Achieving High Performance on AWS HPC Cloud using MVAPICH2-AWS

Presenter: Dhabaleswar K (DK) Panda

The Ohio State University
panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

11/19/19

Agenda

- **Overview of the MVAPICH2 Project**
- Overview of Amazon Elastic Fabric Adapter (EFA)
- Designing MVAPICH2 for EFA
- Performance Results
- Software Release and Future Plans

aws

Overview of the MVAPICH2 MPI Library Project

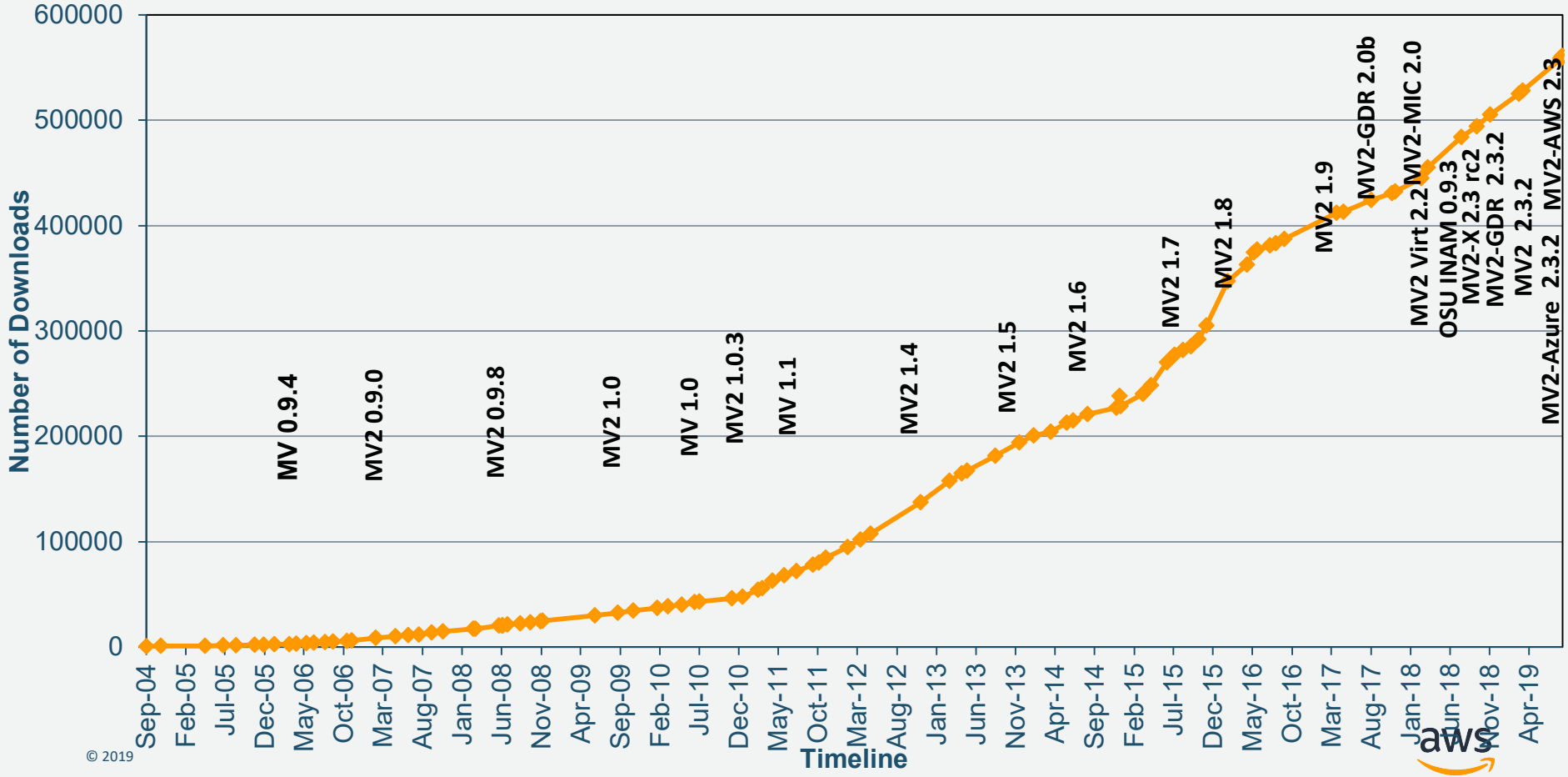
High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)

- 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 (June '19 ranking)
  - 3rd, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China
  - 5th, 448, 448 cores (Frontera) at TACC
  - 8th, 391,680 cores (ABCI) in Japan
  - 15th, 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

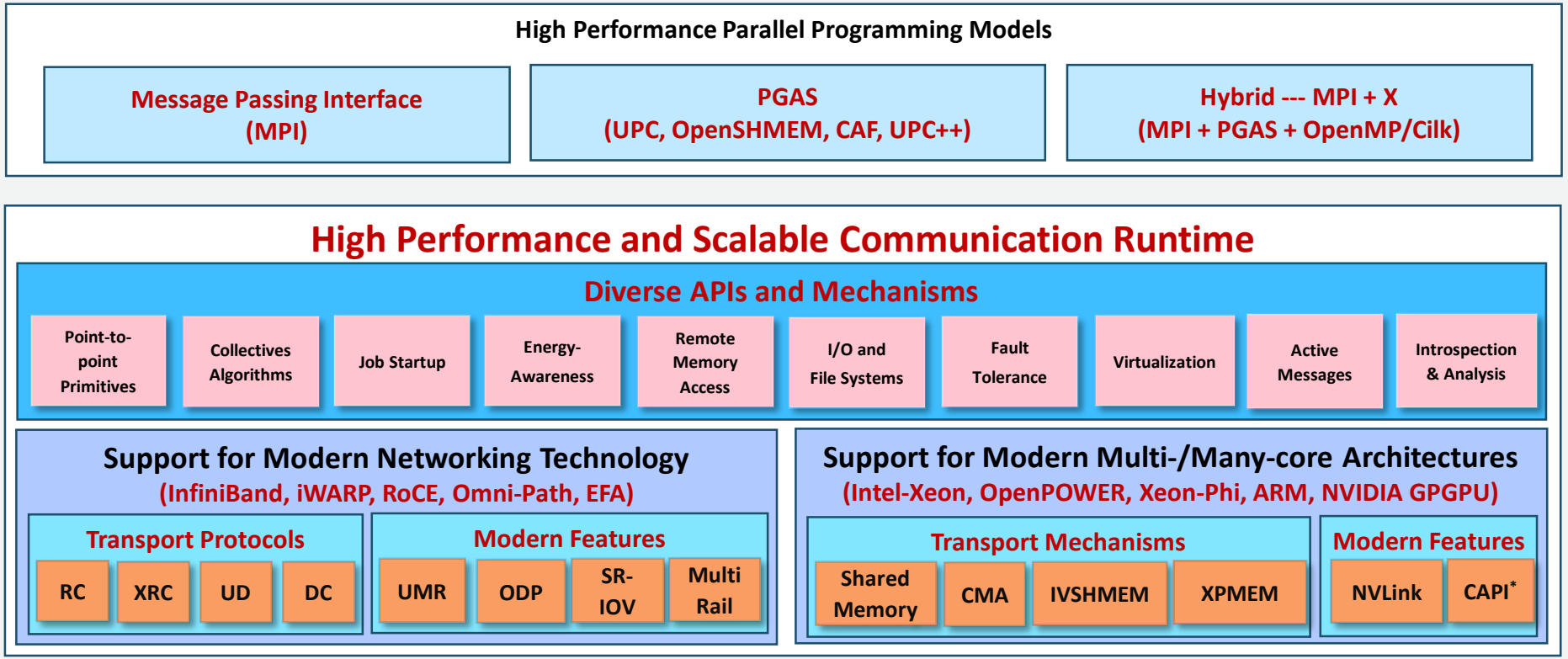
Empowering Top500 systems for over a decade

Partner in the 5th ranked TACC Frontera System

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family



aws

Agenda

- Overview of the MVAPICH2 Project
- **Overview of Amazon Elastic Fabric Adapter (EFA)**
- Designing MVAPICH2 for EFA
- Performance Results
- Software Release and Future Plans

aws

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)

Comparison with IB Transport Types and Trade-offs

Attribute		Reliable Connection	Reliable Datagram	Dynamic Connected	Scalable Reliable Datagram (SRD)	Unreliable Connection	Unreliable Datagram	Raw Datagram
Scalability (M processes, N nodes)		M^2N QPs per HCA	M QPs per HCA	M QPs per HCA	M QPs per HCA	M^2N QPs per HCA	M QPs per HCA	1 QP per HCA
Reliability	**Corrupt data detected**	Yes						
	Data Delivery Guarantee	Data delivered exactly once				No guarantees		
	Data Order Guarantees	Per connection	One source to multiple destinations	Per connection	No	Unordered, duplicate data detected	No	No
	Data Loss Detected	Yes					No	No
	Error Recovery	Errors (retransmissions, alternate path, etc.) handled by transport layer. Client only involved in handling fatal errors (links broken, protection violation, etc.)				Errors are reported to responder	None	None

aws

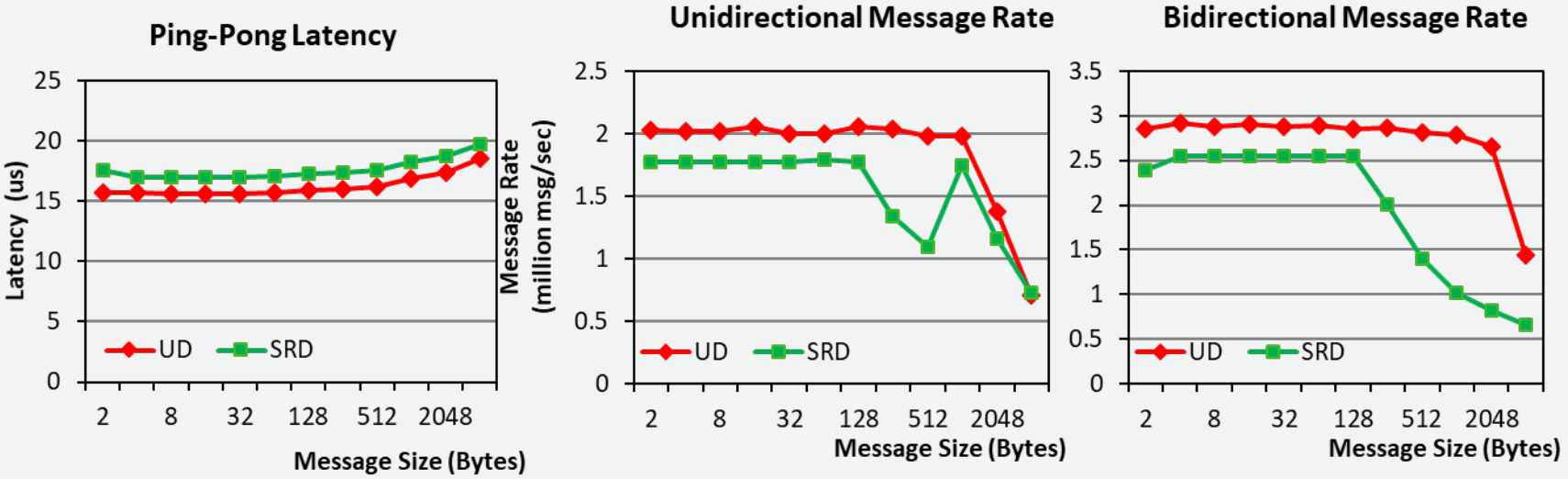
Scalable Reliable Datagrams (SRD): Features & Limitations

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

aws

Verbs level evaluation of EFA performance



- SRD adds 8-10% overhead compared to UD
- Due to hardware based acks used for reliability

- Instance type: c5n.18xlarge
- CPU: Intel Xeon Platinum 8124M @ 3.00GHz

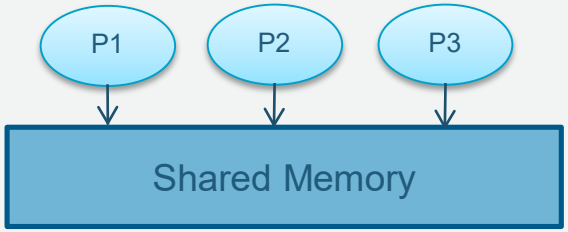
aws

Agenda

- Overview of the MVAPICH2 Project
- Overview of Amazon Elastic Fabric Adapter (EFA)
- **Designing MVAPICH2 for EFA**
- Performance Results
- Software Release and Future Plans

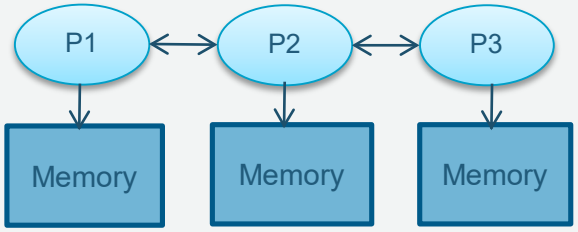
aws

Designing MPI libraries for EFA



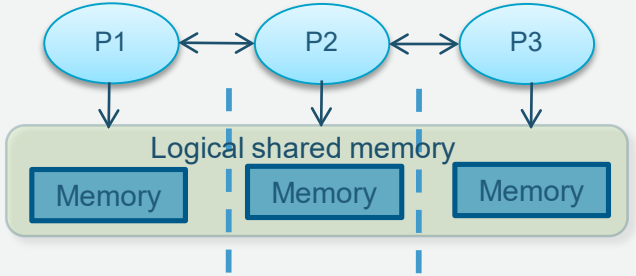
Shared Memory Model

SHMEM, DSM



Distributed Memory Model

MPI (Message Passing Interface)



Partitioned Global Address Space (PGAS)

OpenSHMEM, UPC, UPC++, CAF …

- MPI offer various communication primitives
  - Point-to-point, Collective, Remote Memory Access
  - Provides strict guarantees about reliability and ordering
  - Allows message sizes much larger than allowed by the network
- How to address these semantic mismatches between the network and programming model in a scalable and high-performance manner?

aws

Handled three Major Challenges

- Reliable and in-order delivery
- Zero-copy transmission of large messages
- Handling out-of-order packets for zero-copy transfers

S. Chakraborty, S. Xu, H. Subramoni and D. K. Panda, Designing Scalable and High-Performance MPI Libraries on Amazon Elastic Adapter, Hot Interconnect, 2019

Agenda

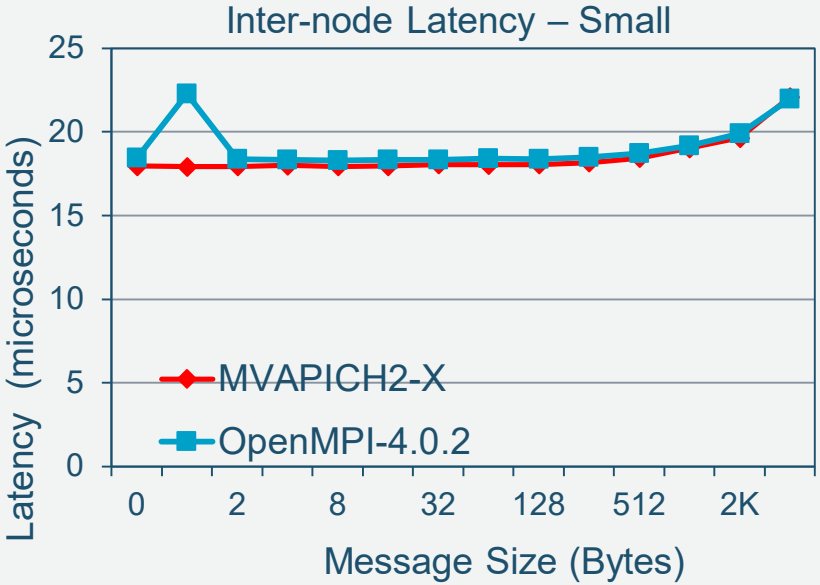
- Overview of the MVAPICH2 Project
- Overview of Amazon Elastic Fabric Adapter (EFA)
- Designing MVAPICH2 for EFA
- **Performance Results**
- Software Release and Future Plans

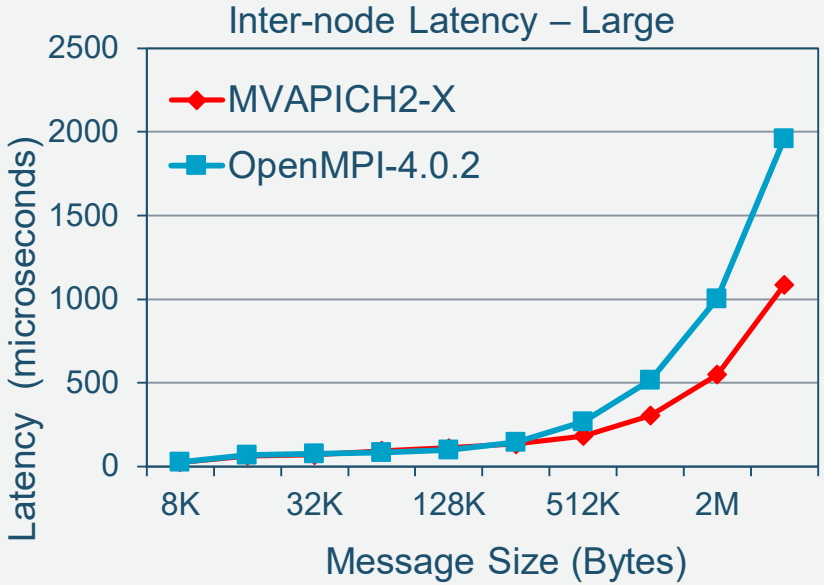
aws

Experimental 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: Latest MVAPICH2-X + SRD support
- OpenMPI version: Open MPI v4.0.2 with libfabric 1.8
- OMB version: OSU Micro-Benchmarks 5.6.2

Point-to-Point Performance

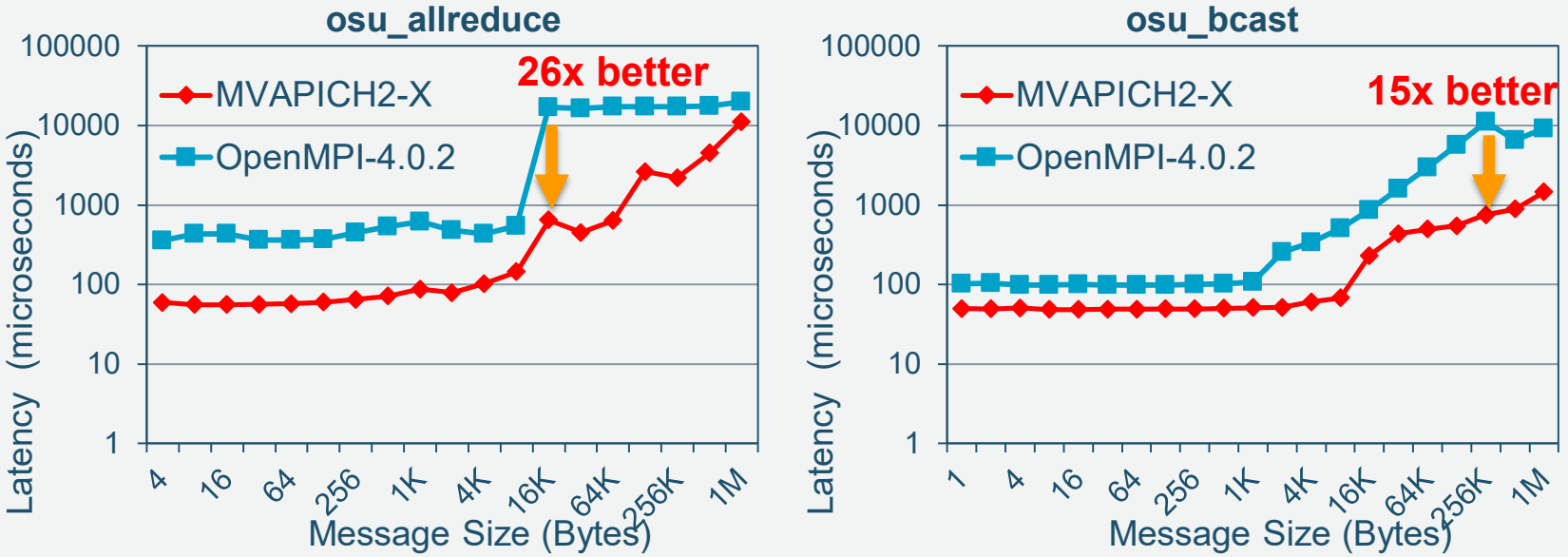




- Up to 1.8x better than OpenMPI on large message sizes

aws

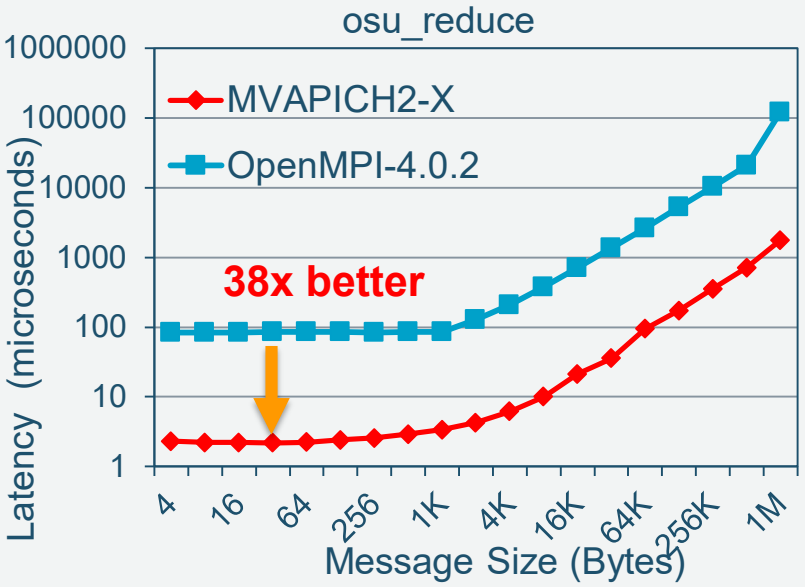
Collectives (4-Node-128-Processes)

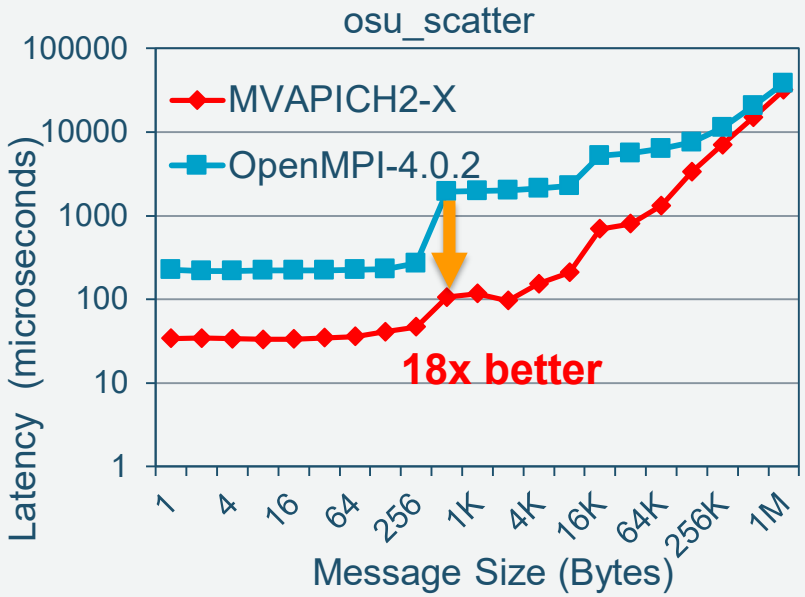


- Up to 26x lower all-reduce latency with MVAPICH2-X than with OpenMPI
- Up to 15x lower bcast latency with MVAPICH2-X than with OpenMPI

aws

Collectives (4-Node-128-Processes) Cont'd

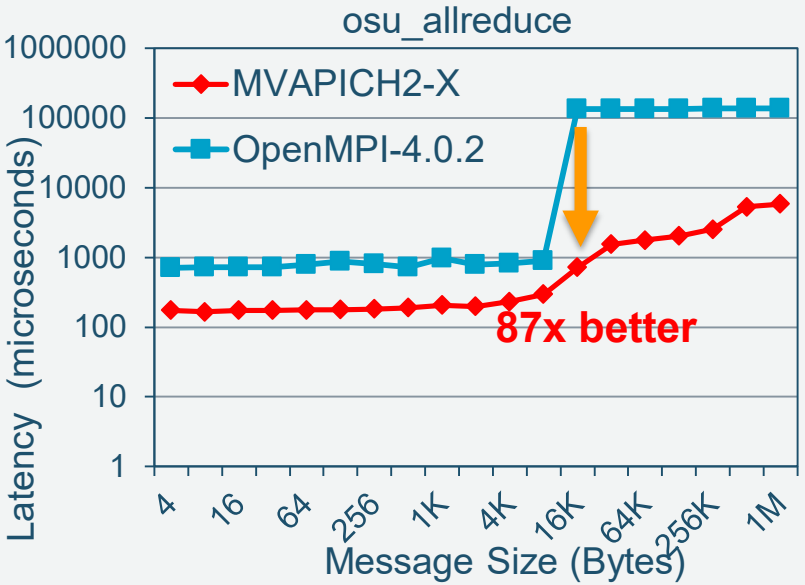


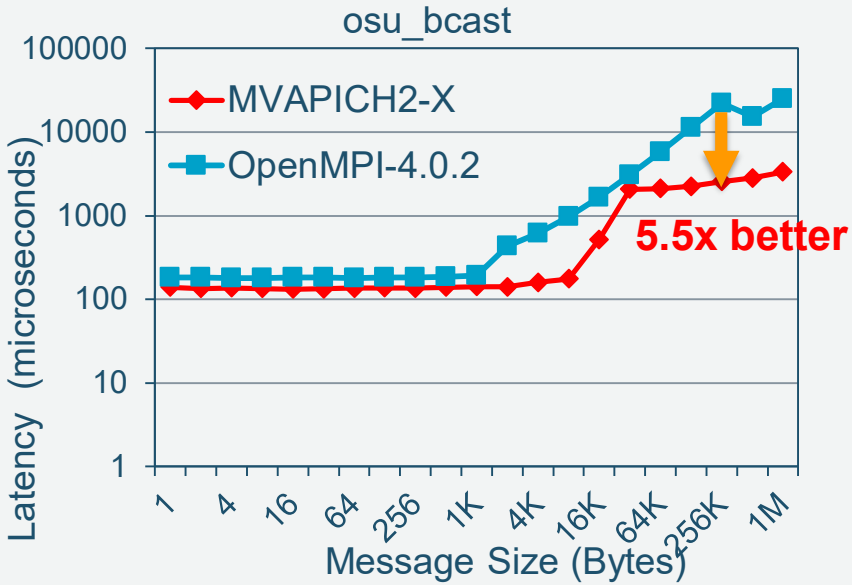


- Up to 38x lower reduce latency with MVAPICH2-X than with OpenMPI
- Up to 18x lower scatter latency with MVAPICH2-X than with OpenMPI

aws

Collectives (16-Node-256-Processes)

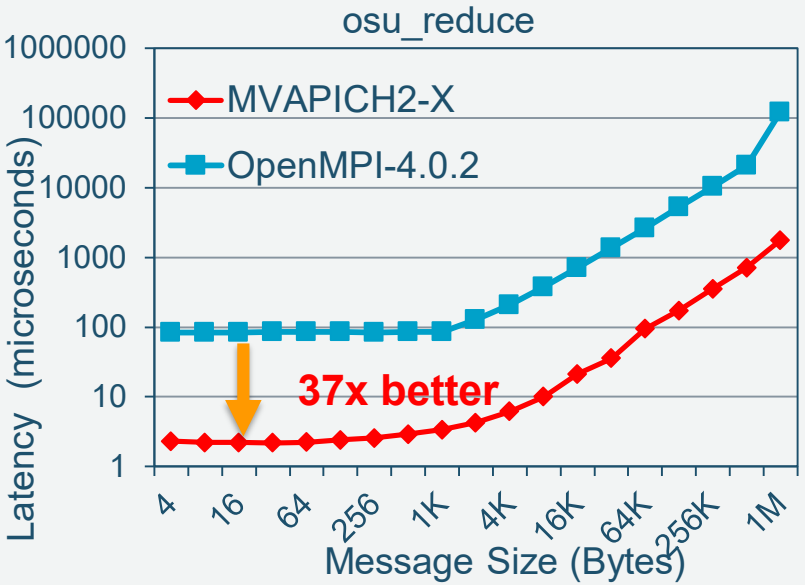


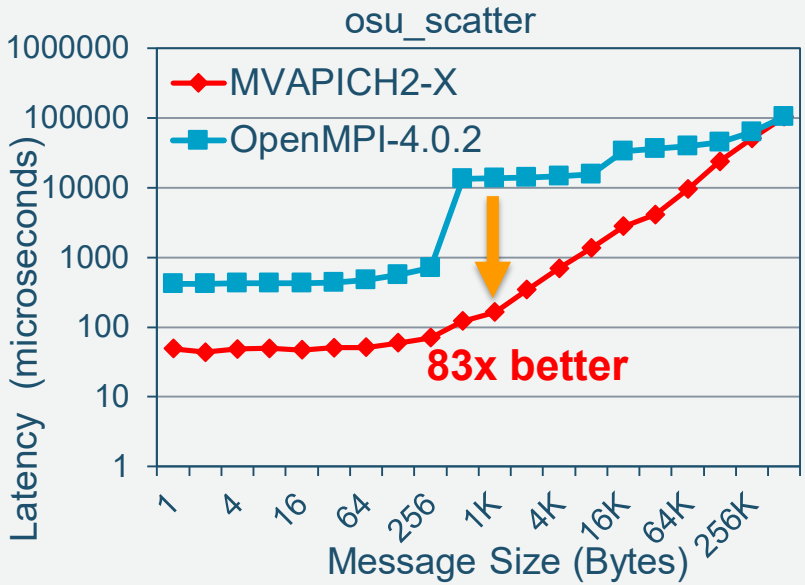


- Up to 87x lower allreduce latency with MVAPICH2-X than with OpenMPI
- Up to 5.5x lower bcast latency with MVAPICH2-X than with OpenMPI

aws

Collectives (16-Node-256-Processes) Cont'd

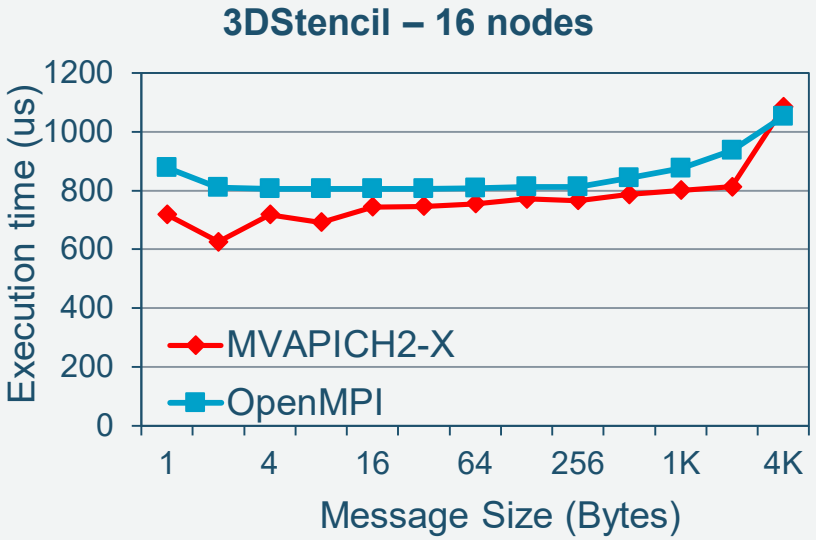


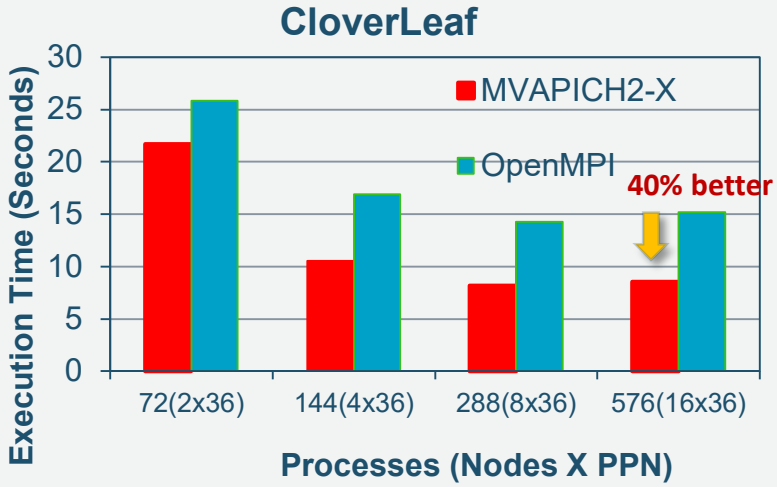


- Up to 37x lower reduce latency with MVAPICH2-X than with OpenMPI
- Up to 83x lower scatter latency with MVAPICH2-X than with OpenMPI

aws

Application Performance





- Up to 23% better performance with 3D-stencil on 16 nodes
- Up to 10% performance improvement for Cloverleaf on 16 nodes

S. Chakraborty, S. Xu, H. Subramoni and D. K. Panda, Designing Scalable and High-Performance MPI Libraries on Amazon Elastic Adapter, Hot Interconnect, 2019

aws

Agenda

- Overview of the MVAPICH2 Project
- Overview of Amazon Elastic Fabric Adapter (EFA)
- Designing MVAPICH2 for EFA
- Performance Results
- **Software Release and Future Plans**

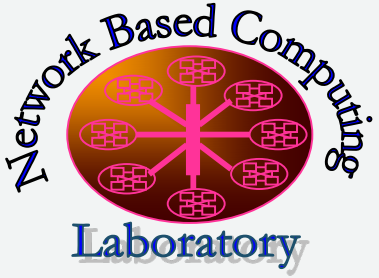
aws

Software Release and Future Plans

- MVAPICH2-X for AWS 2.3 released on 04/12/2019
  - Includes support for SRD and XPMEM based transports
  - Available for download from http://mvapich.cse.ohio-state.edu/downloads/
  - Detailed User Guide: http://mvapich.cse.ohio-state.edu/userguide/mv2x-aws/
- Working on
  - Additional optimizations and tuning, a new version will be released soon
  - Making it available it in an integrated manner in the AWS portal
  - Making it available through AWS Market Place
- Commercial Support available for End-Users, ISVs, and Organizations
  - Through X-ScaleSolutions (http://x-scalesolutions.com)

aws

Thank You!



Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

Follow us on

https://twitter.com/mvapich

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

aws