Advanced GPU Support in MVAPICH-Plus

Presentation at OSU Booth  (SC ‘22)

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
- 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 1.63 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (June ‘22 ranking)
  - 7th, 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 19th, 448, 448 cores (Frontera) at TACC
  - 34th, 288,288 cores (Lassen) at LLNL
  - 46th, 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 16th ranked TACC Frontera system
- Empowering Top500 systems for more than 20 years
MVAPICH2 Release Timeline and Downloads
Production Quality Software Design, Development and Release

• Rigorous Q&A procedure before making a release
  – Exhaustive unit testing
  – Various test procedures on diverse range of platforms and interconnects
  – Test 19 different benchmarks and applications including, but not limited to
    • OMB, IMB, MPICH Test Suite, Intel Test Suite, NAS, Scalapak, and SPEC
  – Spend about 18,000 core hours per commit
  – Performance regression and tuning
  – Applications-based evaluation
  – Evaluation on large-scale systems
• All versions (alpha, beta, RC1 and RC2) go through the above testing
One Runtime to Rule them all!

- **Traditional Scientific Computing**
  - Message Passing Interface, PGAS (UPC, OpenSHMEM, CAF, UPC++), Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)

- **Deep Learning/Machine Learning**
- **Data Science**
- **Big Data**

**High Performance Application Domains**

**MVAPICH-Plus**
(Support for all combinations of CPU, Interconnect, Accelerator, DPU)

**Advanced HPC Hardware**

- **Interconnect Technologies**
  - InfiniBand, Omni-Path, Ethernet, Slingshot 10/11, OPX, Broadcom, Rockport

- **Processor Technologies**
  - x86 (Intel/AMD), ARM, OpenPOWER

- **Accelerator Technologies**
  - GPUs (NVIDIA/AMD), FPGAs

- **Network Offload**
  - Datacenter Processing Units, Switch Offload, Network Adapter Offload
One Runtime to Rule them all!

MVAPICH-Plus
(Support for all combinations of CPU, Interconnect, Accelerator, DPU)

Advanced HPC Hardware

<table>
<thead>
<tr>
<th>Interconnect Technologies</th>
<th>Processor Technologies</th>
<th>Accelerator Technologies</th>
<th>Network Offload</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfiniBand, Omni-Path, Ethernet, Slingshot 10/11, OPX, Broadcom, Rockport</td>
<td>x86 (Intel/AMD), ARM, OpenPOWER</td>
<td>GPUs (NVIDIA/AMD), FPGAs</td>
<td>Datacenter Processing Units, Switch Offload, Network Adapter Offload</td>
</tr>
</tbody>
</table>
MVAPICH-Plus

- Released on 11/11/2022
- Based on MVAPICH 3.0
- Advanced MPI with unified MVAPICH2-GDR and MVAPICH2-X features
- Support for NVIDIA and AMD GPUs
- Optimized designs for HPC, DL, ML, Big Data and Data Science applications
- Added support for the ch4:ucx and ch4:ofi devices
- Added support for the Cray Slingshot 11 interconnect over OFI
  - Supports Cray Slingshot 11 network adapters
- Added support for the Cornelis OPX library over OFI
  - Supports Intel Omni-Path adapters
- Added support for the Intel PSM3 library over OFI
  - Supports Intel Columbiaville network adapters
- Added support for IB verbs over UCX
  - Supports IB and RoCE network adapters
Features of OFI and UCX Support

• Support a broad range of interconnects with widely used libraries
  – Configure with `--with-device=ch4:ofi` or `--with-device=ch4:ucx`

• Runtime provider selection via CVARs
  – `MPIR_CVAR_OFI_USE_PROVIDER=<prov>`

• System default, embedded, or custom installation of OFI/UCX
  – Configure with `--with-libfabric=embedded` or `--with-libfabric=<path>`
  – Configure with `--with-ucx=embedded` or `--with-ucx=<path>`

• Enhanced MVAPICH2 collective designs
MPI Level Latency on Slingshot 11

- **2us** inter-node point-to-point latency for small messages

**Interconnect**: Cray HPE Slingshot 11  
**Library**: MVAPICH2 3.0a  
**CPU**: AMD EPYC 7763 (milan) Processor
MPI Level Bandwidth on Slingshot 11

- **23,985 MB/s** uni-directional peak bandwidth
- **42,034 MB/s** bi-directional peak bandwidth

Interconnect: Cray HPE Slingshot 11 (200 Gbps)
Library: MVAPICH2 3.0a
CPU: AMD EPYC 7763 (milan) Processor
MVAPICH2-3.0a+OPX vs MVAPICH2-2.3.7+PSM2 (Early Performance Results)

**OSU_BIBW (2 Nodes, 1 PPN)**

- MVAPICH2-2.3.7-PSM
- MVAPICH2-3.0a-OPX

**OSU_BW (2 Nodes, 1 PPN)**

- MVAPICH2-2.3.7-PSM
- MVAPICH2-3.0a-OPX

**OSU_Latency (2 Nodes, 1 PPN)**

- MVAPICH2-2.3.7-PSM
- MVAPICH2-3.0a-OPX

**System:**
- Intel Xeon Bronze (Skylake) 3106 CPU @ 1.70GHz (4 nodes, 16 cores/node, 8 x 2 sockets) with Omni-Path 100Gbps
At Sender:

MPI_Send(s_devbuf, size, ...);

At Receiver:

MPI_Recv(r_devbuf, size, ...);

High Performance and High Productivity
MVAPICH-PLUS - Point-to-Point on GPU

- NVIDIA A100 GPUs with CUDA version 11.5 and NCCL version 2.14.3
MVAPICH-PLUS - Collective on GPU (8 nodes, 16 GPU)

- NVIDIA A100 GPUs with CUDA version 11.5 and NCCL version 2.14.3
Point-to-Point Inter-Node Performance on AMD GPUs

Latency:

Latency (us)

Message Size (Bytes)

Bandwidth:

Bandwidth (MB/s)

Message Size (Bytes)

Tioga - ROCm-5.3.0 (MI250-X GPUs)
Collective Performance on AMD GPUs

**Bcast:**

- **Tioga - ROCm-5.3.0 (MI250-X GPUs) – 8 GPUs**

- **MVAPICH-PLUS**
  - Latency (us) vs. Message Size (Bytes)
  - Message Size: 4, 8, 16, 32, 64, 128, 256, 512, 1k, 2k, 4k
  - Latencies: 26.51us, 28.20us, 30.31us, 32.52us, 34.73us, 36.94us, 39.15us, 41.36us, 43.57us, 45.78us

**Alltoall:**

- **MVAPICH-PLUS**
  - Latency (us) vs. Message Size (Bytes)
  - Message Size: 8k, 16k, 32k, 64k, 128k, 256k, 512k, 1M
  - Latencies: 85.22us, 149.43us
Collective Performance on AMD GPUs (Cont.)

**Gather:**

![Graph showing latency vs message size for Gather operation using Tioga - ROCm-5.3.0 (MI250-X GPUs) with 8 GPUs.](image1)

**Allgather:**

![Graph showing latency vs message size for Allgather operation using Tioga - ROCm-5.3.0 (MI250-X GPUs) with 8 GPUs.](image2)

**Tioga - ROCm-5.3.0 (MI250-X GPUs) – 8 GPUs**
Collective Performance on AMD GPUs (Cont.)

Reduce:

Reduce:  

MVAPICH-PLUS

Latency (us)

24.65

Message Size (Bytes)

4 8 16 32 64 128 256 512 1K 2K 4K

Allreduce:

Allreduce:

MVAPICH-PLUS

Latency (us)

75.33

Message Size (Bytes)

8K 16K 32K 64K 128K 256K 512K 1M

Tioga - ROCm-5.3.0 (MI250-X GPUs) – 8 GPUs
MVAPICH2-Plus Upcoming Features for HPC and DL

- On-the-fly Compression for All_Gather Collective
- Scalable Distributed Training with Model-/Hybrid Parallelism for out-of-core DNN Models
- Scaling Single-Image Super-Resolution Training
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
Join us for Multiple Events at SC ‘22

- Presentations at OSU and X-Scale Booth (#4305)
  - 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!

panda@cse.ohio-state.edu

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/