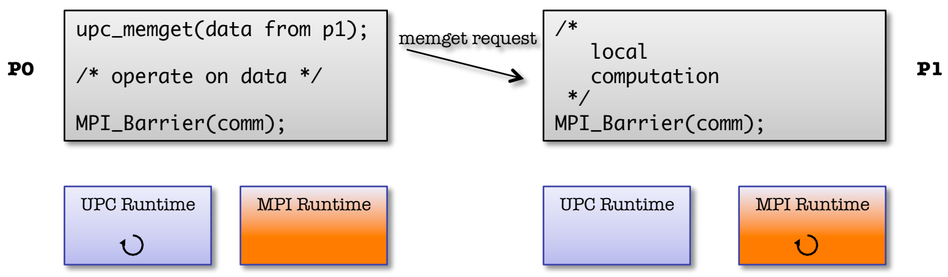
Supporting Hybrid MPI+PGAS Programming Models through Unified Communication Runtime: An MVAPICH2-X Approach

Khaled Hamidouche, Jian Lin, Mingzhe Li, Jie Zhang and D.K. Panda – The Ohio State University
{hamidouc, linjia, limin, zhanjie, panda}@cse.ohio-state.edu

Motivation

Need for a Unified Runtime



- Deadlock when a message is sitting in one runtime, but application calls the other runtime
- Current prescription to avoid this is to barrier in one mode (either PGAS (UPC/OpenSHMEM/CAF) or MPI) before entering the other

  Having multiple runtimes result in bad performance!!!

Coercing UPC/OpenSHMEM/CAF over MPI not Optimal

- MPI does not provide Active Messages
- Current MPI RMA model designed for non cache-coherent machines
  - MPI-3 considering proposal for efficiently supporting cache-coherent machines
- MPI will not support "instant teams"

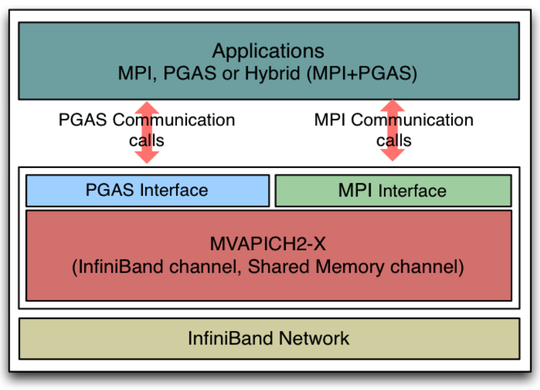
Path forward: unify runtimes, not programming models

Problem Statement

- Can we design a communication library for UPC/OpenSHMEM/CAF?
  - Scalable on large InfiniBand clusters
  - Provides equal or better performance than existing runtime
- Can this library support both MPI and UPC/OpenSHMEM/CAF?
  - Individually, both with great performance

Our Approach - MVAPICH2-X

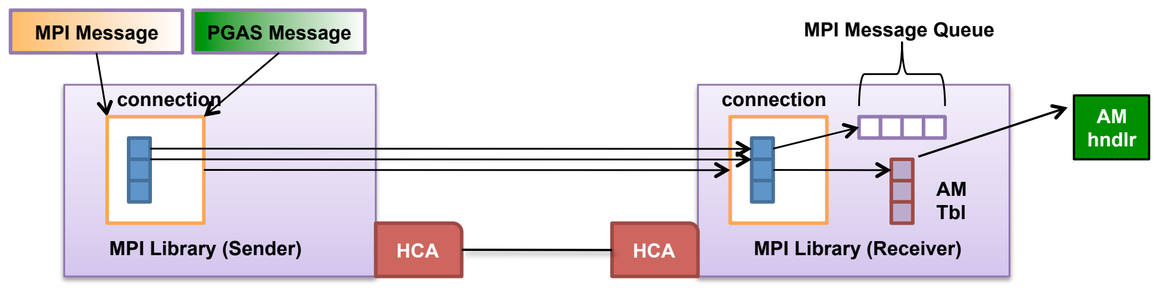
Unified Communication Runtime for MPI and PGAS



- Enables Hybrid (MPI+PGAS) Programming
- Available in MVAPICH2-X 2.0b Release!

- Unified Communication Runtime (UCR) extends MVAPICH2 and provides support for MPI and PGAS (UPC/OpenSHMEM/CAF)
- No deadlock because of single runtime
- Consumes lesser network resources
- MPI Performance not harmed and UPC/OpenSHMEM/CAF performance not penalized

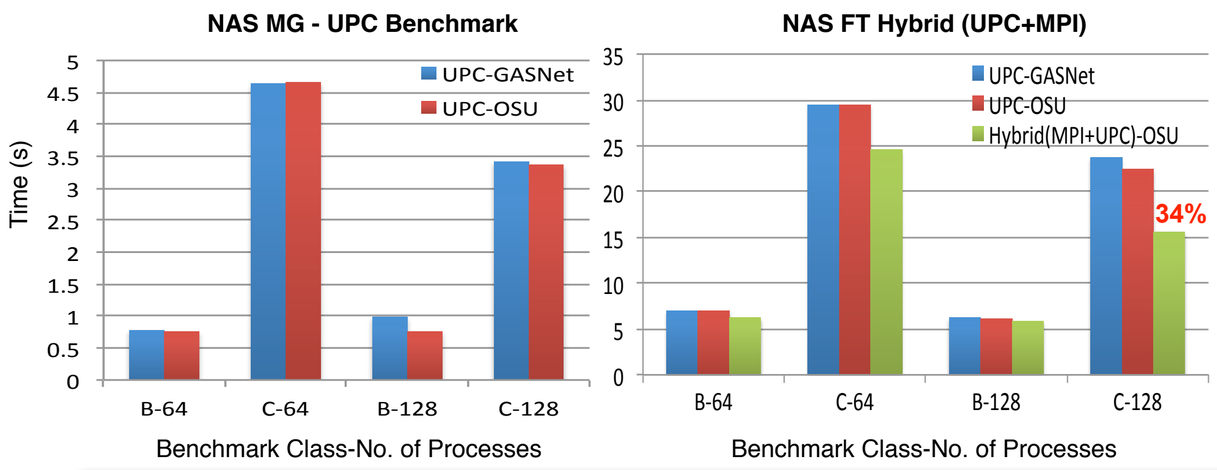
Resources shared between MPI and PGAS



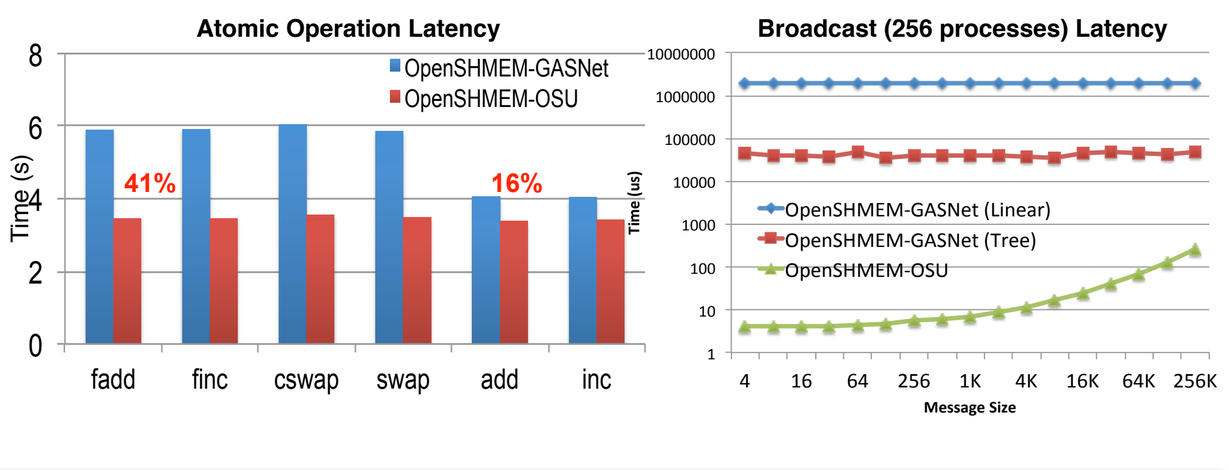
- Resources shared between MPI and UPC/OpenSHMEM/CAF
  - Connections, buffers, memory registrations
  - Schemes for establishing connections (fixed, on-demand)
  - RDMA for large AMs and for PUT, GET

Experimental Results

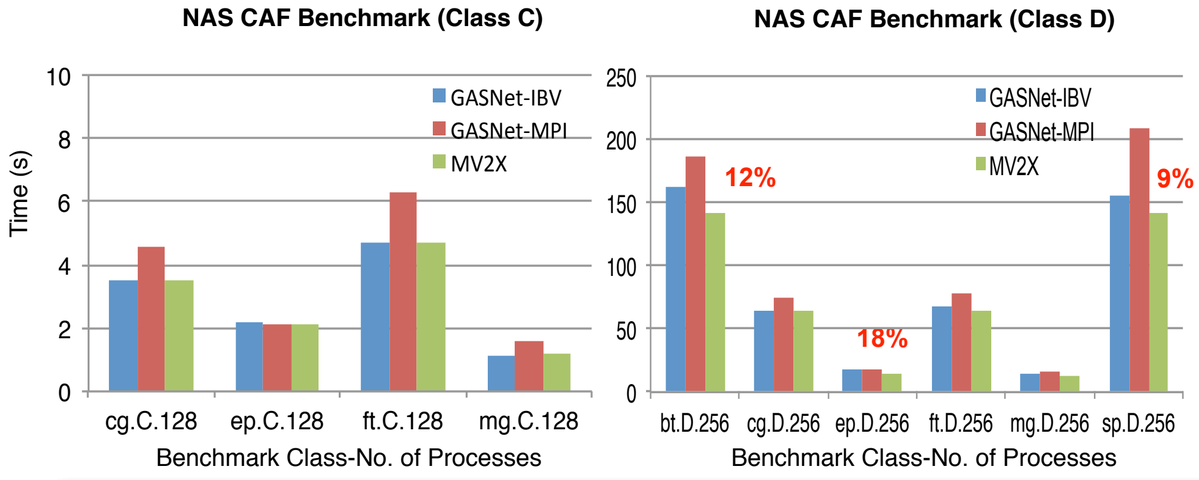
Evaluation using UPC-NAS Benchmarks



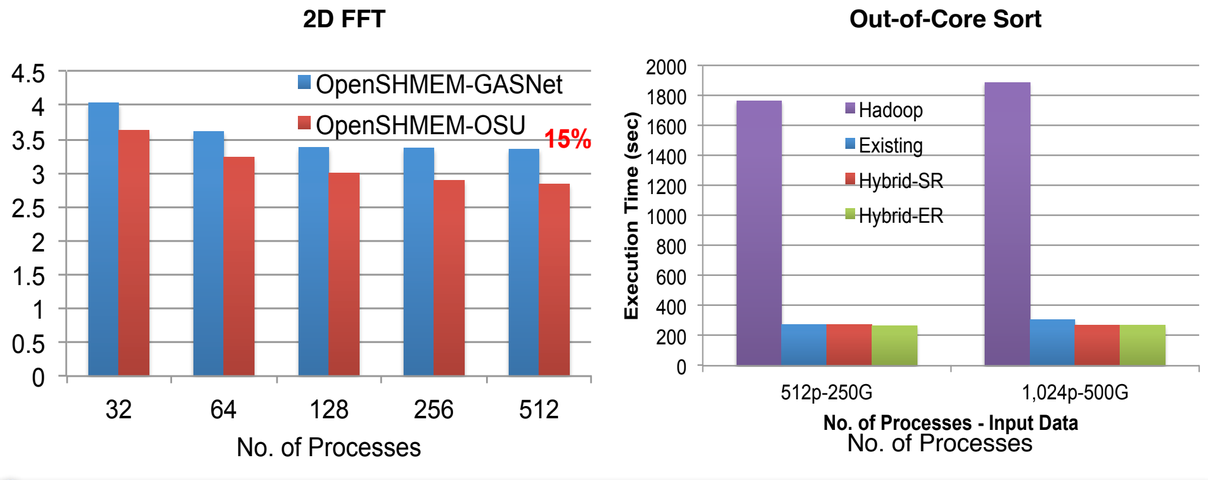
OpenSHMEM Atomics and Broadcast Performance



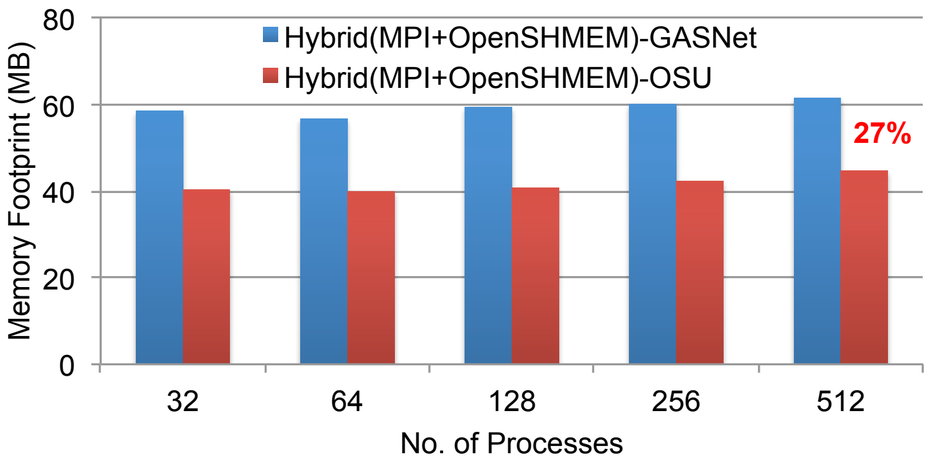
Evaluation using CAF-NAS Benchmarks



Application Evaluation



Network Resource Consumption



- Network resources shared among MPI and PGAS (UPC/OpenSHMEM)
- Lesser resource requirement because of single runtime
- Achieves better scalability

Conclusions

- Promising: MPI communication not harmed; Better performance for UPC/ OpenSHMEM/CAF
- Hybrid MPI+OpenSHMEM Graph500 Benchmark: 13X improvement for 16,384 processes
- Hybrid MPI+UPC FT NAS Benchmark: 34% improvement for Class-C 128 processes
- CAF EP NAS Benchmark: 18% improvement for Class-C 256 processes

Publications:

- J. Jose, S. Potluri, H. Subramon, X. Lu, K. Hamidouche, K. Schulz, H. Sundar and D. K. Panda, Designing Scalable Out-of-core Sorting with Hybrid MPI+PGAS Programming Models (PGAS '14)
- J. Jose, K. Kandalla, S. Potluri, J. Zhang and D. K. Panda, Optimizing Collective Communication in OpenSHMEM, Partitioned Global Address Space Programming Model (PGAS '13)
- J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models , Int'l Super Computing Conference (ISC '13)
- J. Jose, K. Kandalla, M. Luo and D. K. Panda, Supporting MPI & OpenSHMEM over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP '12)
- J. Jose, M. Luo, S. Sur and D. K. Panda, Unifying UPC and MPI Runtimes: Experience with MVAPICH, Partitioned Global Address Space Programming Model (PGAS '10)
- J. Lin, K. Hamidouche, X. Lu, M. Li and D. K. Panda, High-performance Co-array Fortran support with MVAPICH2-X: Initial Experience and evaluation, HIPS 15

Acknowledgements

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

MVAPICH2/MVAPICH2-X: MPI/PGAS over Infiniband,10GE/iWarp & RoCE
http://mvapich.cse.ohio-state.edu/

This research is supported in part by National Science Foundation grants #OCI-0926691, #OCI-1148371 and #CCF-1213084.