The MVAPICH2 Project: Latest Developments and Plans Towards Exascale Computing

Presentation at Mellanox Theatre (SC ‘19)

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/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)
- Available on HPC Clouds, e.g., Amazon EC2, NSF Chameleon, Microsoft Azure, etc.
MPI+X Programming model: Broad Challenges at Exascale

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Scalable job start-up
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Balancing intra-node and inter-node communication for next generation nodes (128-1024 cores)
  - Multiple end-points per node
- Support for efficient multi-threading
- Integrated Support for GPGPUs and FPGAs
- Fault-tolerance/resiliency
- QoS support for communication and I/O
- Support for Hybrid MPI+PGAS programming (MPI + OpenMP, MPI + UPC, MPI+UPC++, MPI + OpenSHMEM, CAF, ...)
- Virtualization
- Energy-Awareness
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 3,050 organizations in 89 countries
- More than 615,000 (> 0.6 million) downloads from the OSU site directly
- Empowering many TOP500 clusters (Jun ‘19 ranking)
  - 3rd, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center
  - 5th, 448,448 cores (Frontera) at TACC
  - 8th, 391,680 cores (ABCI) in Japan
  - 15th, 570,020 cores (Neurion) in South Korea and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)

- http://mvapich.cse.ohio-state.edu Partner in the TACC Frontera System
- Empowering Top500 systems for over a decade
# Architecture of MVAPICH2 Software Family (HPC and DL)

## High Performance Parallel Programming Models

<table>
<thead>
<tr>
<th>Message Passing Interface (MPI)</th>
<th>PGAS (UPC, OpenSHMEM, CAF, UPC++)</th>
<th>Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)</th>
</tr>
</thead>
</table>

## 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)

<table>
<thead>
<tr>
<th>Transport Protocols</th>
<th>Modern Features</th>
<th>Transport Mechanisms</th>
<th>Modern Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>UMR</td>
<td>Shared Memory</td>
<td>Optane*</td>
</tr>
<tr>
<td>SRD</td>
<td>ODP</td>
<td>CMA</td>
<td>NVLink</td>
</tr>
<tr>
<td>UD</td>
<td>SR-IOV</td>
<td>IVSHMEM</td>
<td>CAPI*</td>
</tr>
<tr>
<td>DC</td>
<td>Multi Rail</td>
<td>XPMEM</td>
<td></td>
</tr>
</tbody>
</table>

### Support for Modern Multi-/Many-core Architectures

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

- Optane*  
- NVLink  
- CAPI*  

* Upcoming
## MVAPICH2 Software Family

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI with IB, iWARP, Omni-Path, and RoCE</td>
<td>MVAPICH2</td>
</tr>
<tr>
<td>Advanced MPI Features/Support, OSU INAM, PGAS and MPI+PGAS with IB, Omni-Path, and RoCE</td>
<td>MVAPICH2-X</td>
</tr>
<tr>
<td>MPI with IB, RoCE &amp; GPU and Support for Deep Learning</td>
<td>MVAPICH2-GDR</td>
</tr>
<tr>
<td>HPC Cloud with MPI &amp; IB</td>
<td>MVAPICH2-Virt</td>
</tr>
<tr>
<td>Energy-aware MPI with IB, iWARP and RoCE</td>
<td>MVAPICH2-EA</td>
</tr>
<tr>
<td>MPI Energy Monitoring Tool</td>
<td>OEMT</td>
</tr>
<tr>
<td>InfiniBand Network Analysis and Monitoring</td>
<td>OSU INAM</td>
</tr>
<tr>
<td>Microbenchmarks for Measuring MPI and PGAS Performance</td>
<td>OMB</td>
</tr>
</tbody>
</table>
MVAPICH2 Distributions

• MVAPICH2
  – Basic MPI support for IB, iWARP and RoCE

• MVAPICH2-X
  – Advanced MPI features and support for INAM
  – MPI, PGAS and Hybrid MPI+PGAS support for IB

• MVAPICH2-Virt
  – Optimized for HPC Clouds with IB and SR-IOV virtualization
  – Support for OpenStack, Docker, and Singularity

• OSU Micro-Benchmarks (OMB)
  – MPI (including CUDA-aware MPI), OpenSHMEM and UPC

• OSU INAM
  – InfiniBand Network Analysis and Monitoring Tool

• MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 1:30-2:00pm)
One-way Latency: MPI over IB with MVAPICH2

**Small Message Latency**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.11</td>
</tr>
<tr>
<td>16</td>
<td>1.19</td>
</tr>
<tr>
<td>32</td>
<td>1.11</td>
</tr>
<tr>
<td>64</td>
<td>1.01</td>
</tr>
<tr>
<td>128</td>
<td>1.15</td>
</tr>
<tr>
<td>256</td>
<td>1.04</td>
</tr>
<tr>
<td>512</td>
<td>1.10</td>
</tr>
<tr>
<td>1K</td>
<td>1.10</td>
</tr>
</tbody>
</table>

- **TrueScale-QDR** - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
- **ConnectX-3-FDR** - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
- **ConnectIB-DualFDR** - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
- **ConnectX-4-EDR** - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch
- **Omni-Path** - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch
- **ConnectX-6-HDR** - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

**Large Message Latency**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2K</td>
<td>120</td>
</tr>
<tr>
<td>4K</td>
<td>100</td>
</tr>
<tr>
<td>8K</td>
<td>80</td>
</tr>
<tr>
<td>16K</td>
<td>60</td>
</tr>
<tr>
<td>32K</td>
<td>40</td>
</tr>
<tr>
<td>64K</td>
<td>20</td>
</tr>
<tr>
<td>128K</td>
<td>10</td>
</tr>
<tr>
<td>256K</td>
<td>5</td>
</tr>
</tbody>
</table>

- **TrueScale-QDR**
- **ConnectX-3-FDR**
- **ConnectIB-DualFDR**
- **ConnectX-4-EDR**
- **Omni-Path**
- **ConnectX-6 HDR**
Bandwidth: MPI over IB with MVAPICH2

![Graph showing bandwidth results for different configurations.](image)

- **Unidirectional Bandwidth**
  - Message Size (bytes)
  - Bandwidth (MBytes/sec)
  - TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
  - ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
  - ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
  - ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
  - Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch
  - ConnectX-6-HDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

- **Bidirectional Bandwidth**
  - Message Size (bytes)
  - Bandwidth (MBytes/sec)
  - TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
  - ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
  - ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
  - ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
  - Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch
  - ConnectX-6-HDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
Intra-node Point-to-Point Performance on OpenPOWER

**Intra-Socket Small Message Latency**

- **Graph 1:**
  - X-axis: Message Size (in bytes)
  - Y-axis: Latency (in us)
  - Legend:
    - MVAPICH2-X 2.3.2
    - SpectrumMPI-10.3.0.01

Intra-Socket Latency Comparison:
- MVAPICH2-X 2.3.2: Linear increase with message size.
- SpectrumMPI-10.3.0.01: Steeper increase with message size.

**Observation:**
- 0.25us latency observed at small message sizes.

**Intra-Socket Large Message Latency**

- **Graph 2:**
  - X-axis: Message Size (in bytes)
  - Y-axis: Latency (in us)
  - Legend:
    - MVAPICH2-X 2.3.2
    - SpectrumMPI-10.3.0.01

Intra-Socket Latency Comparison:
- MVAPICH2-X 2.3.2: Steeper increase with message size.
- SpectrumMPI-10.3.0.01: More linear increase with message size.

**Observation:**
- Steeper latency increase observed with larger message sizes.

**Intra-Socket Bandwidth**

- **Graph 3:**
  - X-axis: Message Size (in bytes)
  - Y-axis: Bandwidth (in MB/s)
  - Legend:
    - MVAPICH2-X 2.3.2
    - SpectrumMPI-10.3.0.01

Intra-Socket Bandwidth Comparison:
- MVAPICH2-X 2.3.2: Steeper increase with message size.
- SpectrumMPI-10.3.0.01: More linear increase with message size.

**Observation:**
- Improved bandwidth observed with larger message sizes.

**Intra-Socket Bi-directional Bandwidth**

- **Graph 4:**
  - X-axis: Message Size (in bytes)
  - Y-axis: Bandwidth (in MB/s)
  - Legend:
    - MVAPICH2-X 2.3.2
    - SpectrumMPI-10.3.0.01

Intra-Socket Bi-directional Bandwidth Comparison:
- MVAPICH2-X 2.3.2: Steeper increase with message size.
- SpectrumMPI-10.3.0.01: More linear increase with message size.

**Observation:**
- Steeper bi-directional bandwidth increase observed with larger message sizes.

**Platform:**
- Two nodes of OpenPOWER (Power9-ppc64le) CPU using Mellanox EDR (MT4121) HCA
Point-to-point: Latency & Bandwidth (Intra-socket) on ARM

Latency - Small Messages
- MVAPICH2-X-Next
- HPCX
- OpenMPI+UCX

Latency - Medium Messages
- MVAPICH2-X-Next
- HPCX
- OpenMPI+UCX

Latency - Large Messages
- MVAPICH2-X-Next
- HPCX
- OpenMPI+UCX

Bandwidth - Small Messages
- MVAPICH2-X-Next
- HPCX
- OpenMPI+UCX

Bandwidth - Medium Messages
- MVAPICH2-X-Next
- HPCX
- OpenMPI+UCX

Bandwidth - Large Messages
- MVAPICH2-X-Next
- HPCX
- OpenMPI+UCX

70% better
Startup Performance on TACC Frontera

MPI_Init on Frontera

- MPI_Init takes 3.9 seconds on 57,344 processes on 1,024 nodes
- All numbers reported with 56 processes per node

New designs available in MVAPICH2-2.3.2
Bcast with RDMA_CM Hardware Multicast on Frontera

- MPI_Bcast shows flat scalability for increasing number of nodes
- All numbers reported with 56 processes per node
Evaluation of SHArP based Non Blocking Allreduce

- Complete offload of Allreduce collective operation to Switch helps to have much higher overlap of communication and computation

Available since MVAPICH2 2.3a
Benefits of SHARP Allreduce at Application Level

Avg DDOT Allreduce time of HPCG

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV2_ENABLE_SHARP=1</td>
<td>Enables SHARP-based collectives</td>
<td>Disabled</td>
</tr>
<tr>
<td>--enable-sharp</td>
<td>Configure flag to enable SHARP</td>
<td>Disabled</td>
</tr>
</tbody>
</table>

SHARP support available since MVAPICH2 2.3a

- Refer to Running Collectives with Hardware based SHARP support section of MVAPICH2 user guide for more information
- [http://mvapich.cse.ohio-state.edu/static/media/mvapich/mvapich2-2.3-userguide.html#x1-990006.26](http://mvapich.cse.ohio-state.edu/static/media/mvapich/mvapich2-2.3-userguide.html#x1-990006.26)
Optimized CMA-based Collectives for Large Messages

- Significant improvement over existing implementation for Scatter/Gather with 1MB messages (up to 4x on KNL, 2x on Broadwell, 14x on OpenPower)
- New two-level algorithms for better scalability
- Improved performance for other collectives (Bcast, Allgather, and Alltoall)


Available since MVAPICH2-X 2.3b
Shared Address Space (XPMEM)-based Collectives Design

**OSU_Allreduce (Broadwell 256 procs)**

- MVAPICH2-2.3b
- IMPI-2017v1.132
- MVAPICH2-X-2.3rc1

**OSU_Reduce (Broadwell 256 procs)**

- MVAPICH2-2.3b
- IMPI-2017v1.132
- MVAPICH2-2.3rc1

- **“Shared Address Space”**-based true zero-copy Reduction collective designs in MVAPICH2
- Offloaded computation/communication to peers ranks in reduction collective operation
- Up to 4X improvement for 4MB Reduce and up to 1.8X improvement for 4M AllReduce


Available in MVAPICH2-X 2.3rc1
Minimizing Memory Footprint by Direct Connect (DC) Transport

- Constant connection cost *(One QP for any peer)*
- Full Feature Set (RDMA, Atomics etc)
- Separate objects for send (DC Initiator) and receive (DC Target)
  - DC Target identified by “DCT Number”
  - Messages routed with (DCT Number, LID)
  - Requires same “DC Key” to enable communication
- Available since MVAPICH2-X 2.2a

User-mode Memory Registration (UMR)

- Introduced by Mellanox to support direct local and remote noncontiguous memory access
- Avoid packing at sender and unpacking at receiver
- Available since MVAPICH2-X 2.2b

Connect-IB (54 Gbps): 2.8 GHz Dual Ten-core (IvyBridge) Intel PCI Gen3 with Mellanox IB FDR switch

On-Demand Paging (ODP)

- Applications no longer need to pin down underlying physical pages
- Memory Region (MR) are *NEVER* pinned by the OS
  - Paged in by the HCA when needed
  - Paged out by the OS when reclaimed
- ODP can be divided into two classes
  - Explicit ODP
    - Applications still register memory buffers for communication, but this operation is used to define access control for IO rather than pin-down the pages
  - Implicit ODP
    - Applications are provided with a special memory key that represents their complete address space, does not need to register any virtual address range
- Advantages
  - Simplifies programming
  - Unlimited MR sizes
  - Physical memory optimization


Available since MVAPICH2-X 2.3b
Impact of Zero Copy MPI Message Passing using HW Tag Matching (Point-to-point)

Removal of intermediate buffering/copies can lead up to 35% performance improvement in latency of medium messages on TACC Frontera
Benefits of the New Asynchronous Progress Design: Broadwell + InfiniBand

Up to 44% performance improvement in P3DFFT application with 448 processes
Up to 19% and 9% performance improvement in HPL application with 448 and 896 processes

A. Ruhela, H. Subramoni, S. Chakraborty, M. Bayatpour, P. Kousha, and D.K. Panda, Efficient Asynchronous Communication Progress for MPI without Dedicated Resources, EuroMPI 2018

Available in MVAPICH2-X 2.3rc1
Evaluation of Applications on Frontera (Cascade Lake + HDR100)

Performance of MILC and WRF2 applications scales well with increase in system size.
MVAPICH2 Distributions

- **MVAPICH2**
  - Basic MPI support for IB, iWARP and RoCE
- **MVAPICH2-X**
  - Advanced MPI features and support for INAM
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- **MVAPICH2-Virt**
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - Support for OpenStack, Docker, and Singularity
- **OSU Micro-Benchmarks (OMB)**
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- **OSU INAM**
  - InfiniBand Network Analysis and Monitoring Tool
- **MVAPICH2-GDR and Deep Learning** (Will be presented on Thursday at 10:30am)
Can HPC and Virtualization be Combined?

• 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

J. Zhang, X. Lu, J. Jose, R. Shi and D. K. Panda, Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? EuroPar'14
J. Zhang, X. Lu, J. Jose, M. Li, R. Shi and D.K. Panda, High Performance MPI Library over SR-IOV enabled InfiniBand Clusters, HiPC’14
Application-Level Performance on Chameleon

- 32 VMs, 6 Core/VM
- Compared to Native, 2-5% overhead for Graph500 with 128 Procs
- Compared to Native, 1-9.5% overhead for SPEC MPI2007 with 128 Procs
Performance of Radix on Microsoft Azure

Total Execution Time on HC (Lower is better)

- **MVAPICH2-X**
- **HPCx**

3x faster

Number of Processes (Nodes X PPN)

<table>
<thead>
<tr>
<th>Number of Processes</th>
<th>Execution Time (Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16(1x16)</td>
<td>20</td>
</tr>
<tr>
<td>32(1x32)</td>
<td>15</td>
</tr>
<tr>
<td>44(1X44)</td>
<td>14</td>
</tr>
<tr>
<td>88(2X44)</td>
<td>13</td>
</tr>
<tr>
<td>176(4X44)</td>
<td>12</td>
</tr>
<tr>
<td>352(8X44)</td>
<td>11</td>
</tr>
</tbody>
</table>

Total Execution Time on HB (Lower is better)

- **MVAPICH2-X**
- **HPCx**

38% faster

Number of Processes (Nodes X PPN)

<table>
<thead>
<tr>
<th>Number of Processes</th>
<th>Execution Time (Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60(1X60)</td>
<td>5</td>
</tr>
<tr>
<td>120(2X60)</td>
<td>10</td>
</tr>
<tr>
<td>240(4X60)</td>
<td>15</td>
</tr>
</tbody>
</table>
Amazon Elastic Fabric Adapter (EFA)

- Enhanced version of Elastic Network Adapter (ENA)
  - Allows OS bypass, up to 100 Gbps bandwidth
- New QP type: **Scalable Reliable Datagram (SRD)**
  - Network aware multi-path routing - low tail latency
  - Guaranteed Delivery, no ordering guarantee
- Exposed through verbs and libfabric interfaces

---

### Feature Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>UD</th>
<th>SRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send/Recv</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Send w/ Immediate</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>RDMA</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>Read/Write/Atomic</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>Scatter Gather Lists</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Shared Receive Queue</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>Reliable Delivery</td>
<td>✖</td>
<td>✔</td>
</tr>
<tr>
<td>Ordering</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>Inline Sends</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>Global Routing Header</td>
<td>✔</td>
<td>✖</td>
</tr>
<tr>
<td>MTU Size</td>
<td>4KB</td>
<td>8KB</td>
</tr>
</tbody>
</table>

---

**Verbs-level Latency**

- UD: 15.69 us, 16.95 us
- SRD: 15.69 us

**Unidirectional Message Rate**

- UD: 1.77 million msg/sec, 2.02 million msg/sec
- SRD: 1.77 million msg/sec
**MPI-level Performance with SRD**

Scatterv – 8 node 36 ppn

Gatherv – 8 node 36 ppn

Allreduce – 8 node 36 ppn

---

**miniGhost** 10% better

**CloverLeaf**

Instance type: c5n.18xlarge
CPU: Intel Xeon Platinum 8124M @ 3.00GHz
MVAPICH2 version: MVAPICH2-X 2.3rc2 + SRD support
OpenMPI version: Open MPI v3.1.3 with libfabric 1.7

*S. Chakraborty, S. Xu, H. Subramoni, and D. K. Panda, Designing Scalable and High-performance MPI Libraries on Amazon Elastic Fabric Adapter, to be presented at the 26th Symposium on High Performance Interconnects, (HOTI ’19)*
MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE

- MVAPICH2-X
  - Advanced MPI features and support for INAM
  - MPI, PGAS and Hybrid MPI+PGAS support for IB

- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - Support for OpenStack, Docker, and Singularity

- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC

- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 10:30am)
OSU Microbenchmarks

- Available since 2004
- Suite of microbenchmarks to study communication performance of various programming models

  - Benchmarks available for the following programming models
    - Message Passing Interface (MPI)
    - Partitioned Global Address Space (PGAS)
      - Unified Parallel C (UPC)
      - Unified Parallel C++ (UPC++)
      - OpenSHMEM

  - Benchmarks available for multiple accelerator based architectures
    - Compute Unified Device Architecture (CUDA)
    - OpenACC Application Program Interface

- Part of various national resource procurement suites like NERSC-8 / Trinity Benchmarks
- Please visit the following link for more information
  - http://mvapich.cse.ohio-state.edu/benchmarks/
MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - Advanced MPI features and support for INAM
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - Support for OpenStack, Docker, and Singularity
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPICH2-GDR and Deep Learning (Will be presented on Wednesday at 1:30 pm)
Overview of OSU INAM

- A network monitoring and analysis tool that is capable of analyzing traffic on the InfiniBand network with inputs from the MPI runtime
  - [http://mvapich.cse.ohio-state.edu/tools/osu-inam/](http://mvapich.cse.ohio-state.edu/tools/osu-inam/)
- Monitors IB clusters in real time by querying various subnet management entities and gathering input from the MPI runtimes
- Capability to analyze and profile node-level, job-level and process-level activities for MPI communication
  - Point-to-Point, Collectives and RMA
- Ability to filter data based on type of counters using “drop down” list
- Remotely monitor various metrics of MPI processes at user specified granularity
- "Job Page" to display jobs in ascending/descending order of various performance metrics in conjunction with MVAPICH2-X
- Visualize the data transfer happening in a “live” or “historical” fashion for entire network, job or set of nodes
- OSU INAM 0.9.4 released on 11/10/2018
  - Enhanced performance for fabric discovery using optimized OpenMP-based multi-threaded designs
  - Ability to gather InfiniBand performance counters at sub-second granularity for very large (>2000 nodes) clusters
  - Redesign database layout to reduce database size
  - Enhanced fault tolerance for database operations
    - Thanks to Trey Dockendorf @ OSC for the feedback
  - OpenMP-based multi-threaded designs to handle database purge, read, and insert operations simultaneously
  - Improved database purging time by using bulk deletes
  - Tune database timeouts to handle very long database operations
  - Improved debugging support by introducing several debugging levels
OSU INAM Features

Comet@SDSC --- Clustered View

(1,879 nodes, 212 switches, 4,377 network links)

- Show network topology of large clusters
- Visualize traffic pattern on different links
- Quickly identify congested links/links in error state
- See the history unfold – play back historical state of the network

Finding Routes Between Nodes
OSU INAM Features (Cont.)

- **Visualizing a Job (5 Nodes)**
  - Job level view
    - Show different network metrics (load, error, etc.) for any live job
    - Play back historical data for completed jobs to identify bottlenecks
  - Node level view - details per process or per node
    - CPU utilization for each rank/node
    - Bytes sent/received for MPI operations (pt-to-pt, collective, RMA)
    - Network metrics (e.g. XmitDiscard, RcvError) per rank/node

- **Estimated Process Level Link Utilization**
  - Estimated Link Utilization view
    - Classify data flowing over a network link at different granularity in conjunction with MVAPICH2-X 2.2rc1
      - Job level and
      - Process level
Applications-Level Tuning: Compilation of Best Practices

- MPI runtime has many parameters
- Tuning a set of parameters can help you to extract higher performance
- Compiled a list of such contributions through the MVAPICH Website
  - [http://mvapich.cse.ohio-state.edu/best_practices/](http://mvapich.cse.ohio-state.edu/best_practices/)
- Initial list of applications
  - Amber
  - HoomDBlue
  - HPCG
  - Lulesh
  - MILC
  - Neuron
  - SMG2000
  - Cloverleaf
  - SPEC (LAMMPS, POP2, TERA_TF, WRF2)
- Soliciting additional contributions, send your results to mvapich-help at cse.ohio-state.edu.
- We will link these results with credits to you.
Commercial Support for MVAPICH2 Libraries

• Supported through X-ScaleSolutions (http://x-scalesolutions.com)

• Benefits:
  – Help and guidance with installation of the library
  – Platform-specific optimizations and tuning
  – Timely support for operational issues encountered with the library
  – Web portal interface to submit issues and tracking their progress
  – Advanced debugging techniques
  – Application-specific optimizations and tuning
  – Obtaining guidelines on best practices
  – Periodic information on major fixes and updates
  – Information on major releases
  – Help with upgrading to the latest release
  – Flexible Service Level Agreements

• Support provided to Lawrence Livermore National Laboratory (LLNL) during last two years
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
One More Presentation

• Wednesday (11/21/19) at 1:30pm

MVAPICH2-GDR: Pushing the Frontier of HPC and Deep Learning
Join us for Multiple Events at SC ‘19

• Presentations at OSU and X-Scale Booth (#2094)
  – 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/752/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/