Overview of the MVAPICH Project: Latest Status and Future Roadmap

MVAPICH2 User Group (MUG) Meeting

by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda
Drivers of Modern HPC Cluster Architectures

- Multi-core processors are ubiquitous
- InfiniBand very popular in HPC clusters
- Accelerators/Coprocessors becoming common in high-end systems
- Pushing the envelope for Exascale computing

Tianhe – 2 (1)  Titan (2)  Stampede (6)  Tianhe – 1A (10)
Parallel Programming Models Overview

- Programming models provide abstract machine models
- Models can be mapped on different types of systems
  - e.g. Distributed Shared Memory (DSM), MPI within a node, etc.
- PGAS models and Hybrid MPI+PGAS models are gradually receiving importance
Supporting Programming Models for Multi-Petaflop and Exaflop Systems: Challenges

Application Kernels/Applications

Middleware

Programming Models
MPI, PGAS (UPC, Global Arrays, OpenSHMEM), CUDA, OpenACC, Cilk, Hadoop, MapReduce, etc.

Co-Design Opportunities and Challenges across Various Layers

Communication Library or Runtime for Programming Models

Point-to-point Communication (two-sided & one-sided)

Collective Communication

Synchronization & Locks

I/O & File Systems

Fault Tolerance

Networking Technologies (InfiniBand, 10/40GigE, Aries, BlueGene)

Multi/Many-core Architectures

Accelerators (NVIDIA and MIC)
Designing (MPI+X) at Exascale

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Extremely minimum memory footprint

- Hybrid programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, ...)

- Balancing intra-node and inter-node communication for next generation multi-core (128-1024 cores/node)
  - Multiple end-points per node

- Support for efficient multi-threading

- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
  - Power-aware

- Support for MPI-3 RMA Model

- Support for GPGPUs and Accelerators

- Fault-tolerance/resiliency

- QoS support for communication and I/O
MVAPICH2/MVAPICH2-X Software

- [http://mvapich.cse.ohio-state.edu](http://mvapich.cse.ohio-state.edu)

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2012
  - Used by more than 2,200 organizations (HPC Centers, Industry and Universities) in 73 countries

MVAPICH Team Projects

Timeline

MVAPICH
Oct-00

OMB
Nov-04

MVAPICH2
Jan-10

MVAPICH2-GDR
Nov-12

MVAPICH2-X
Aug-14

EOL
MVAPICH/MVAPICH2 Release Timeline and Downloads

- Download counts from MVAPICH2 website
MVAPICH2/MVAPICH2-X Software (Cont’d)

• Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)

• Empowering many TOP500 clusters
  – 7th ranked 519,640-core cluster (Stampede) at TACC
  – 13th, 74,358-core (Tsubame 2.5) at Tokyo Institute of Technology
  – 23rd, 96,192-core (Pleiades) at NASA and many others

• Partner in the U.S. NSF-TACC Stampede System
Strong Procedure for Design, Development and Release

• Research is done for exploring new designs
• Designs are first presented to conference/journal publications
• Best performing designs are incorporated into the codebase
• Rigorous Q&A procedure before making a release
  – Exhaustive unit testing
  – Various test procedures on diverse range of platforms and interconnects
  – Performance tuning
  – Applications-based evaluation
  – Evaluation on large-scale systems
• Even alpha and beta versions go through the above testing
**MVAPICH2 Architecture (Latest Release 2.0)**

- Major Computing Platforms: IA-32, EM64T, Ivybridge, Westmere, Sandybridge, Opteron, Magny, ..

- All Different PCI interfaces

- MVAPICH User Group Meeting 2014
MVAPICH2 2.0 and MVAPICH2-X 2.0

- Released on 06/20/2014

- Major Features and Enhancements
  - Based on MPICH-3.1
  - Extended support for MPI-3 RMA in OFA-IB-CH3, OFA-IWARP-CH3, and OFA-RoCE-CH3 interfaces
  - Multiple optimizations and performance enhancements for RMA in OFA-IB-CH3 channel
  - MPI-3 RMA support for CH3-PSM channel
  - CMA support is now enabled by default
  - Enhanced intra-node SMP performance
  - Dynamic CUDA initialization
  - Support for running on heterogeneous clusters with GPU and non-GPU nodes
  - Multi-rail support for GPU communication and UD-Hybrid channel
  - Improved job-startup performance for large-scale mpirun_rsh jobs
  - Reduced memory footprint
  - Updated compiler wrappers to remove application dependency on network and other extra libraries
  - Updated hwloc to version 1.9

- MVAPICH2-X 2.0GA supports hybrid MPI + PGAS (UPC and OpenSHMEM) programming models.
  - Based on MVAPICH2 2.0GA including MPI-3 features; Compliant with UPC 2.18.0 and OpenSHMEM v1.0f
  - Enhanced optimization of OpenSHMEM collectives and Optimized shm allocate routine
  - Optimized UPC collectives and support for GUPC translator
One-way Latency: MPI over IB with MVAPICH2

![Small Message Latency Graph](image)

![Large Message Latency Graph](image)

- DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch
- FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
Bandwidth: MPI over IB with MVAPICH2

Unidirectional Bandwidth

- DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch
- FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

Bidirectional Bandwidth

- DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch
- FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
MVAPICH2 Two-Sided Intra-Node Performance
(Shared memory and Kernel-based Zero-copy Support (LiMIC and CMA))

Latest MVAPICH2 2.0
Intel Ivy-bridge

Latency

Latency (us)

Message Size (Bytes)

Intra-Socket
Inter-Socket

0.18 us
0.45 us

Bandwidth (Intra-socket)

Bandwidth (Inter-socket)

Bandwidth (MB/s)

Message Size (Bytes)

intra-Socket-CMA
intra-Socket-Shmem
intra-Socket-LiMIC

inter-Socket-CMA
inter-Socket-Shmem
inter-Socket-LiMIC

14,250 MB/s
13,749 MB/s

MVAPICH User Group Meeting 2014
Recent measurements of Inter-node Get/Put Latency and Bandwidth for the Latest MVAPICH2 2.0 with Intel Sandy-bridge and Connect-IB (single-port) have shown the following improvements:

- **Inter-node Get/Put Latency**:
  - Get: 2.04 µs
  - Put: 1.56 µs

- **Inter-node Get/Put Bandwidth**:
  - Get: 6881 Mbytes/sec
  - Put: 6876 Mbytes/sec

- **Intra-socket Get/Put Latency**:
  - Get: 0.08 µs

- **Intra-socket Get/Put Bandwidth**:
  - Get: 15364 Mbytes/sec
  - Put: 14926 Mbytes/sec

These results highlight the performance gains achieved with the latest version of MVAPICH2 and the Intel Sandy-bridge with Connect-IB.
MVAPICH2-X for Hybrid MPI + PGAS Applications

- Unified communication runtime for MPI, UPC, OpenSHMEM available with MVAPICH2-X 1.9 onwards!
  - [http://mvapich.cse.ohio-state.edu](http://mvapich.cse.ohio-state.edu)

- Feature Highlights
  - Supports MPI(+OpenMP), OpenSHMEM, UPC, MPI(+OpenMP) + OpenSHMEM, MPI(+OpenMP) + UPC
  - MPI-3 compliant, OpenSHMEM v1.0 standard compliant, UPC v1.2 standard compliant (with initial support for UPC 1.3)
  - Scalable Inter-node and intra-node communication – point-to-point and collectives
Hybrid (MPI+PGAS) Programming

• Application sub-kernels can be re-written in MPI/PGAS based on communication characteristics

• Benefits:
  – Best of Distributed Computing Model
  – Best of Shared Memory Computing Model

• Exascale Roadmap*:
  – “Hybrid Programming is a practical way to program exascale systems”

Hybrid MPI+OpenSHMEM Graph500 Design

- Performance of Hybrid (MPI+OpenSHMEM) Graph500 Design
  - 8,192 processes
    - 2.4X improvement over MPI-CS0
    - 7.6X improvement over MPI-Simple
  - 16,384 processes
    - 1.5X improvement over MPI-CS0
    - 13X improvement over MPI-Simple

J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models, International Supercomputing Conference (ISC’13), June 2013
J. Jose, K. Kandalla, M. Luo and D. K. Panda, Supporting Hybrid MPI and OpenSHMEM over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP '12), September 2012
MVAPICH2 2.1a and MVAPICH2-X 2.1a

• Will be available later this week or next week

• Features
  – Based on MPICH 3.1.2
  – PMI-2-based startup with SLURM
  – Improved job-startup, etc.
MVAPICHD2-GDR 2.0

• Released on 08/23/2014

• Major Features and Enhancements
  – Based on MVAPICHD2 2.0 (OFA-IB-CH3 interface)
  – High performance RDMA-based inter-node point-to-point communication (GPU-GPU, GPU-Host and Host-GPU) using GPUDirect RDMA and pipelining
  – Support for MPI-3 RMA (one Sided) communication and atomics using GPU buffers with GPUDirect RDMA and pipelining
  – Efficient small message inter-node communication using the new NVIDIA GDRCOPY module
  – Efficient small message inter-node communication using loopback design
  – Multi-rail support for inter-node point-to-point GPU communication
  – High performance intra-node point-to-point communication for multi-GPU adapters/node (GPU-GPU, GPU-Host and Host-GPU) using CUDA IPC and pipelining
  – Automatic communication channel selection for different GPU communication modes (DD, HH and HD) in different configurations (intra-IOH and inter-IOH)
  – Automatic selection and binding of the best GPU/HCA pair in multi-GPU/HCA system configuration
  – Optimized and tuned support for collective communication from GPU buffers
  – Enhanced and Efficient support for datatype processing on GPU buffers including support for vector/h-vector, index/h-index, array and subarray
  – Dynamic CUDA initialization. Support GPU device selection after MPI_Init
  – Support for non-blocking streams in asynchronous CUDA transfers for better overlap
  – Efficient synchronization using CUDA Events for pipelined device data transfers
Performance of MVAPICH2 with GPU-Direct-RDMA: Latency

Small Message Latency

- MV2-GDR2.0
- MV2-GDR2.0b
- MV2 w/o GDR

Message Size (bytes) vs. Latency (us)

MVAPICH2-GDR-2.0
Intel Ivy Bridge (E5-2680 v2) node with 20 cores
NVIDIA Tesla K40c GPU, Mellanox Connect-IB Dual-FDR HCA
CUDA 6.5, Mellanox OFED 2.1 with GPU-Direct-RDMA
Performance of MVAPICH2 with GPU-Direct-RDMA: Bandwidth

**GPU-GPU Internode MPI Uni-Directional Bandwidth**

**Small Message Bandwidth**

- MV2-GDR2.0
- MV2-GDR2.0b
- MV2 w/o GDR

**Message Size (bytes)**

- 1
- 4
- 16
- 64
- 256
- 1K
- 4K

**Bandwidth (MB/s)**

- 0
- 500
- 1000
- 1500
- 2000
- 2500
- 3000

**MVAPICH2-GDR-2.0**

Intel Ivy Bridge (E5-2680 v2) node with 20 cores
NVIDIA Tesla K40c GPU, Mellanox Connect-IB Dual-FDR HCA
CUDA 6.5, Mellanox OFED 2.1 with GPU-Direct-RDMA
Performance of MVAPICH2 with GPU-Direct-RDMA: Bi-Bandwidth

GPU-GPU Internode MPI Bi-directional Bandwidth

Small Message Bi-Bandwidth

- MV2-GDR2.0
- MV2-GDR2.0b
- MV2 w/o GDR

Message Size (bytes)

MVAPICH2-GDR-2.0

Intel Ivy Bridge (E5-2680 v2) node with 20 cores
NVIDIA Tesla K40c GPU, Mellanox Connect-IB Dual-FDR HCA
CUDA 6.5, Mellanox OFED 2.1 with GPU-Direct-RDMA
Performance of MVAPICH2 with GPU-Direct-RDMA: MPI-3 RMA

GPU-GPU Internode MPI Put latency (RMA put operation Device to Device)

MPI-3 RMA provides flexible synchronization and completion primitives

![Small Message Latency Graph]

- MV2-GDR2.0
- MV2-GDR2.0b

MVAPICH2-GDR-2.0
Intel Ivy Bridge (E5-2680 v2) node with 20 cores
NVIDIA Tesla K40c GPU, Mellanox Connect-IB Dual-FDR HCA
CUDA 6.5, Mellanox OFED 2.1 with GPU-Direct-RDMA
Application-Level Evaluation (HOOMD-blue)

HOOMD-blue Strong Scaling

- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- MV2-GDR 2.0 (released on 08/23/14) : try it out !
  - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0
    MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768
    MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1
    MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384

Weak Scalability of HOOMD–Blue with 2K particles/GPU Node

Strong Scalability of HOOMD–Blue with 256K particles

Number of GPU Nodes

Average TPS

Number of GPU Nodes

Average TPS
MPI Applications on MIC Clusters

• MPI (+X) continues to be the predominant programming model in HPC
• Flexibility in launching MPI jobs on clusters with Xeon Phi

Multi-core Centric

Host-only

Offload (/reverse Offload)

Symmetric

Coprocessor-only

Many-core Centric

Xeon

Xeon Phi

Host-only

Offload

Symmetric

Coprocessor-only

Many-core Centric

Offloaded Computation

MPI Program

MPI Program

MPI Program

MPI Program
MVAPICH2-MIC Design for Clusters with IB and MIC

- Offload Mode
- Intranode Communication
  - Coprocessor-only Mode
  - Symmetric Mode
- Internode Communication
  - Coprocessors-only
  - Symmetric Mode
- Multi-MIC Node Configurations
Latest Status on MVAPICH2-MIC

• Running on three major systems
  – Stampede : module swap mvapich2 mvapich2-mic/20130911
  – Blueridge (Virginia Tech) : module swap mvapich2 mvapich2-mic/1.9
  – Beacon (UTK) : module unload intel-mpi; module load mvapich2-mic/1.9

• A new version based on MVAPICH2 2.0 is being worked out
• Will be available in a few weeks
Optimized MPI Collectives for MIC Clusters (Allgather & Alltoall)

32-Node-Allgather (16H + 16 M)
Small Message Latency

- MV2-MIC
- MV2-MIC-Opt

32-Node-Allgather (8H + 8 M)
Large Message Latency

- MV2-MIC
- MV2-MIC-Opt

32-Node-Alltoall (8H + 8 M)
Large Message Latency

- MV2-MIC
- MV2-MIC-Opt

P3DFFT Performance

- Communication
- Computation


MVAPICH User Group Meeting 2014
OSU Micro-Benchmarks (OMB)

• Started in 2004 and continuing steadily
• Allows MPI developers and users to
  – Test and evaluate MPI libraries
• Has a wide-range of benchmarks
  – Two-sided (MPI-1, MPI-2 and MPI-3)
  – One-sided (MPI-2 and MPI-3)
  – RMA (MPI-3)
  – Collectives (MPI-1, MPI-2 and MPI-3)
  – Extensions for GPU-aware communication (CUDA and OpenACC)
  – UPC (Pt-to-Pt)
  – OpenSHMEM (Pt-to-Pt and Collectives)
• Widely-used in the MPI community
OMB (OSU Micro-Benchmarks) 4.4

- Released on 08/23/2014
- Include benchmarks for MPI, OpenSHMEM, and UPC Programming Models
- MPI
  - Support MPI-1, MPI-2 and MPI-3 standards
  - Point-to-Point Benchmarks: Latency, Multi-threaded latency, Multi-pair latency, Multiple bandwidth / Message rate test bandwidth, Bidirectional bandwidth
  - Collective Benchmarks: Collective latency test for MPI_Allgather, MPI_Alltoall, MPI_Allreduce, MPI_Barrier, MPI_Bcast, MPI_Gather, MPI_Reduce, MPI_Reduce_Scatter, MPI_Scatter, and vector collectives
  - One-sided Benchmarks: Put latency, Put bandwidth, Put bidirectional bandwidth, Get latency, Get bandwidth, Accumulate latency, Atomics latency, and Get_accumulate latency
  - CUDA Extensions to evaluate MPI communication performance from/to buffers on NVIDIA GPU devices
  - OpenACC Extensions to evaluate MPI communication performance from/to buffers on NVIDIA GPU devices
  - Support for MPI-3 RMA operations using GPU buffers
- OpenSHMEM
  - Point-to-Point benchmarks: Put latency, Get latency, Put message rate, and Atomics latency
  - Collective benchmarks: Collect latency, FCollect latency, Broadcast Latency, Reduce latency, and Barrier latency
- Unified Parallel C (UPC)
  - Point-to-Point benchmarks: Put latency, Get latency
  - Collective benchmarks: Barrier latency, Exchange latency, Gatherall latency, Gather latency, Reduce latency, and Scatter latency
Virtualization Support with MVAPICH2 using SR-IOV
Intra-node Inter-VM Point-to-Point Latency and Bandwidth

- 1 VM per Core
- MVAPICH2-SR-IOV-IB brings only 3-7% (latency) and 3-8% (BW) overheads, compared to MAPICH2 over Native InfiniBand Verbs (MVAPICH2-Native-IB)
Performance Evaluations with NAS and Graph500

- 8 VMs across 4 nodes, 1 VM per Socket, 64 cores totally
- MVAPICH2-SR-IOV-IB brings 3-7% and 3-9% overheads for NAS Benchmarks and Graph500, respectively, compared to MVAPICH2-Native-IB
Performance Evaluation with LAMMPS

- 8 VMs across 4 nodes, 1 VM per Socket, 64 cores totally
- MVAPICH2-SR-IOV-IB brings 7% and 9% overheads for LJ and CHAIN in LAMMPS, respectively, compared to MVAPICH2-Native-IB
MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 500K-1M cores
  - Dynamically Connected Transport (DCT) service with Connect-IB
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF ...)
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of Collective Offload framework
  - Including support for non-blocking collectives (MPI 3.0)
  - Core-Direct support
- Extended RMA support (as in MPI 3.0)
- Extended topology-aware collectives
- Power-aware collectives
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended Checkpoint-Restart and migration support with SCR
- Low-overhead and high-performance Virtualization support
Web Pointers

NOWLAB Web Page
http://nowlab.cse.ohio-state.edu

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu