OPENFABRICS ALLIANCE

2022 OFA Virtual Workshop

MPI Library Performance on AWS Arm-based HPC Cloud with Elastic Fabric Adapter

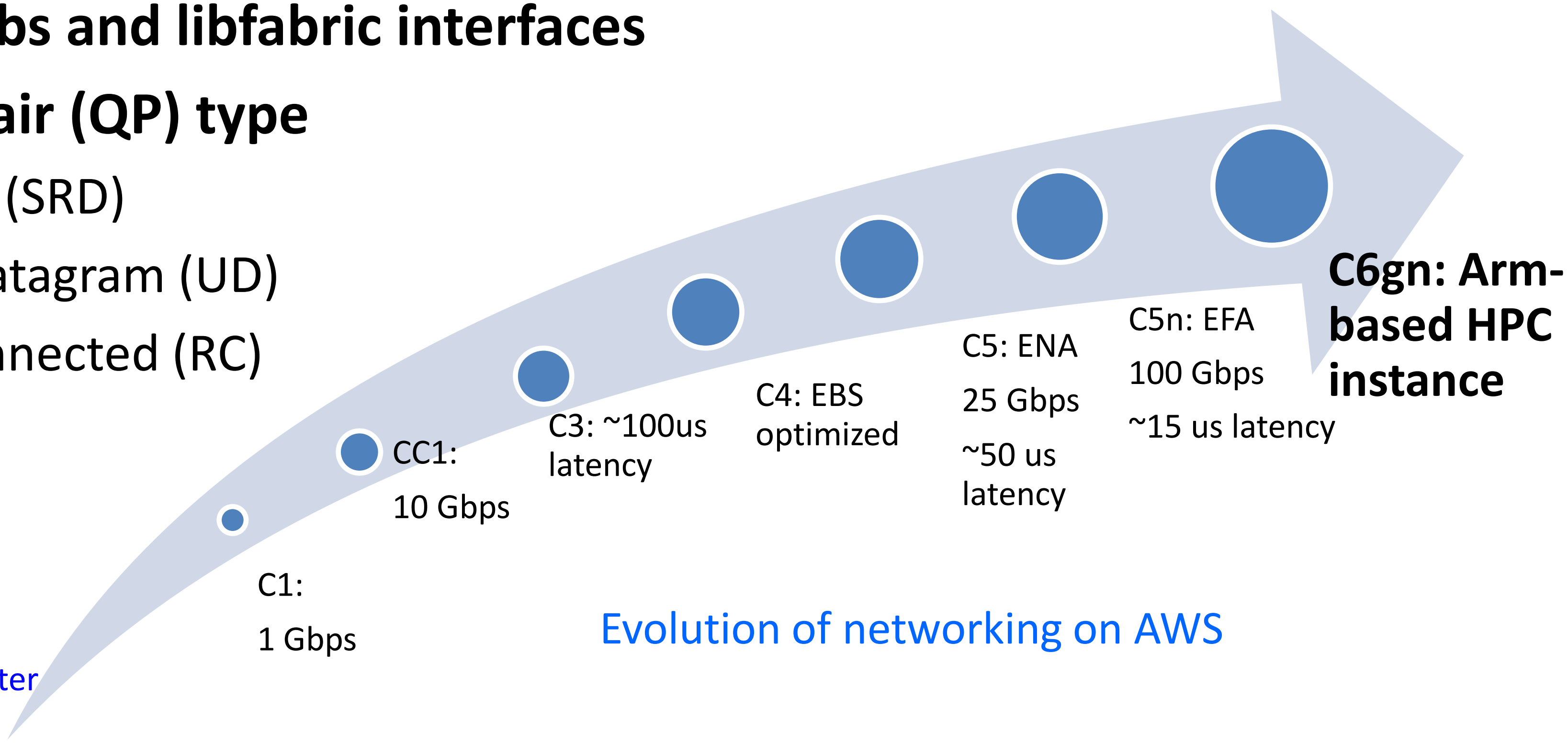
Shulei Xu, **Aamir Shafi**, Hari Subramoni, and Dhabaleswar K. (DK) Panda
The Ohio State University
panda@cse.ohio-state.edu

OUTLINE

- **Introduction**
- MPI Optimization
- Performance Evaluation
  - Micro-benchmark level Performance
  - Application level Performance
- Conclusion

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 Connected (RC)



Evolution of networking on AWS

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

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

RECENT UPDATES IN AWS EC2 INSTANCES FOR HPC WORKLOADS

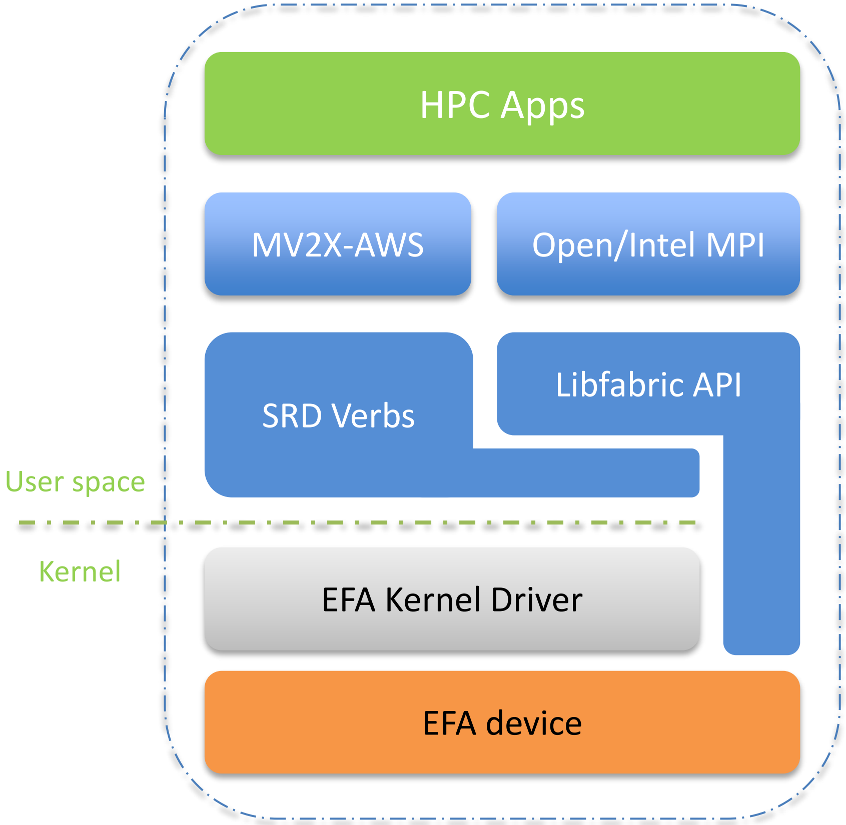
- **Various hardware selection**
  - Support both x86 (Intel/AMD) & Arm based CPU types
  - Support multiple hardware configuration choices including vCPUs count, storage and network bandwidth

- **Recent supported Arm-based HPC instances**
  - Custom-built by AWS using 64-bit Arm Neoverse cores to enable the best price performance for workloads running in Amazon EC2
  - Support up to 100 Gbps networking bandwidth, 38 Gbps Elastic Block Store (EBS) bandwidth

- **Quickly deploy HPC environments with AWS Parallelcluster**
  - Support multiple instance types and job schedulers like Slurm
  - Support OS type Amazon Linux2, CentOS 7, Ubuntu 18.04 and 20.04
  - Flexible cost-effective allocation, launch when need

MPI LIBRARIES ON AWS EC2 HPC INSTANCES

- **Supports MPI libraries on instances with EFA support**
- **OpenMPI & IntelMPI are based on Libfabric API**
  - Libfabric Bypass the OS kernel and can communicate directly with EFA device
- **MVAPICH2-X-AWS is based directly on SRD verbs API**
  - Different to Open MPI and IntelMPI, directly invokes SRD verbs API to implement MPI level communication
  - Detail design is included in this paper:
    - Designing Scalable and High-performance MPI Libraries on Amazon Elastic Fabric Adapter, S. Chakraborty , S. Xu , H. Subramoni , DK Panda, HotI 19, Aug 2019



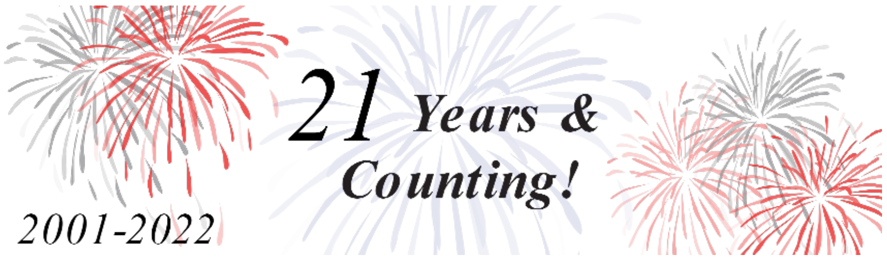
Get started with EFA and MPI, https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/efa-start.html

OUTLINE

- Introduction
- **MPI Optimization**
- Performance Evaluation
  - Micro-benchmark level Performance
  - Application level Performance
- Conclusion

OVERVIEW OF THE MVAPICH2 PROJECT

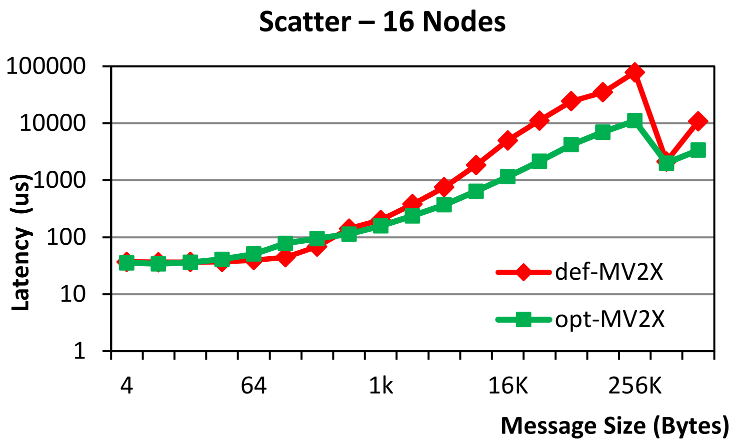
- **High Performance open-source MPI Library**
- **Support for multiple interconnects**
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, Rockport Networks, and Slingshot
- **Support for multiple platforms**
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and 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 (since 2014) and AMD (since 2020) GPUs
  - 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

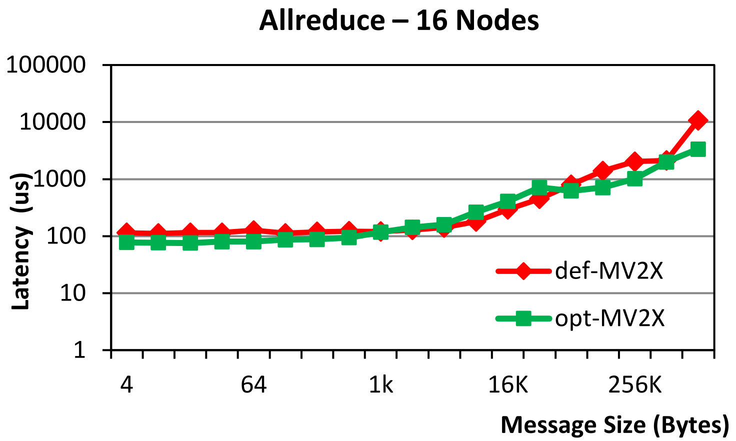


- Used by more than 3,200 organizations in 89 countries
- More than 1.57 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '21 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 13th, 448, 448 cores (Frontera) at TACC
  - 26th, 288,288 cores (Lassen) at LLNL
  - 38th, 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 13th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

MPI OPTIMIZATION

- **Collective algorithm tuning**
  - Systematically iterate through different MVAPICH2 collective algorithms for all ***number_of_nodes*** x ***ppn*** combinations, and determine algorithms with best performance for each scenario.
- **XPMEM kernel module optimization**
  - User-level API for multiple processes share address space
  - Automatically detect XPMEM module in OS, and apply optimization if it is loaded.
    - Using *dlopen* to open *libxpmem* on runtime
  - Improve point-to-point & collective intra-node large message communication performance.
- **Examples of collective performance difference are shown on the right**





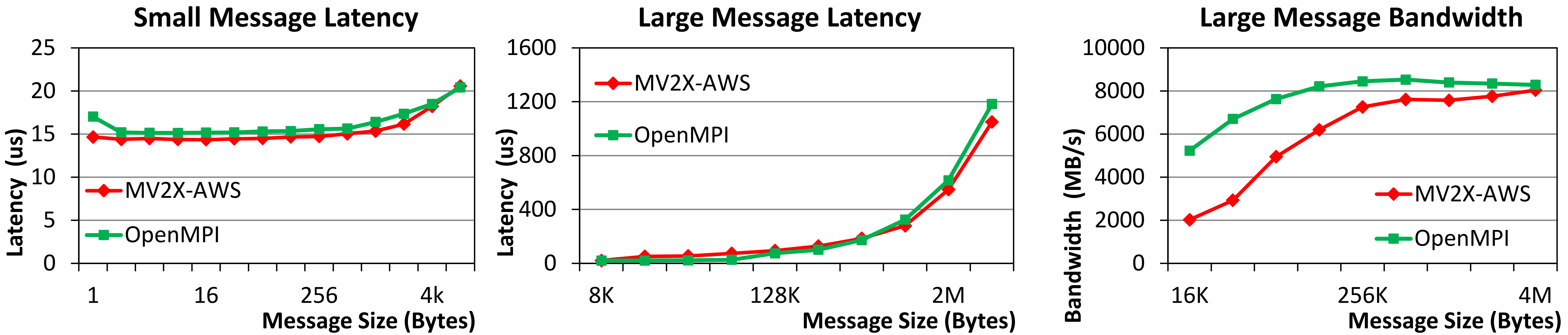
OUTLINE

- Introduction
- MPI Optimization
- **Performance Evaluation**
  - **Experimental Setups**
  - **Micro-benchmark level Performance**
  - **Application level Performance**
- Conclusion

EXPERIMENTAL SETUP

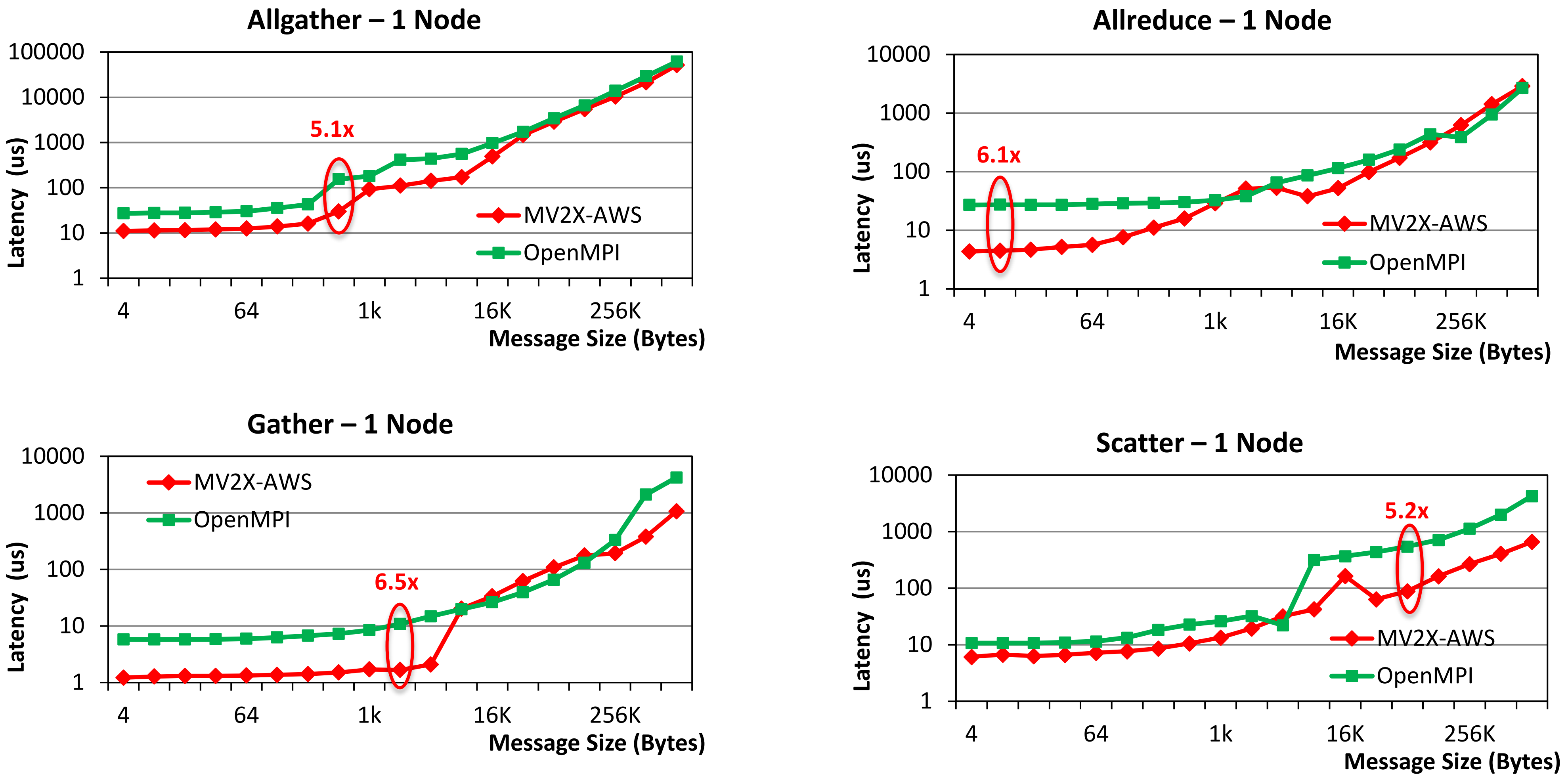
- **Experiment System Specification**
  - Instance Type: c6gn.16xlarge
  - RAM (DDR4): 128 GB
  - Libfabric version: 1.13.2
  - Parallel cluster: 3.0.2
- **MPI libraries & benchmark Specification:**
  - MVAPICH2: Latest Mvapich2-X-AWS
  - OpenMPI: 4.1.0 (Parallelcluster built-in)
  - OSU Micro-benchmarks: 5.8

PERFORMANCE EVALUATION

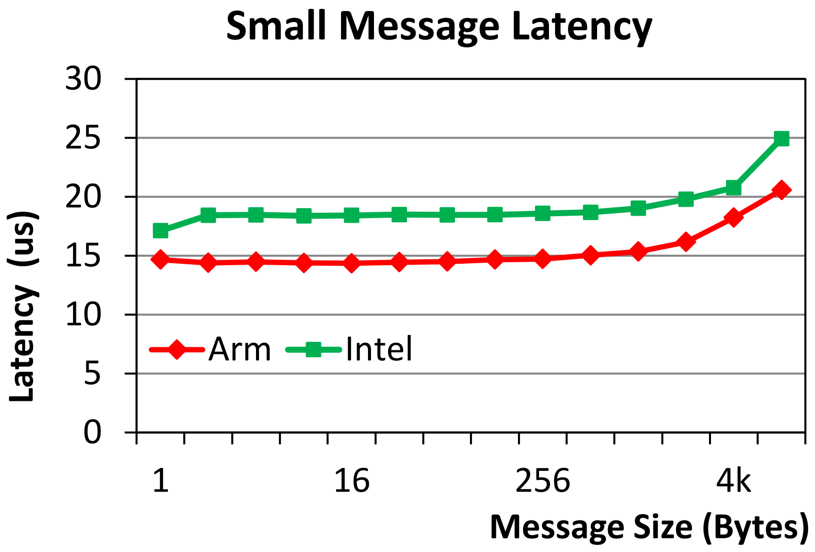


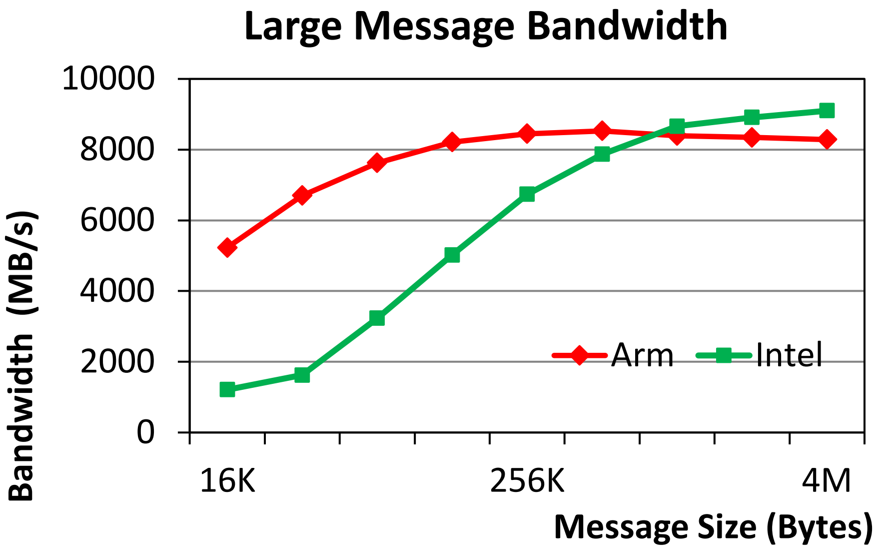
- **Point-to-point communication performance**

SINGLE NODE COLLECTIVE PERFORMANCE



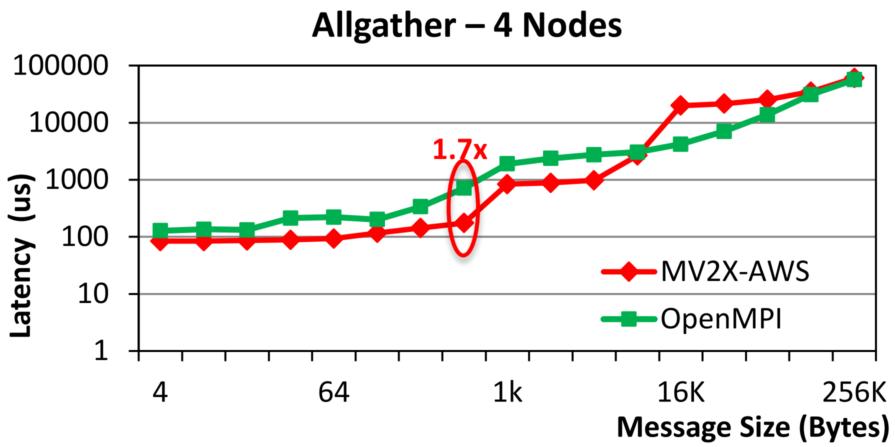
MVAPICH2-X-AWS CROSS ARCHITECTURE COMPARISON

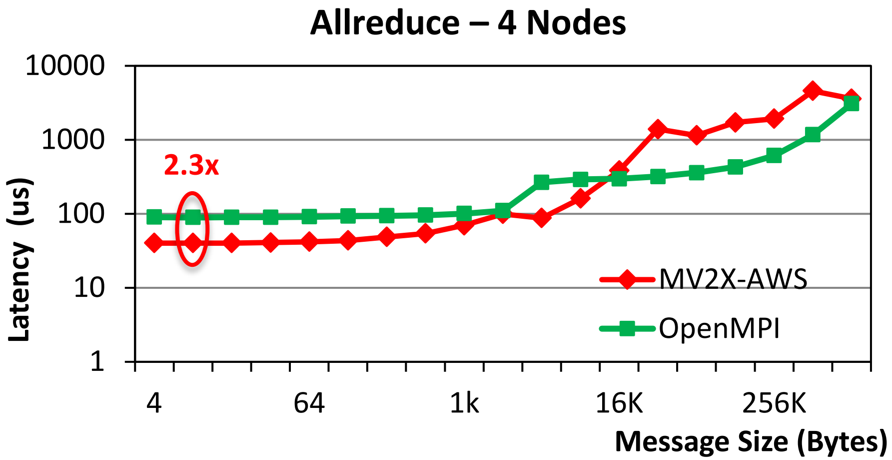


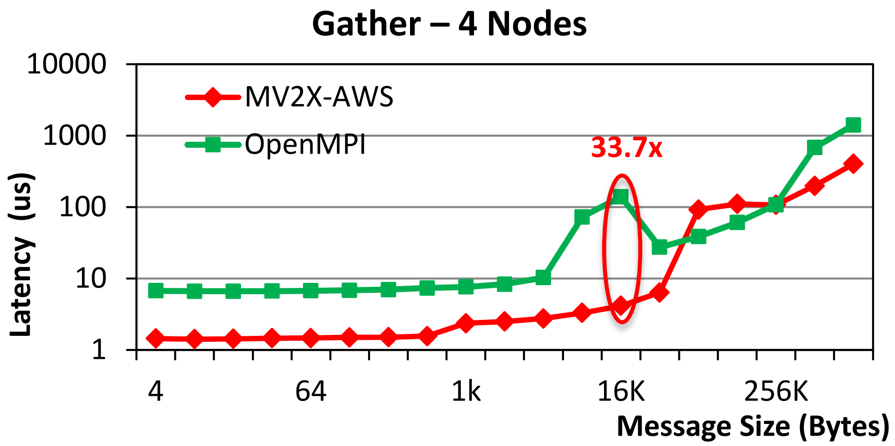


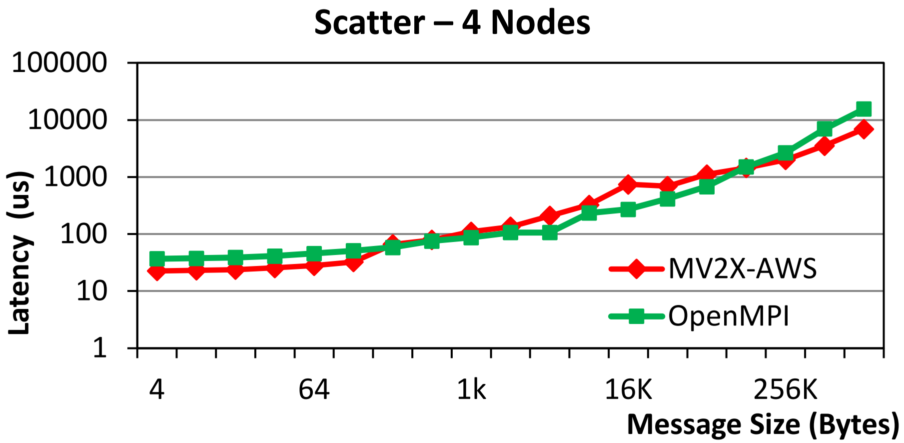
- Comparing basic MPI point-to-point performance on AWS Arm (c6gn.16xlarge) vs. x86 (c5n.18xlarge)
- AWS Arm system has similar point-to-point latency performance trend, there is a small gap which is due to different resource allocation
- MVAPICH2-X-AWS has higher point-to-point bandwidth in medium message sizes on Arm systems, and higher large message bandwidth with large message sizes (>= 1MB)

4 NODES COLLECTIVE PERFORMANCE

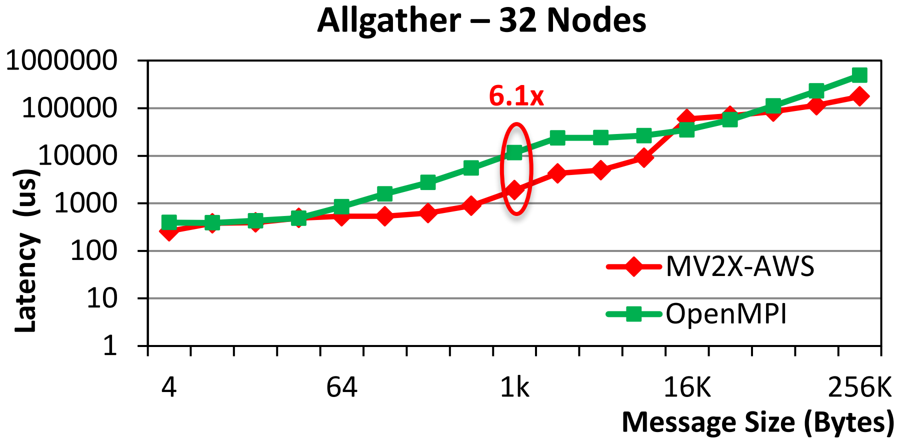


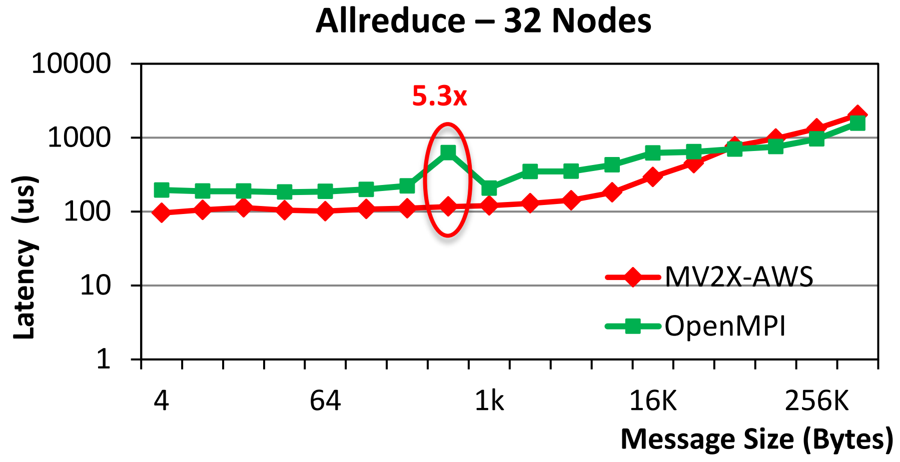


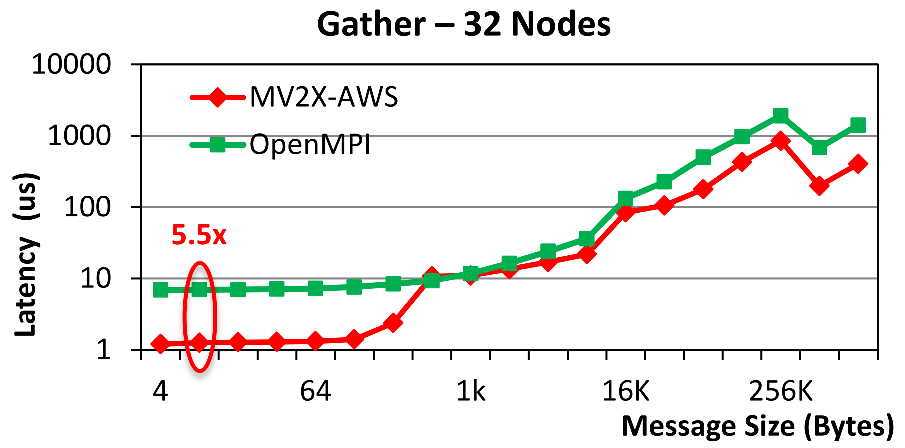


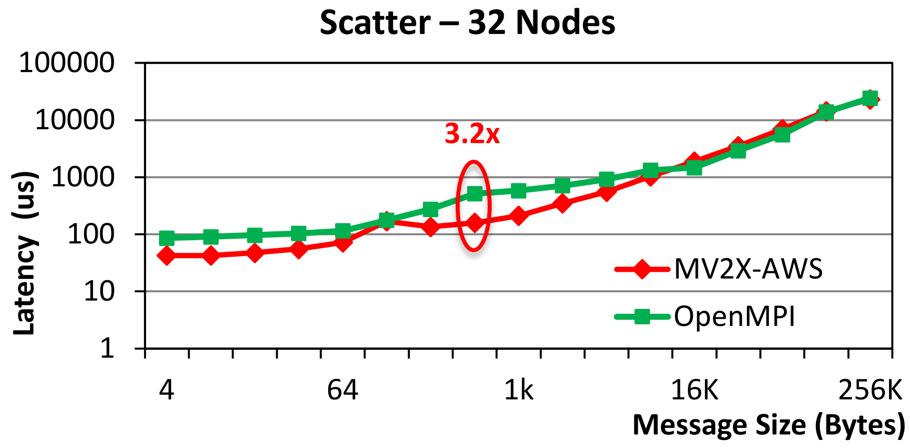


32 NODES COLLECTIVE PERFORMANCE



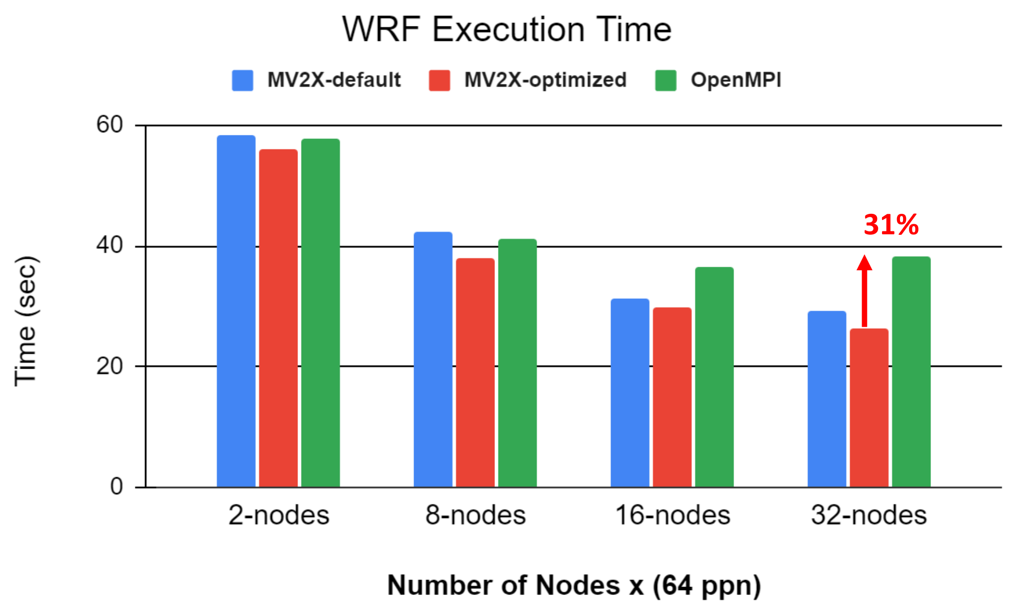


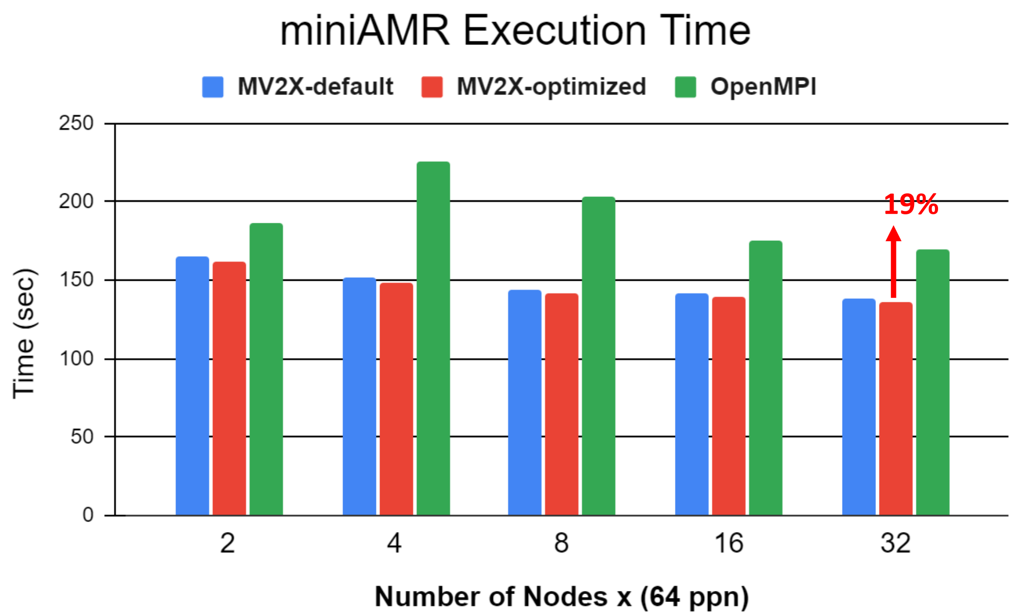




APPLICATION PERFORMANCE

- **Application level performance comparison:**
  - WRF with strong scaling input dataset from 12km resolution case over the Continental U.S. domain
  - miniAMR using default benchmarking input mesh size





RELATED PUBLICATION

Additional Information, Results, and Evaluation can be found in the paper below at APDCM workshop of IPDPS22:

Arm meets Cloud: A Case Study of MPI Library Performance on AWS Arm-based HPC Cloud with Elastic Fabric Adapter. S. Xu, A. Shafi, H. Subramoni, D. Panda. 24th Workshop on Advances in Parallel and Distributed Computational Models (APDCM 22), May 2022.

OUTLINE

- Introduction
- MPI Optimization
- Performance Evaluation
  - Experimental Setups
  - Micro-benchmark level Performance
  - Application level Performance
- **Conclusion & Future Plans**

CONCLUSION AND FUTURE PLANS

- Arm-based Cloud Systems has become a competitive option for HPC application users with compute-intensive workloads

- Performance optimization for MPI libraries leads to significant improvements as well as traditional HPC systems with x86 CPU

- Future Plans:
  - Further performance optimization on coming Graviton Gen3 System on AWS
  - Similar performance optimization for MVAPICH2 on other HPC cloud systems
  - Performance optimization for Arm-based GPU systems on AWS or other cloud systems

THANK YOU!

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 Big Data Project
http://hibd.cse.ohio-state.edu/

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