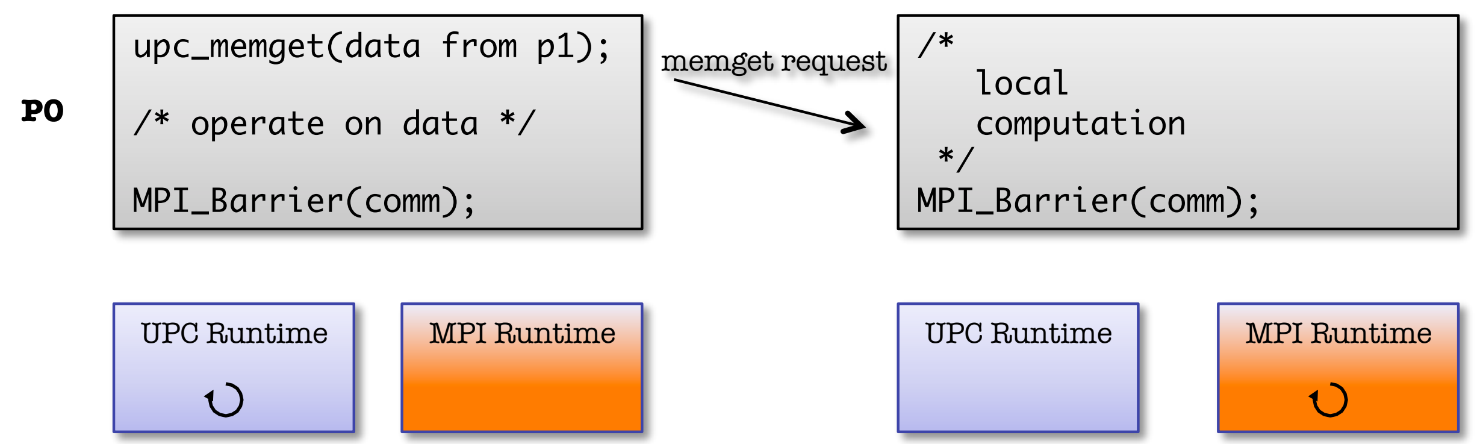
MVAPICH2-X: Unified Communication Runtime for Efficient Hybrid MPI+PGAS Programming Models

Khaled Hamidouche and D.K. Panda – The Ohio State University
{hamidouc, panda}@cse.ohio-state.edu

Motivation

Need for a Unified Runtime



```
upc_memget(data from p1);

/* operate on data */

MPI_Barrier(comm);
```

```
/*
   local
   computation
 */
MPI_Barrier(comm);
```

- 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/CAF/UPC++/OpenSHMEM) or MPI) before entering the other

Having multiple runtimes result in bad performance!!!

Coercing UPC/CAF/UPC++/OpenSHMEM over MPI is 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"

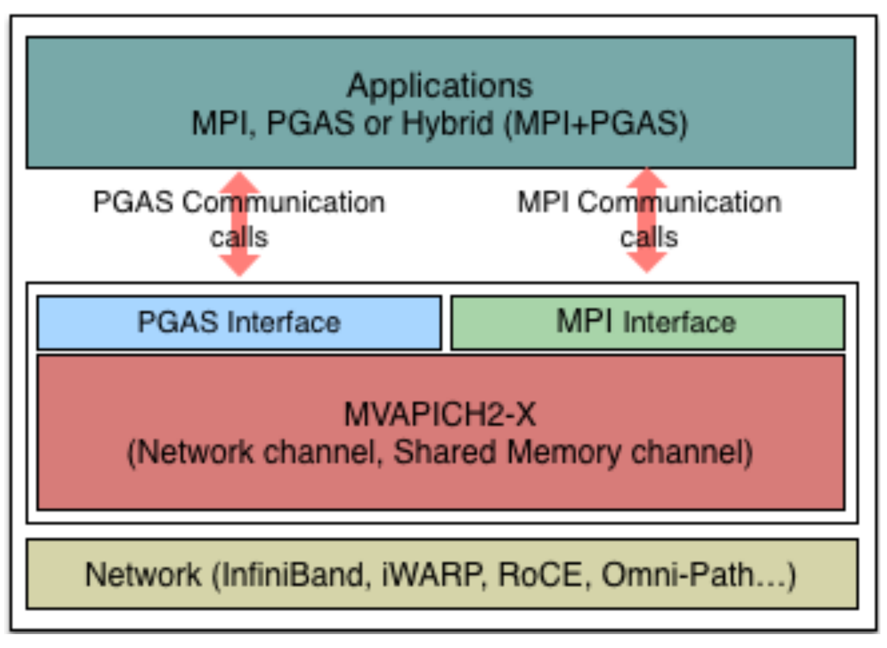
Path forward: unify runtimes, not programming models

Problem Statement

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

MVAPICH2-X: Unified MPI+PGAS Communication Runtime for Exascale Systems

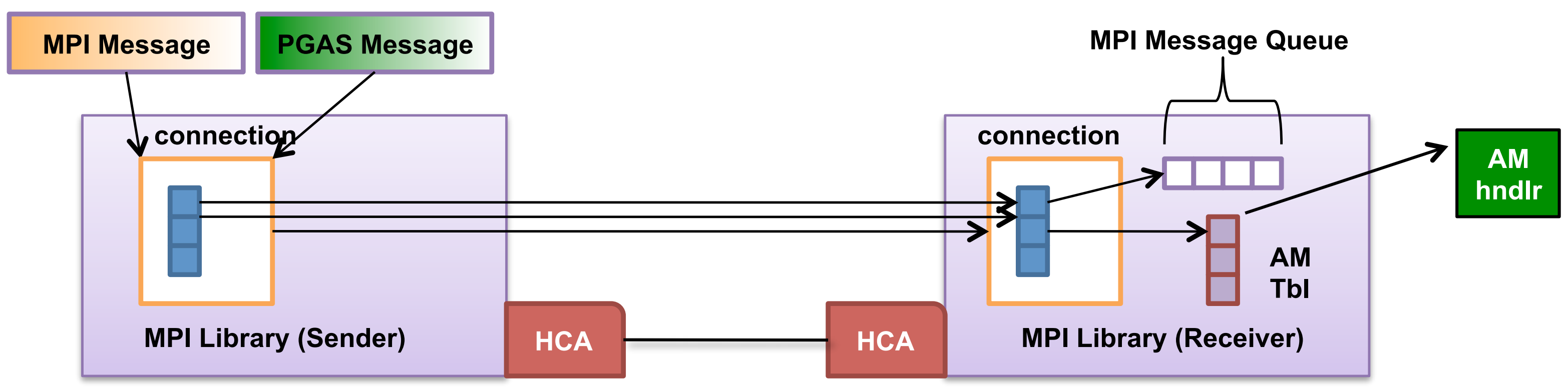
Unified Communication Runtime


Applications
MPI, PGAS or Hybrid (MPI+PGAS)
PGAS Communication calls
MPI Communication calls
PGAS Interface
MPI Interface
MVAPICH2-X
(Network channel, Shared Memory channel)
Network (InfiniBand, iWARP, RoCE, Omni-Path…)

- Supports UPC, UPC++, CAF, OpenSHMEM
- Enables Hybrid MPI+PGAS i.e., MPI+{UPC, UPC++, CAF, OpenSHMEM} Programming
- MPI-3 compliant
- Available since MVAPICH2 1.9 (2012)

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

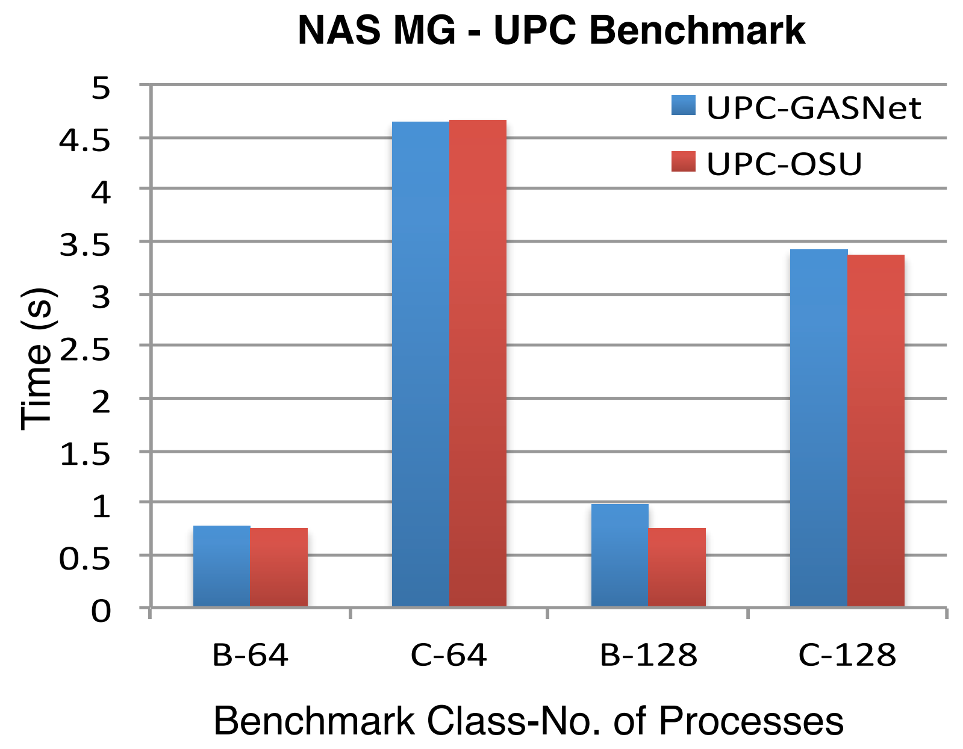
Resources shared between MPI and PGAS

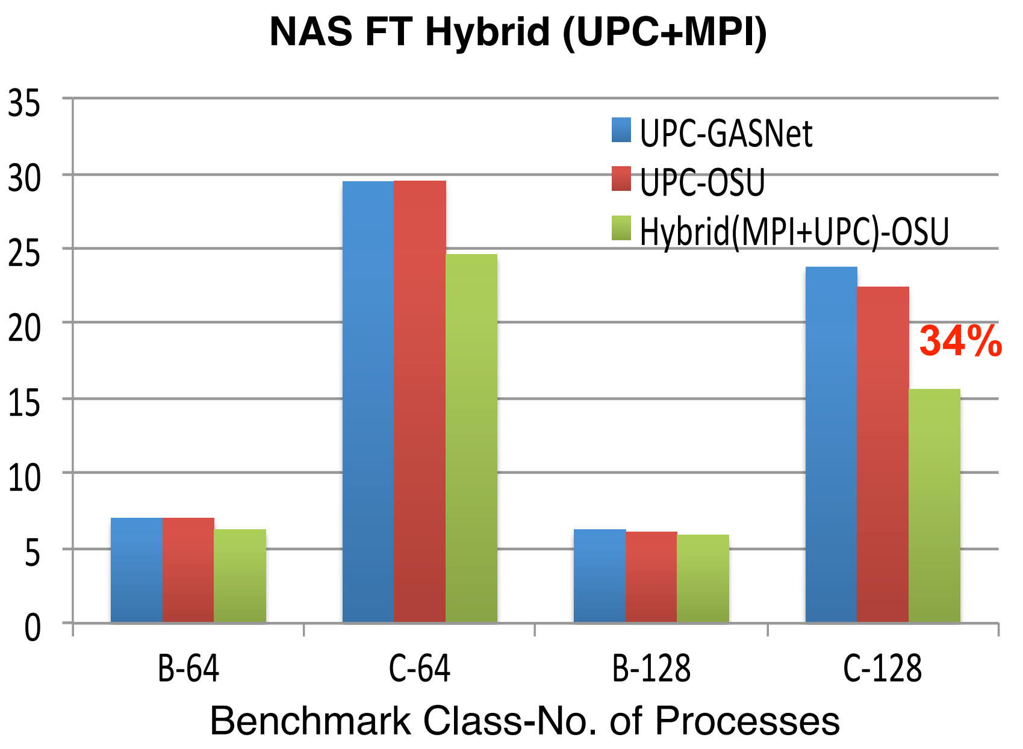


- Resources shared between MPI and UPC/CAF/UPC++/OpenSHMEM
  - Connections, buffers, memory registrations
  - Schemes for establishing connections (fixed, on-demand)
  - RDMA for large AMs and for PUT, GET

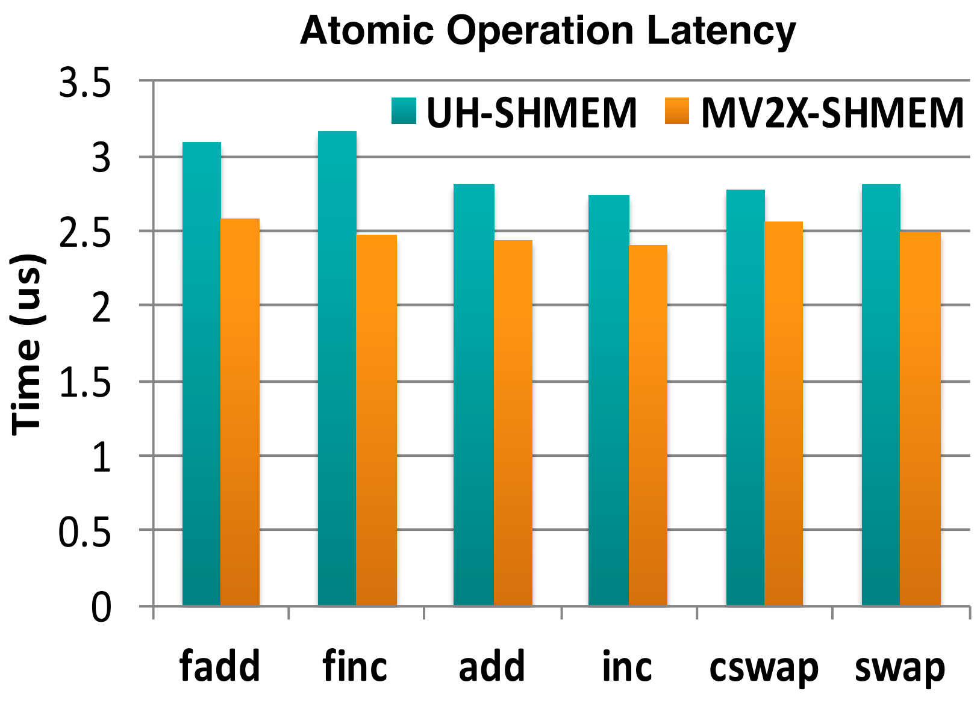
Experimental Results

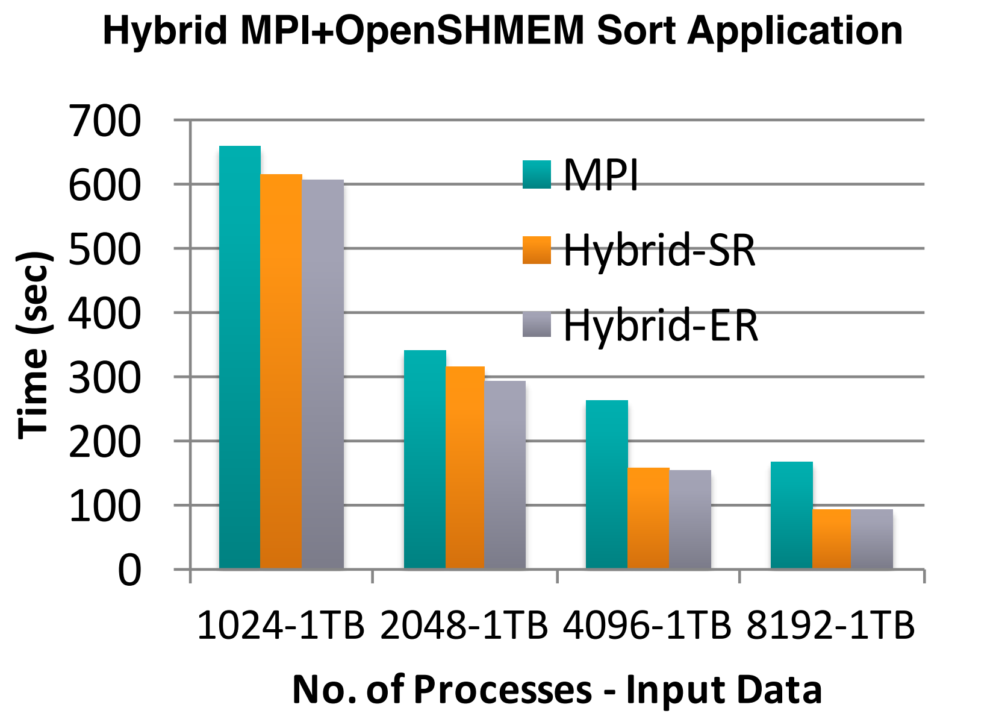
UPC-NAS Benchmarks Performance



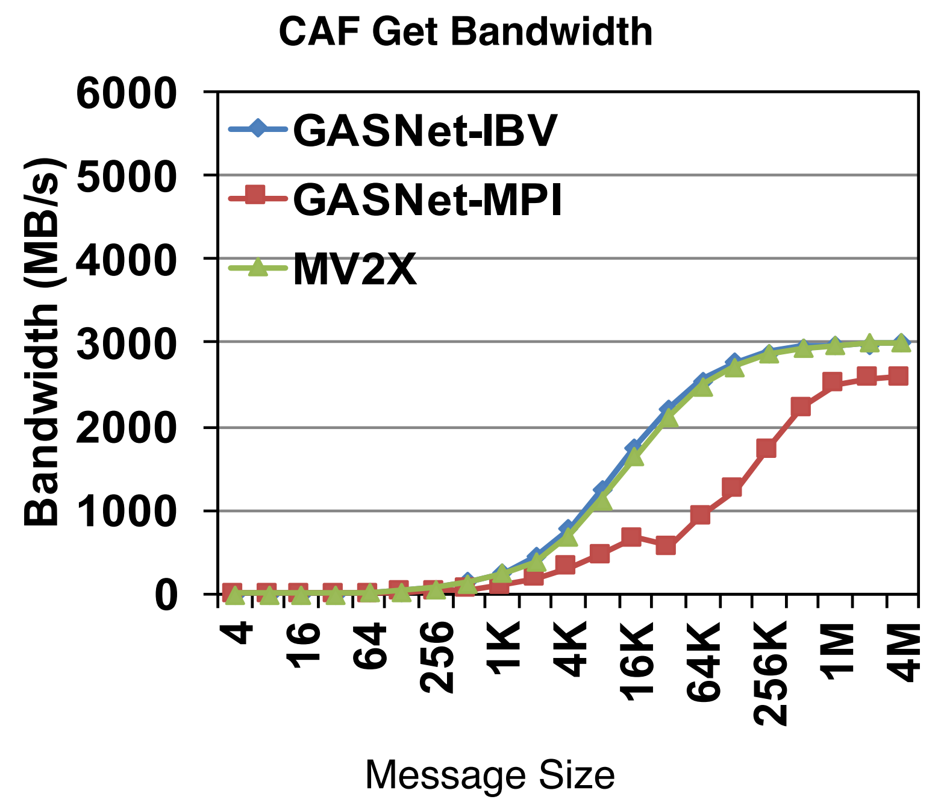


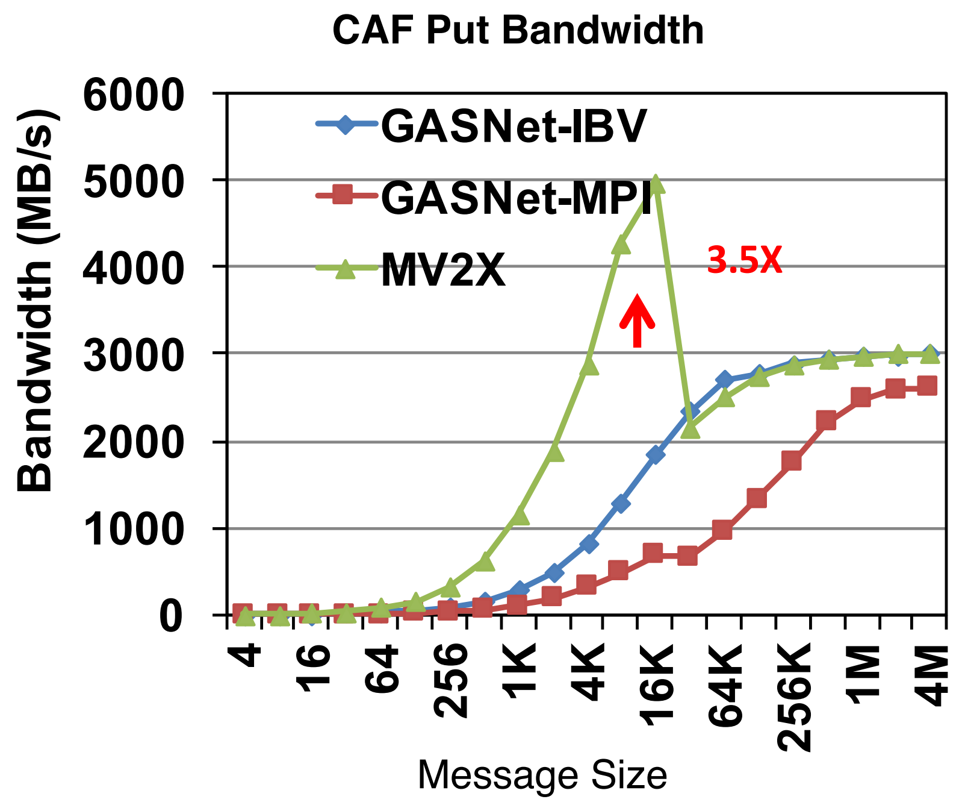
OpenSHMEM Atomics and Hybrid MPI+OpenSHMEM



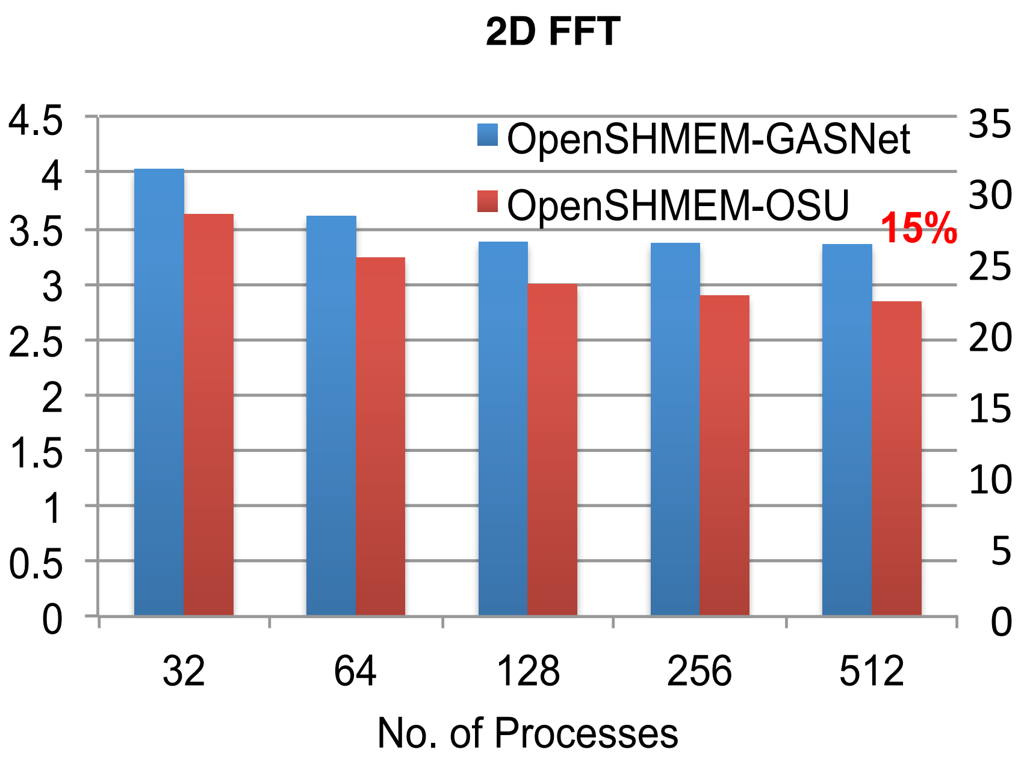


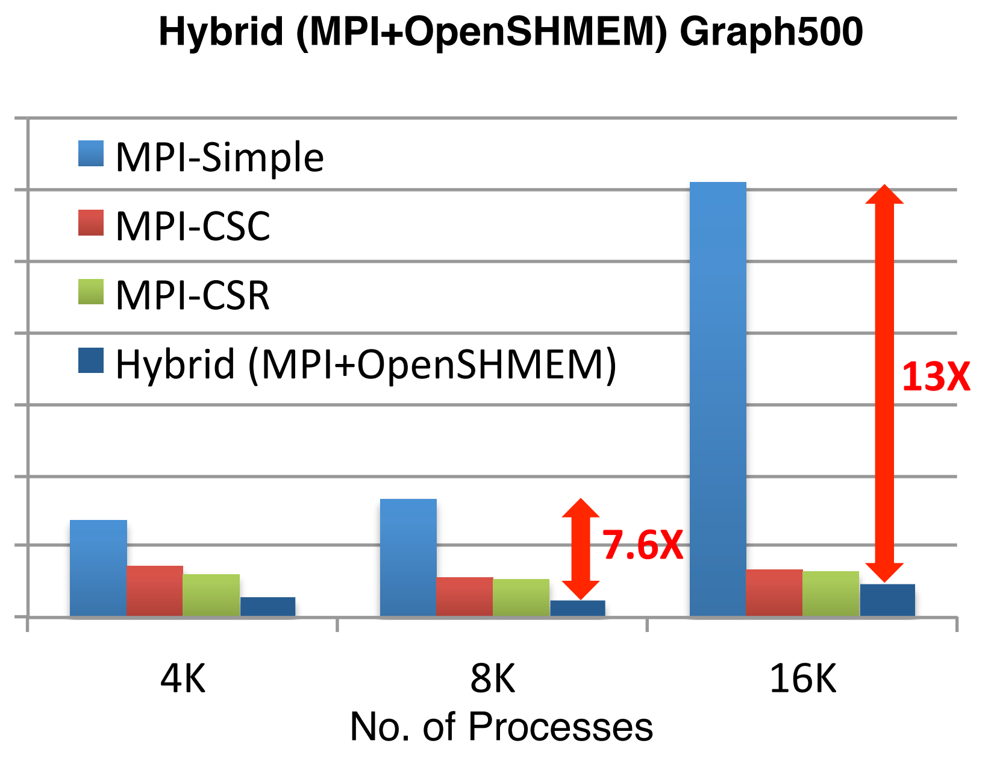
CAF mem_get and mem_put Bandwidth



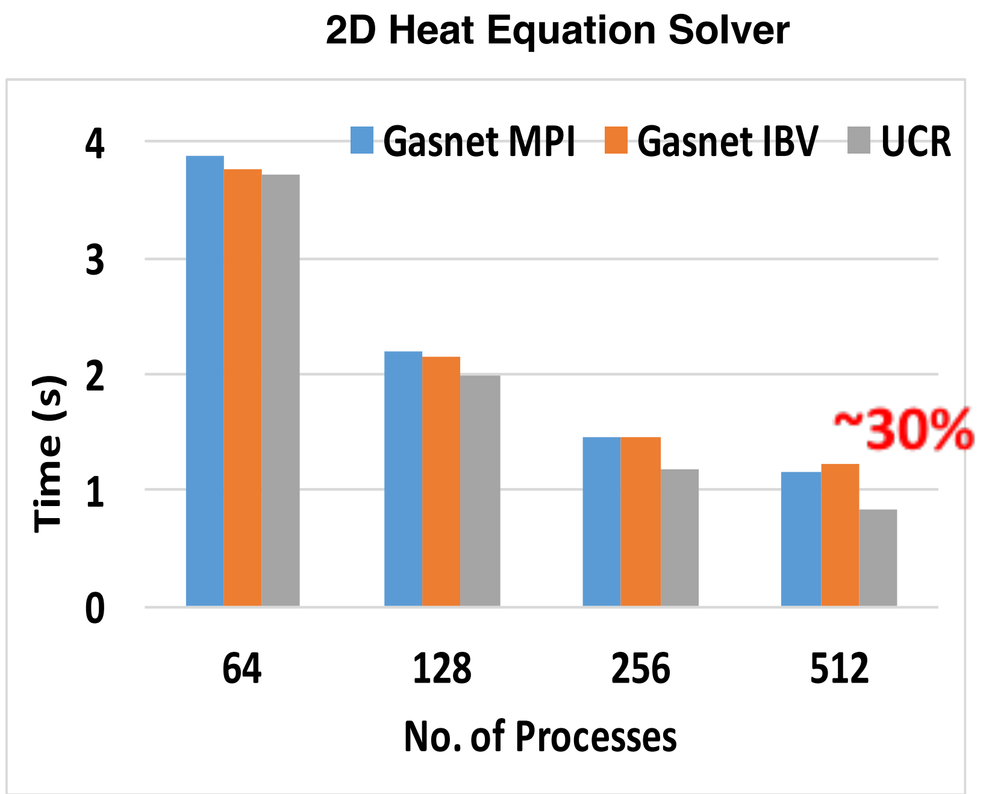


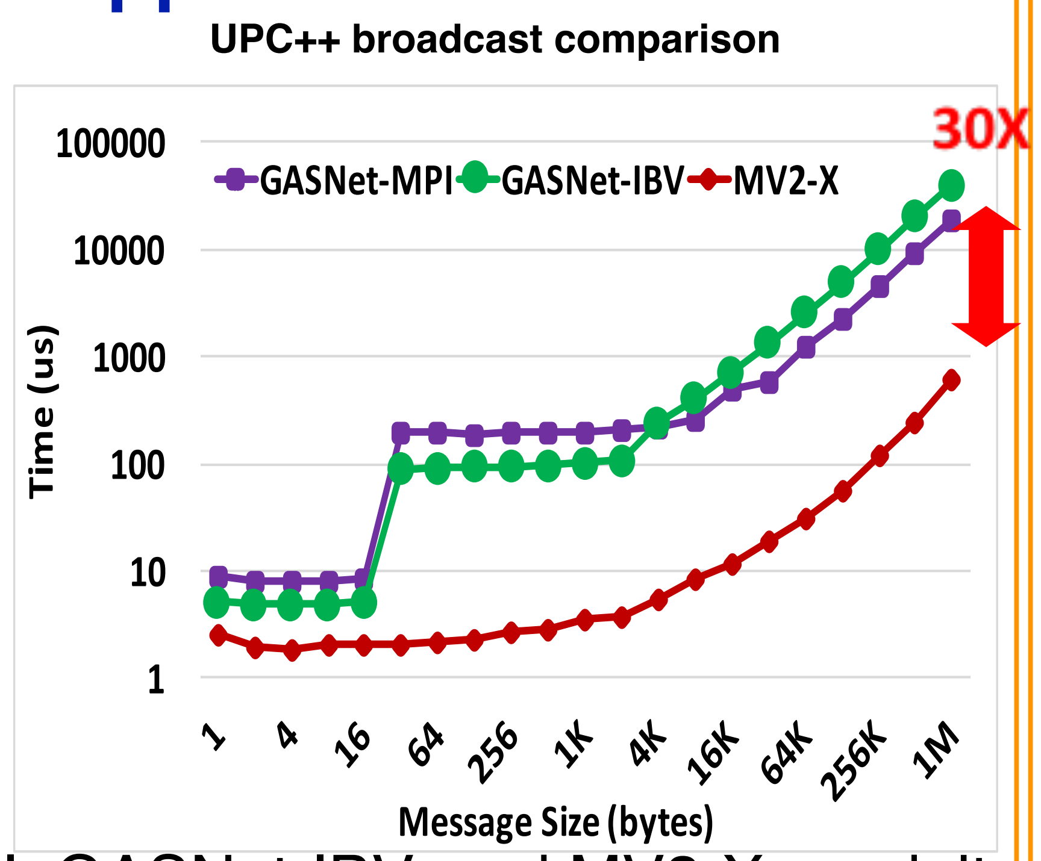
OpenSHMEM Application Evaluation





UPC++ Collectives and Application Evaluation





- UPC++ running on GASNet-MPI, GASNet-IBV, and MV2-X conduits
- Takes advantage of unified runtime and improved collectives to provide better performance

Conclusions

- MVAPICH2-X: Unified Communication Runtime for Hybrid Programming
- Promising: MPI communication not harmed; Better performance for UPC/CAF/UPC++/OpenSHMEM
- Hybrid MPI+OpenSHMEM Graph500 Benchmark: **13X** improvement for 16,384 processes
- Hybrid MPI+UPC FT NAS Benchmark: **34%** improvement for Class-C 128 processes
- UPC++ 2D-Heat with UCR provided **30**% improvement on 512 processes

Publications

- 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)

Acknowledgements

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

MVAPICH

MVAPICH2/MVAPICH2-X: MPI/PGAS over InfiniBand, Omni-Path, Ethernet/iWARP, and 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.