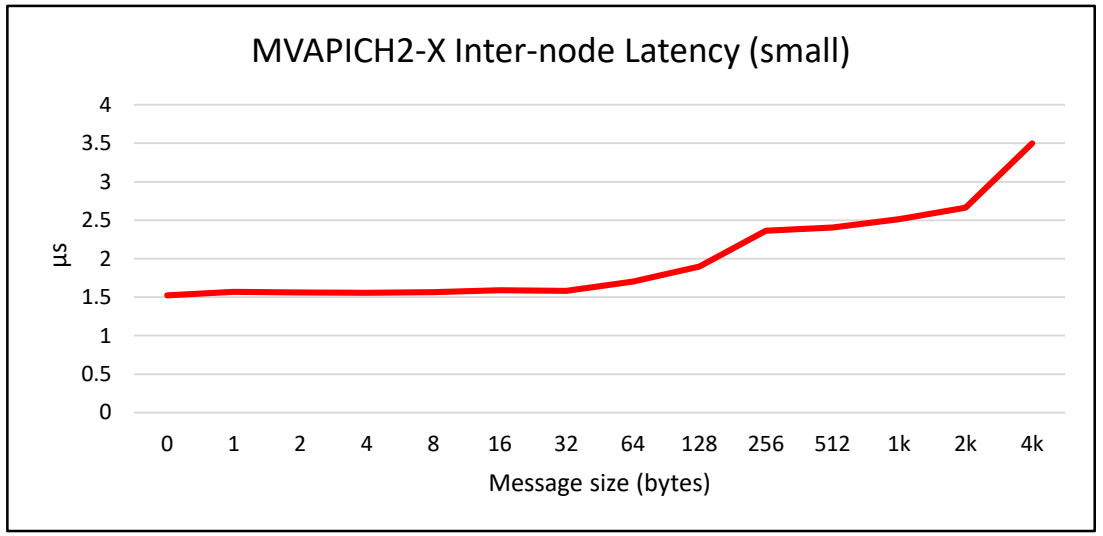


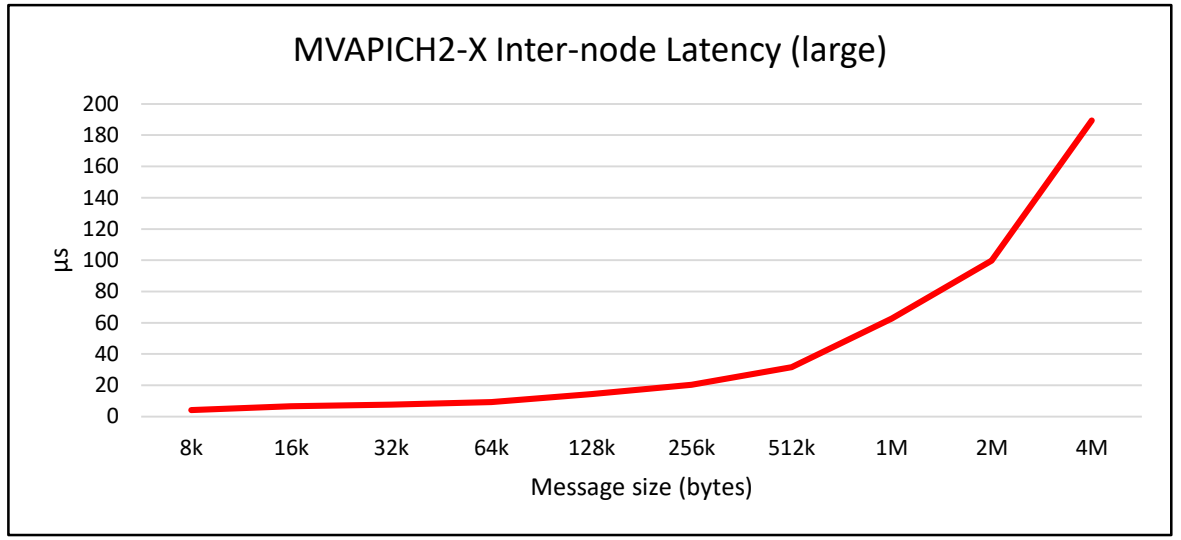


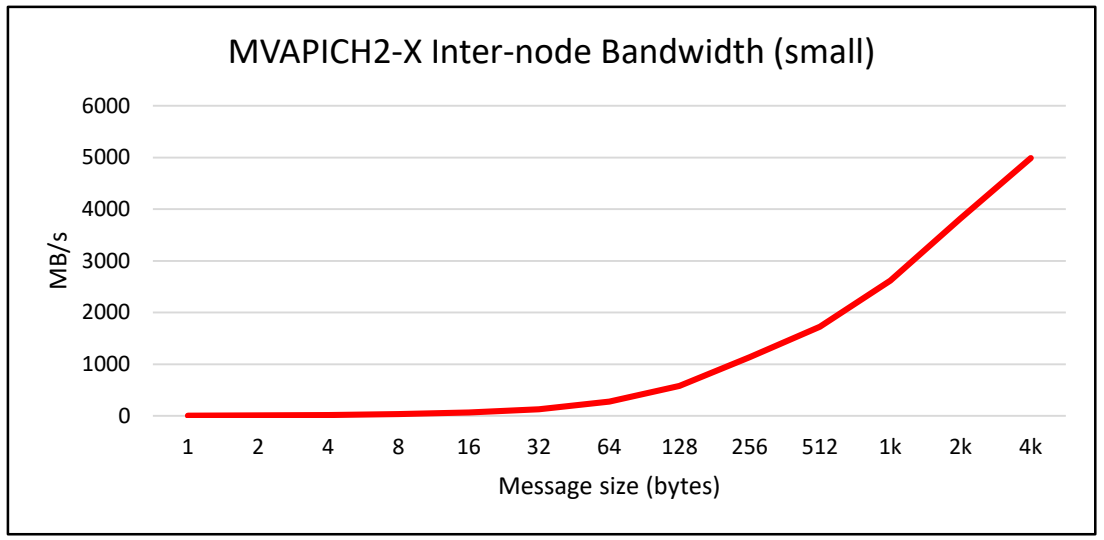


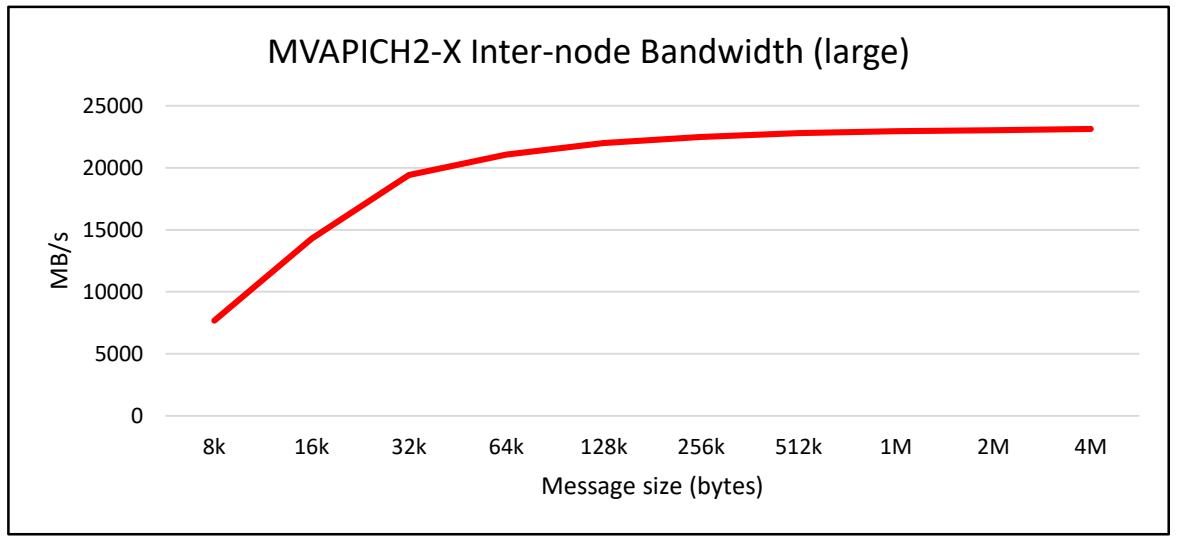


PERFORMANCE ON HBV2 INSTANCES – MVAPICH2-X 2.3.RC3









MVAPICH2-AZURE DEPLOYMENT

- Released on 05/20/2020
- Integrated Azure CentOS HPC Images
  - https://github.com/Azure/azhpc-images/releases/tag/centos-7.6-hpc-20200417
- MVAPICH2 2.3.3
  - CentOS Images (7.6, 7.7 and 8.1)
  - Tested with multiple VM instances
- MVAPICH2-X 2.3.RC3
  - CentOS Images (7.6, 7.7 and 8.1)
  - Tested with multiple VM instances
- More details from Azure Blog Post
  - https://techcommunity.microsoft.com/t5/azure-compute/mvapich2-on-azure-hpc-clusters/ba-p/1404305

AGENDA

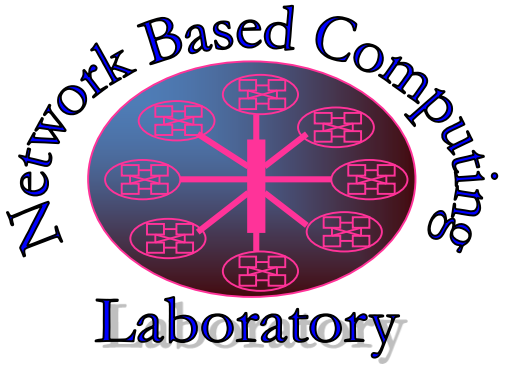
- Introduction & Motivation
- Support to AWS EFA
  - Overview
  - Challenges & Solutions for MPI Libraries Design on EFA
  - Experimental Evaluation
- Support to Azure VM
  - Dedicated Performance Evaluation & Tuning
  - One-click Quick Deployment
- Conclusions & Future Plans

CONCLUSION & FUTURE PLANS

- **HPC workloads are being run on cloud environments**
- **MVAPICH2 deployments are available on AWS and Azure for users to take advantage of high-performance and scalability**
- **On AWS EFA**
  - MVAPICH2-X for AWS 2.3 available
  - Includes support for SRD and XPMEM based transports
    - Available for download from http://mvapich.cse.ohio-state.edu/downloads/
  - Latest MVAPICH2-X 2.3.3 GA with performance enhancements is being tested for deployment
- **On Azure:**
  - MVAPICH2 2.3.3 and MVAPICH2-X 2.3.rc3 are available as Integrated CentOS images
  - Latest MVAPICH2 2.3.4 and MVAPICH2-X 2.3.3 GA will be available soon
- **These versions will be available through respective Market Places soon**
- **Commercial Support available for End-Users, ISVs, and Organizations through X-Scale Solutions (http://x-scalesolutions.com)**

THANK YOU!

{xu.2452, ghazimirsaeed.3, subramoni.1} @osu.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/