The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC ‘21)

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), and AWS EFA
• 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 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 3,200 organizations in 89 countries
• More than 1.52 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
Architecture of MVAPICH2 Software Family for HPC, DL/ML, and Data Science

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 GPGPU)

- Transport Mechanisms: Shared Memory, CMA, IPC, XPMEM
- Modern Features: BlueField2, NVLink, CAPI*

* Upcoming
# MVAPICH2 Software Family

## High-Performance Parallel Programming Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Description</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-Virt</td>
<td>High-performance and scalable MPI for hypervisor and container based HPC cloud</td>
</tr>
<tr>
<td>MVAPICH2-EA</td>
<td>Energy aware and High-performance MPI</td>
</tr>
<tr>
<td>MVAPICH2-MIC</td>
<td>Optimized MPI for clusters with Intel KNC</td>
</tr>
</tbody>
</table>

## Microbenchmarks

<table>
<thead>
<tr>
<th>Microbenchmark</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMB</td>
<td>Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs</td>
</tr>
</tbody>
</table>

## Tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSU INAM</td>
<td>Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration</td>
</tr>
<tr>
<td>OEMT</td>
<td>Utility to measure the energy consumption of MPI applications</td>
</tr>
</tbody>
</table>
Network Based Computing Laboratory

MPICH BoF (SC’21)

MVAPICH2 – Basic MPI

**MPI_Init on Frontera (Small Scale)**

- **MVAPICH2 2.3.4**
- **Intel MPI 2020**

**MPI_Init on Frontera (Large Scale)**

- **MVAPICH2 2.3.4**
- **Intel MPI 2020**

**Intra-node Latency on OpenPOWER**

- **MVAPICH2 2.3.5**
- **SpectrumMPI-10.3.1.00**

**Allreduce Latency on OpenPOWER**

- 1 Node 40 PPN

**MPI_Bcast using Multicast on Frontera**

- 2048 Nodes

**MPI_Bcast using Multicast on Frontera**

- 32 Bytes
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)

**Impact of Transport Protocol Selection**

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

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

**P3DFFT using BlueField-2 DPU on HPCAC**
MVAPICHER2-GDR – Optimized MPI for clusters with NVIDIA and AMD GPUs

Best Performance for GPU-based Transfers

TensorFlow Training with MVAPICHER2-GDR on Summit

Enhanced Alltoall on DGX2-A100

Best Performance for GPU-based Transfers

MVAPICHER2-GDR (NO-GDR) and MVAPICHER2-GDR-2.3.5

Latency (us) vs Message Size (Bytes)

1.85us latency at 10x speedup

Time/epoch = 3 seconds
Total Time (90 epochs) = 3 x 90 = 270 sec = 4.5 minutes!

MVAPICHER2-GDR 2.3.5

ROCm Support for AMD GPUs (Available with MVAPICHER2-GDR 2.3.6)

LLNL Corona Cluster - ROCm-4.3.0 (mi50 AMD GPUs)

Allreduce 128 GPUs (16 Nodes, 8 GPN)

Broadcast 128 GPUs (16 Nodes, 8 GPN)

"On-the-fly" Compression Support

AWP-ODC Earthquake Sim App

"On-the-fly" Compression Support

DASK for Data Science

Baseline (No compression)
MPC-OPT
ZFP-OPT (rate:16) +35%
ZFP-OPT (rate:8) +18%
Baseline (No compression)
MPC-OPT
ZFP-OPT (rate:16) +1.56x
ZFP-OPT (rate:8)

Network Based Computing Laboratory
MPICH BoF (SC’21)
**MVAPICH2-X Advanced Support for HPC-Clouds**

**Performance on Amazon EFA**
WRF 3.6 Execution Time

- OpenMPI
- Intel MPI
- MVAPICH2-X-AWS

Instance type: c5n.18xlarge
CPU: Intel Xeon Platinum 8124M @ 3.00GHz
MVAPICH2 version: MVAPICH2-X-aws v2.3
OpenMPI version: Open MPI v4.0.3 with libfabric 1.9
IntelMPI version: Intel MPI 2019.7.217

**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
  - MVAPICH2-Azure 2.3.3
  - Integrated Azure CentOS HPC Images:
    https://github.com/Azure/azhpc-images/releases/tag/centos-7.6-hpc-20200417
MVAPICH2 – Future Roadmap and Plans for Exascale

- Update to MPICH 3.4.2 CH3 channel
  - 2021
- Initial support for the CH4 channel
  - Mid 2022
- Making CH4 channel default
  - Late 2022 / 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.)

Current Software Engineers
- B. Seeds
- N. Shineman

Current Students (Undergrads)
- M. Lieber
- L. Xu

Current Research Scientists
- A. Shafi
- H. Subramoni

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.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- J. Hashmi (Ph.D.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- P. Lal (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.)
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- H. Wang
- A. Ruhela
- J. Vienne

Past Post-Docs
- D. Banerjee
- X. Besseron
- M. S. Ghazimeersaede
- A. Ruhela
- J. Vienne
- H. Wang

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

Current Research Scientists
- H. Subramoni
- R. Zhou (Ph.D.)
- K. Al Attar (M.S.)
- N. Sarkauskas (M.S.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Sarkauskas (B.S.)
- N. Senthil Kumar (M.S.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Srivastava (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.)

Current Research Specialist
- R. Motlagh

Current Students (Undergrads)
- M. Lieber
- L. Xu

Past Research Scientists
- A. Shafi
- H. Subramoni

Current Software Engineers
- B. Seeds
- N. Shineman

Current Research Scientists
- A. Shafi
- H. Subramoni

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
Thank You!

subramoni.1@osu.edu
https://web.cse.ohio-state.edu/~subramoni.1/

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/