Enhancing MPI for HPC Clouds: Case Study with Amazon AWS and Microsoft Azure

Shulei Xu, Mahdieh Ghazimirsaeed Jahanzeb Hashmi and Dhabaleswar K. Panda (Advisor)

More Information in: http://mvapich.cse.ohio-state.edu/

THE OHIO STATE UNIVERSITY

Overview

New Trend of Cloud VM in HPC:
- Cloud VMs will significantly change future of High-Performance Computing (HPC)
- AWS EFA targeted towards tightly coupled HPC workloads
- New instances released with support of MPI

MPI Support for AWS and Azure:
- Implement MPI support for AWS EFA
- Propose designs based on AWS EFA SRD & UD
- Add support & performance optimization for Azure HB & HC high performance virtual machines
- MVAPICH2 with XPMEM on AWS & Azure

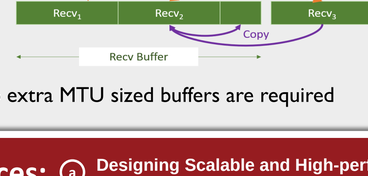
Performance Evaluation:
- Performance evaluation of UD & SRD design on AWS
- Performance comparing evaluation between MVAPICH2 (optimized & non-optimized) and OpenMPI on AWS & Azure clusters

Challenges

Support for AWS EFA Instances

Elastic Fabric Adapter (EFA) for Tightly-Coupled HPC Workloads
- Provide a new transport mode called Scalable Reliable Datagram (SRD)
  - Similar to UD but with reliable delivery
  - SRD supports a larger MTU (Maximum Transmission Unit) (8KB) than UD (4KB)
  - Some features on other IB adapters not supported

HPC Software Stack with EFA



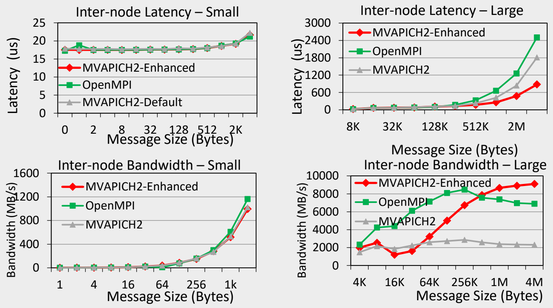
Limitations of EFA

Support for Azure VM

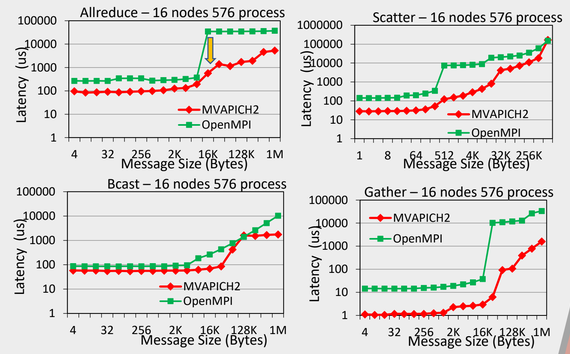
In-depth Performance Evaluation
- Various MPI libraries available on Azure HPC VM
  - MVAPICH2
  - HPCX-2.3.0
  - OpenMPI-4.0.1 with UCX
  - Intel MPI 2019
- In-depth performance evaluation on HB & HC instances
  - Pt-to-pt performance
  - Collective performance
  - Application level performance
- Performance improvement
  - XPMEM support for MVAPICH2
  - Dedicated pt-to-pt & collective tuning
  - NUMA mapping policy for collective communication
- Easy deployment for user
  - Already included as built-in software in Azure HPC following CentOS image:
    - Azure CentOS 7.6 HPC Image
    - Azure CentOS 7.7 HPC Image
    - Azure CentOS 8.1 HPC Image

MPI Level Performance Evaluation

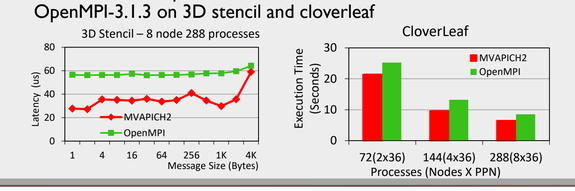
AWS EFA Point-to-Point Evaluation
- MVAPICH2 vs. AWS build-in OpenMPI-4.0.2 on instance c5n18xlarge
- Inter-node pt-to-pt evaluation to designs based on SRD
- MVAPICH2-Enhanced: pt-to-pt operations are tuned to select best designs
- Up to 65% better for latency of large message size
- Up to 3x better for bandwidth with enhanced Rendezvous design



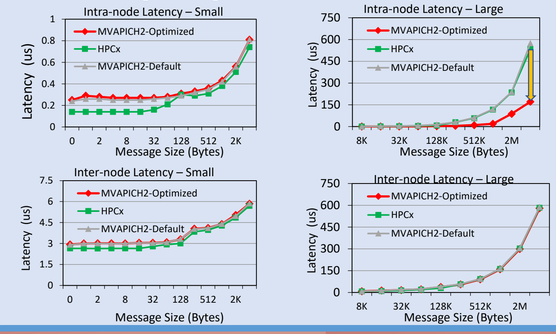
AWS EFA Collective Evaluation
- MVAPICH2 vs. AWS build-in OpenMPI-4.0.2 on instance c5n.18xlarge
- Multi-node collective evaluation to designs based on SRD
- MVAPICH2: collective operations are tuned to select best designs
- MVAPICH2 has up to 25x better performance in Allreduce



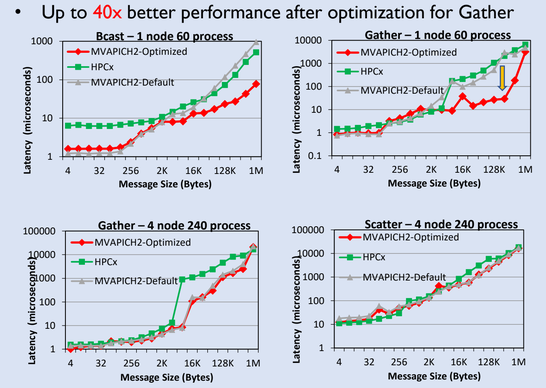
AWS EFA Application Evaluation
- Performance comparison of MVAPICH2 vs. AWS build-in OpenMPI-3.1.3 on 3D stencil and cloverleaf



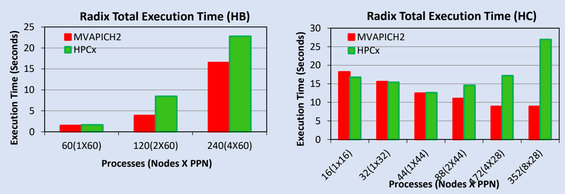
Azure HB & HC Point-to-Point Evaluation
- Performance comparison of MVAPICH2 vs. HPCX-2.3.0
- MVAPICH2-Optimized: pt-to-pt operations are tuned to select best designs
- MVAPICH2 has up to 3x better performance than unoptimized version in large message sizes intra-node latency



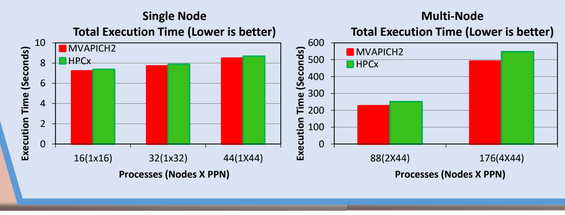
Azure HB & HC Collective Evaluation
- Single and multi-node collective evaluation of MVAPICH2
- MVAPICH2-Optimized: Collective operations are tuned to select best designs
- Up to 40x better performance after optimization for Gather



Azure HB & HC Application Evaluation
- Performance comparison of MVAPICH2 vs. HPCX on radix

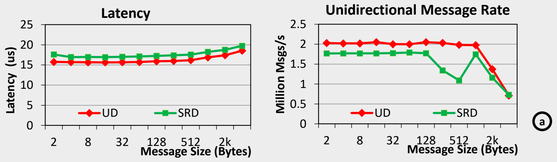


- Performance comparison of MVAPICH2 vs. HPCX on FDS



Solution

Single QP Communicates Multiple Peers
- UD can make use of global routing header (GRH)



[a]

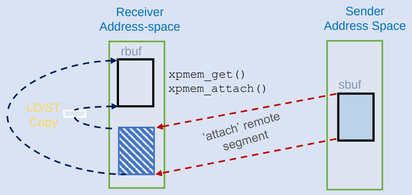
Sliding Window-based Receiver Design
- Packet in-order is delivered to receiver
- Packet out-of-order is temporarily stored

Zero-copy Rendezvous Protocol
- Reorder packets on receiver side
- Out-of-order messages are stored in temporary buffer

[a]

- Two extra MTU sized buffers are required

Advanced XPMEM Support
- Kernel module with user-level API for sharing address spaces among processes
- Multiple processes share their address-spaces
- The communication is realized via accessing the data from remote process' memory



Details of MPI Libraries used for Evaluation
- MVAPICH2
  - Optimized XPMEM & CMA (Cross Memory Attach) support
  - Intra-node co-operative (COOP) Rendezvous protocol for pt-to-pt
  - NUMA-domain aware process mapping
- HPCX v2.3.0
  - Based on OpenMPI-4.0 (03cf3e4)
  - Includes UCX-1.5.0 & HCOLL-4.2.2554

References:
[a] Designing Scalable and High-performance MPI Libraries on Amazon Elastic Fabric Adapter S. Chakraborty, S. Xu, H. Subramoni, D. Panda HOT Interconnects 26, Aug 2019.
[b] EFA for Tightly-Coupled HPC Workloads https://aws.amazon.com/blogs/aws/now-available-elastic-fabric-adapter-efa-for-tightly-coupled-hpc-workloads/