The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC ‘18)
by

Hari Subramoni
The Ohio State University
E-mail: subramon@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~subramon
Overview of the MVAPICH2 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
  - 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 2,950 organizations in 86 countries
  - More than 505,000 (> 0.5 million) downloads from the OSU site directly
  - Empowering many TOP500 clusters (Nov ‘18 ranking)
    - 3rd ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
    - 14th, 556,104 cores (Oakforest-PACS) in Japan
    - 17th, 367,024 cores (Stampede2) at TACC
    - 27th, 241,108-core (Pleiades) at NASA and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  - Partner in the upcoming TACC Frontera System
  - http://mvapich.cse.ohio-state.edu

- Empowering Top500 systems for over a decade
MVAPICH2 Release Timeline and Downloads

Network Based Computing Laboratory  MPICH BoF (SC’18)
Architecture of MVAPICH2 Software Family

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)

- Transport Protocols
  - RC
  - XRC
  - 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
  - IVSHMEM
  - XPMEM
  - MCDRAM*
  - 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>Benchmark</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>
**MVAPICH2 – Basic MPI**

**Fast Startup on Emerging Many-Cores**

TACC Stampede2

- **Enhanced Intra-node Performance for ARM**

- **Enhanced Intra-node Performance for OpenPOWER**

---

**Advanced Allreduce with SHARP**

- **Major Features and Enhancements in MVAPICH2 2.3 released on 07/23/2018**
  - Improved job startup time for OFA-IB-CH3, PSM-CH3, and PSM2-CH3
  - Enhanced performance of point-to-point operations for CH3-Gen2 (InfiniBand), CH3-PSM, and CH3-PSM2 (Omni-Path) channels
  - Enhanced performance for Allreduce, Reduce_scatter_block, Allgather, Allgatherv, Iallreduce
  - Enhanced process mapping strategies and automatic architecture/network detection
  - Enhanced support for MPI_T PVARs, CVARs and debugging abilities
MVAPICH2-X – Advanced MPI + PGAS + Tools

CMA-Aware MPI_Bcast

Power8 (160 Processes)

Use CMA

Use SHMEM

Message Size

Latency (us)

1

10

100

1000

10000

100000

1M

2M

4M

8M

Performance of P3DFFT Optimized Async Progress •

Shared Address Space (XPMEM)-based Collectives Design

Latency (us)

1

10

100

1000

10000

1M

2M

4M

8M

Message Size

16K 32K 64K 128K 256K 512K 1M 2M 4M

OSU_Allreduce

(Broadwell 256 procs)

OSU_Reduce

(Broadwell 256 procs)

Major Features and Enhancements in MVAPICH2-X 2.3rc1 released on 09/21/2018

– MPI Features
  • Based on MVAPICH2 2.3GA
  • Optimized support for Skylake, ARM, and OpenPOWER architecture

– MPI (Advanced) Features
  • Support for XPMEM-based point-to-point operations and collective operations (Reduce and All-Reduce)
  • Enhanced asynchronous progress designs for progressing non-blocking point-to-point and collective operations

– UPC Features
  • Support Contention Aware Kernel-Assisted MPI collectives

– OpenSHMEM Features
  • Support Contention Aware Kernel-Assisted MPI collectives

Network Based Computing Laboratory

MPICH BoF (SC’18)
MVAPICH2-GDR – Optimized MPI for clusters with NVIDIA GPUs

Best Performance for GPU-based Transfers

Exploiting CUDA-Aware MPI for TensorFlow (Horovod)

GPU-Based MPI_Allreduce

Enhanced Kernel-based Datatype Processing

- Major Features and Enhancements in MVAPICH2-GDR 2.3 released on 11/10/2018
  - Support for CUDA 10, 9.2, 9.0, Volta (V100) GPU, and OpenPOWER with NVLink
  - Efficient Multiple CUDA stream-based IPC communication
  - Enhanced performance of GPU-based point-to-point communication
  - Leverage Linux CMA feature for enhanced host-based communication
  - InfiniBand Multicast based designs for GPU-based broadcast and streaming applications
  - Efficient reduce, allreduce, and broadcast designs for Deep Learning applications
  - Enhanced collective tuning on Xeon, OpenPOWER, and NVIDIA DGX-1 systems
• Virtualization has many benefits
  – Fault-tolerance
  – Job migration
  – Compaction

• Have not been very popular in HPC due to overhead associated with Virtualization
• New SR-IOV (Single Root – IO Virtualization) support available with Mellanox InfiniBand adapters changes the field
• Enhanced MVAPICH2 support for SR-IOV
• MVAPICH2-Virt 2.2 supports:
  – OpenStack, Docker, and singularity
MVAPICH2 – Plans for Exascale

- 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*
- 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
Thank You!

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

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://hidal.cse.ohio-state.edu/