The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC‘22)

by

Hari Subramoni
The Ohio State University
E-mail: subramon@cse.ohio-state.edu

https://web.cse.ohio-state.edu/~subramoni.1/
History of MVAPICH

• A long time ago, in a galaxy far, far away…. (actually 22 years ago), there existed...
• MPICH
  – High performance and widely portable implementation of MPI standard
  – From ANL
• MVICH
  – Implementation of MPICH ADI-2 for VIA
  – VIA – Virtual Interface Architecture (precursor to InfiniBand)
  – From LBL
• VAPI
  – Verbs level API
  – Initial InfiniBand API from IB Vendors (older version of OFED/IB verbs)

MPICH + MVICH + VAPI = MVAPICH
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 Slingshot10/11, Broadcom, Cornelis Networks OPX

• 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 3,275 organizations in 90 countries

• More than 1.64 Million downloads from the OSU site directly

• Empowering many TOP500 clusters (Nov ‘22 ranking)
  – 7th, 10,649,600-core (Sunway TaihuLight) at NSC in Wuxi, China
  – 19th, 448,448 cores (Frontera) at TACC
  – 34th, 288,288 cores (Lassen) at LLNL
  – 46th, 570,020 cores (Nurion) in South Korea

• Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)

• Partner in the 16th ranked TACC Frontera system

• Empowering Top500 systems for more than 20 years
MVAPICH2 Release Timeline and Downloads
Architecture of MVAPICH2 Software Family for HPC and DL/ML

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
  - SRD
  - UD
  - DC

- Modern Features
  - UMR
  - ODP
  - SR-IOV
  - Multi Rail

Support for Modern Multi-/Many-core Architectures

(Intel-Xeon, OpenPOWER, Xeon-Phi, ARM, NVIDIA/AMD GPGPU)

- Transport Mechanisms
  - Shared Memory
  - CMA
  - IVSHMEM
  - XPMEM

- Modern Features
  - Optane*
  - NVLink
  - CAPI* (* Upcoming)
# MVAPICH2 Software Family

<table>
<thead>
<tr>
<th><strong>High-Performance Parallel Programming Libraries</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH2</td>
<td>Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE</td>
</tr>
<tr>
<td>MVAPICH2-X</td>
<td>Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime</td>
</tr>
<tr>
<td>MVAPICH2-GDR</td>
<td>Optimized MPI for clusters with NVIDIA GPUs</td>
</tr>
<tr>
<td>MVAPICH2-Plus</td>
<td>Advanced MPI with unified MVAPICH2-GDR and MVAPICH2-X features for HPC, DL, ML, Big Data and Data Science applications</td>
</tr>
<tr>
<td>MVAPICH2-J</td>
<td>Java bindings for MVAPICH2 family of libraries</td>
</tr>
</tbody>
</table>

## Microbenchmarks

| **OMB**                                           | Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs |

## Tools

<table>
<thead>
<tr>
<th><strong>OSU INAM</strong></th>
<th>Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OEMT</strong></td>
<td>Utility to measure the energy consumption of MPI applications</td>
</tr>
</tbody>
</table>
MVAPICH2-3.0a Point-to-Point

Latency on Slingshot 11

Latency on OPX

Latency on IB

Bandwidth on Slingshot 11

Bandwidth on OPX

Bandwidth on IB
MVAPICH2-X – Advanced MPI + PGAS + Tools

**MPI_Allreduce using SHARP on Frontera** (1ppn, 7,861 nodes)

**MPI_Barrier using SHARP on Frontera** (1ppn, 7,861 nodes)

**Total Execution Time, BF-2 (osu_ibcast)**

**Total Execution Time, BF-2 (osu_allgather)**

**Impact of Transport Protocol Selection**

**P3DFFT using BlueField-2 DPU on HPCAC**

Network Based Computing Laboratory

MPICH BoF (SC’22)
MVAPICH2-GDR – Optimized MPI for clusters with NVIDIA and AMD GPUs

Best Performance for GPU-based Transfers

TensorFlow Training with MVAPICH2-GDR on Summit

Enhanced Alltoall on DGX2-A100

MVAPICH2-GDR on Slingshot-10 and AMD GPUs

OLCF Spock Cluster – ROCm 5.0 + MI100 GPUs (4 Nodes 4 PPN – 16 GPUs)

Allreduce 16 GPUs (4 Nodes, 4 GPN)

“On-the-fly” Compression Support (AWP-ODC Earthquake Sim App)

“On-the-fly” Compression Support (DASK for Data Science)

Network Based Computing Laboratory

MPICH BoF (SC’22)
MVAPICH-PLUS – NVIDIA GPU Performance + IB

Intra-Node Point-to-Point

Latency

Bandwidth

Bi-Directional Bandwidth

Inter-Node Point-to-Point

Latency

Bandwidth

Bi-Directional Bandwidth

MPI Collectives – 16 GPUs

MPI_Bcast

MPI_Allreduce

NVIDIA A100 GPUs, InfiniBand Networking, and CUDA 11.5
MVAPICH-PLUS – AMD GPU Performance + IB

Intra-Node Point-to-Point

Latency

Bandwidth

Bi-Directional Bandwidth

Inter-Node Point-to-Point

Latency

Bandwidth

Bi-Directional Bandwidth

AMD MI-100 GPUs, InfiniBand Networking, and ROCm 5.1.3

MPI_Allreduce – 16 GPUs

Small Message Allreduce

Large Message Allreduce
MVAPICH-PLUS – GPU Performance + Slingshot-11

Inter-Node Point-to-Point

**Small Message Latency**

- MVAPICH-PLUS
  - Latency (us)
  - Message Size (Bytes)

**Large Message Latency**

- MVAPICH-PLUS
  - Latency (us)
  - Message Size (Bytes)

**Bandwidth**

- MVAPICH-PLUS
  - Bandwidth (MB/s)
  - Message Size (Bytes)

**Bi-Directional Bandwidth**

- MVAPICH-PLUS
  - Bandwidth (MB/s)
  - Message Size (Bytes)

MPI Collectives – 8 GPUs

**Small Message Broadcast**

- MVAPICH-PLUS
  - Latency (us)
  - Message Size (Bytes)

**Large Message Broadcast**

- MVAPICH-PLUS
  - Latency (us)
  - Message Size (Bytes)

**Small Message Allreduce**

- MVAPICH-PLUS
  - Latency (us)
  - Message Size (Bytes)

**Large Message Allreduce**

- MVAPICH-PLUS
  - Latency (us)
  - Message Size (Bytes)

AMD MI250-X GPUs, Slingshot-11 Networking, ROCm-5.3.0
**MVAPICH2 Java Bindings (MVAPICH2-J)**

- **MVAPICH2-J** is an effort to produce Java bindings for the MVAPICH2 library
- **Features**
  - Provides Java bindings to the MVAPICH2 family of libraries
  - Support for communication of basic Java datatypes and Java new I/O (NIO) package direct ByteBuffers
  - Support for blocking and non-blocking point-point communication protocols
  - Support for blocking collective and strided collective communication protocols
- **Results** *(Performance is evaluated against Open MPI’s Java bindings)*
  - **Broadcast Performance**
    - For both buffer and Java arrays, MVAPICH2-J outperforms by 6.2x and 2.2x on average, respectively
  - **AllReduce Performance**
    - For both buffer and Java arrays, MVAPICH2-J outperforms by 2.8x and 1.6x on average, respectively

*Layered Architecture of the Java Bindings for the MVAPICH2 Library*
MVAPICH2-X Advanced Support for HPC-Clouds

Performance on Amazon EFA
WRF 3.6 Execution time

- Master-x
- Ompi

- 2-nodes
- 8-nodes
- 16-nodes
- 32-nodes

Instance type: c6gn.16xlarge
CPU: Amazon Graviton 2 @ 2.50GHz (64 cores per node)
MVAPICH2 version: MVAPICH2-X-AWS-2.3.7 (aarch64)
OpenMPI version: Open MPI v4.1.0 with libfabric 1.13.2

Performance on Oracle HPC Shapes
Bcast – 36-ppn

- MVAPICH2-X
- HPCx

Performance of WRF on Microsoft Azure
WRF 3.6 Execution time

- MVAPICH2
- MVAPICH2-X+XPMEM

VM type: HBv2
CPU: AMD EPYC 7V12 @ 2.45GHz
MVAPICH2 version: MVAPICH2-Azure 2.3.3
MVAPICH2-X version: MVAPICH2-X (2.3rc3)

- Releases
  - MVAPICH2-X-AWS 2.3.7
  - MVAPICH2-Azure 2.3.6
  - Integrated Azure CentOS HPC Images:
    [https://github.com/Azure/azhpc-images/releases/tag/centos-hpc-20220112](https://github.com/Azure/azhpc-images/releases/tag/centos-hpc-20220112)
MVAPICH2 – Future Roadmap and Plans for Exascale

- Making CH4 channel default
  - Early 2023
- Performance and Memory scalability toward 1M-10M cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF ...)
  - MPI + Task*
- Enhanced Optimization for GPUs and FPGAs*
- Taking advantage of advanced features of Mellanox InfiniBand
  - Tag Matching*
  - Adapter Memory*
- Enhanced communication schemes for upcoming architectures
  - NVLINK*
  - CAPI*
  - Bluefield2*
- Extended topology-aware collectives
- Extended Energy-aware designs and Virtualization Support
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended FT support
- Support for * features will be available in future MVAPICH2 Releases
Acknowledgments to all the Heroes (Past/Current Students and Staffs)

**Current Students (Graduate)**
- N. Alnaasan (Ph.D.)
- Q. Anthony (Ph.D.)
- C.-C. Chun (Ph.D.)
- N. Contini (Ph.D.)
- A. Jain (Ph.D.)
- K. S. Khorassani (Ph.D.)
- A. H. Tu (Ph.D.)
- H. Ahn (Ph.D.)
- G. Kuncham (Ph.D.)
- R. Vaidya (Ph.D.)
- J. Yao (Ph.D.)
- M. Han (M.S.)
- A. Guptha (M.S.)
- T. Gangadharpappa (M.S.)
- K. Gopalakrishnan (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Potluri (Ph.D.)
- K. Raj (M.S.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Sarkauskas (B.S. and M.S.)
- N. Senthil Kumar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)
- J. Zhang (Ph.D.)
- M. Abduljabbar
- A. Shafi

**Past Students**
- A. Awan (Ph.D.)
- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- M. Bayatpour (Ph.D.)
- R. Biswas (M.S.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- S. Chakraborthy (Ph.D.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- C.-H. Chu (Ph.D.)
- D. Banerjee
- X. Besseron
- M. S. Ghazimeersaeed
- T. Gangadharpappa
- K. Gopalakrishnan
- J. Hashmi
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- M. Kedia (M.S.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- H.-W. Jin
- E. Mancini
- K. Manian
- M. Luo
- S. Marcarelli
- A. Ruhela
- J. Vienne
- H. Wang

**Past Post-Docs**
- K. Khorassani (Ph.D.)
- P. Kousha (Ph.D.)
- B. Michalowicz (Ph.D.)
- Q. Zhou (Ph.D.)
- K. Al Attar (M.S.)
- L. Xu (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Potluri (Ph.D.)
- K. Raj (M.S.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Senthil Kumar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)
- J. Zhang (Ph.D.)
- M. Abduljabbar
- A. Shafi

**Current Students (Undergrads)**
- V. Shah
- T. Chen
- B. Seeds
- N. Pavuk
- N. Shineman
- M. Lieber

**Past Research Scientists**
- K. Hamidouche
- S. Sur
- X. Lu

**Past Senior Research Associate**
- J. Hashmi

**Past Programmers**
- A. Reifsteck
- D. Bureddy
- J. Perkins

**Current Research Scientists**
- M. Abduljabbar
- A. Shafi

**Current Faculty**
- H. Subramoni

**Current Software Engineers**
- B. Seeds
- N. Pavuk
- N. Shineman
- M. Lieber

**Current Research Specialist**
- R. Motlagh

**Past Research Scientists**
- K. Hamidouche
- S. Sur
- X. Lu

**Past Senior Research Associate**
- J. Hashmi

**Past Programmers**
- A. Reifsteck
- D. Bureddy
- J. Perkins

**Past Research Specialist**
- M. Arnold
- J. Smith
Join us for Multiple Events at SC ‘22

• Presentations at OSU and X-Scale Booth (#4035)
  – 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/904/talks/
Thank You!

subramon@cse.ohio-state.edu

Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

Follow us on

https://twitter.com/mvapich

MVAPICH

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

HiBD

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

HiDL

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