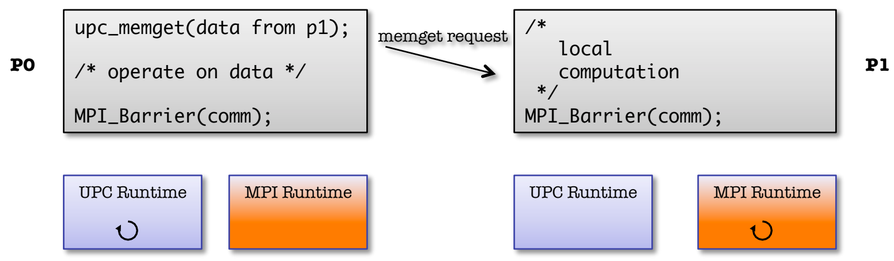
Unifying UPC and MPI Runtimes: Experience with MVAPICH

Jithin Jose, Miao Luo, Sayantan Sur, Dhabaleswar K. Panda
The Ohio State University

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 UPC or MPI) before entering the other

Bad performance!!!

Coercing UPC 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"

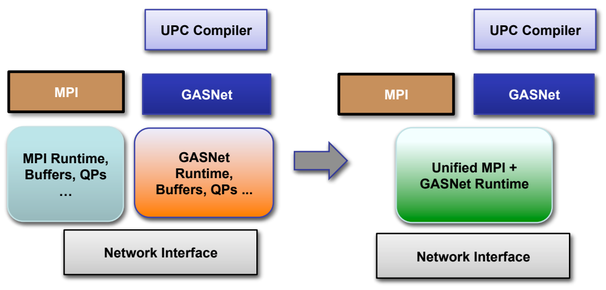
Path forward: unify runtimes, not programming models

Problem Statement

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

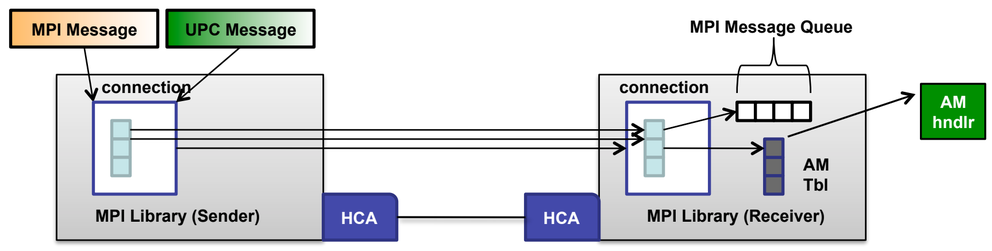
Our Approach

Unified Runtime provides APIs for both MPI & GASNet



- Different AM APIs based on size for optimization
  - Short AM (with/without data payload)
  - Medium AM (bounce buffer using RDMA FP)
  - Large AM (RDMA Put, on-demand connections)
- GASNet Extended interface for efficient RMA
  - Inline Put/Put/Put Bulk/Get (RDMA operations)

Resources shared between MPI & UPC

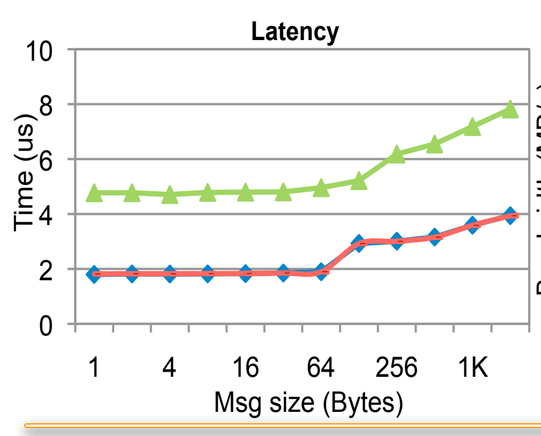


- All resources are shared between MPI and UPC
  - Connections, buffers, memory registrations
  - Schemes for establishing connections (fixed, on-demand)
  - RDMA for large AMs and for PUT, GET

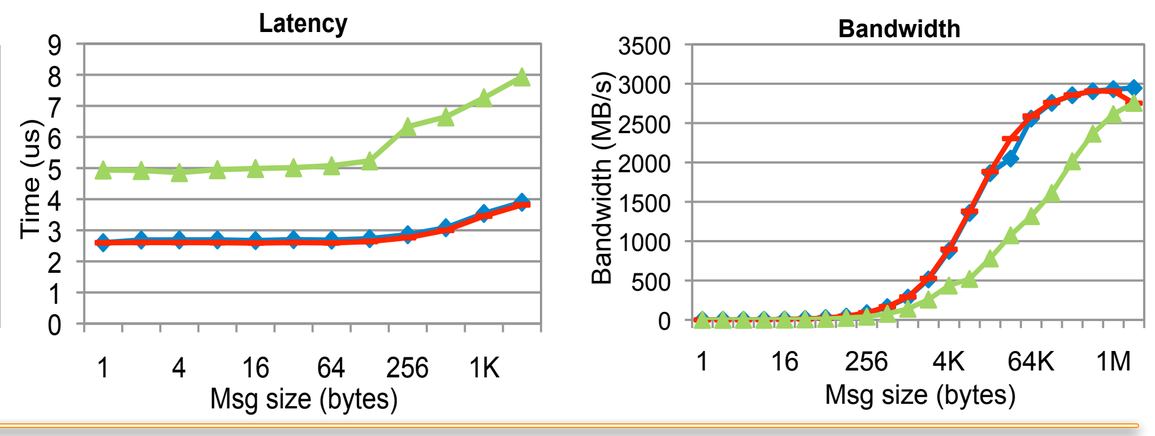
Experimental Results

GASNet-INCR GASNet-IBV GASNet-MPI

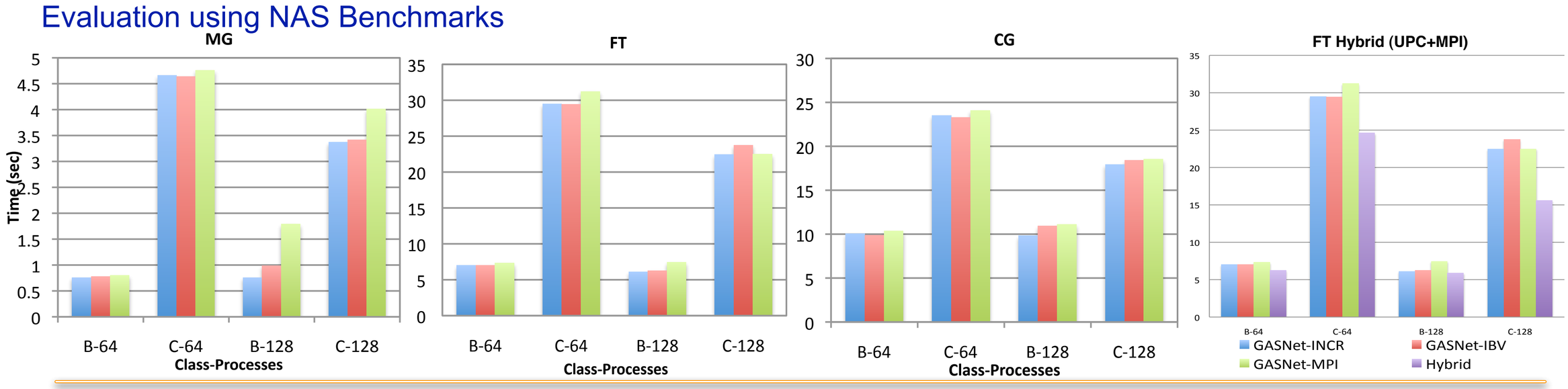
Microbenchmark - upc_memput



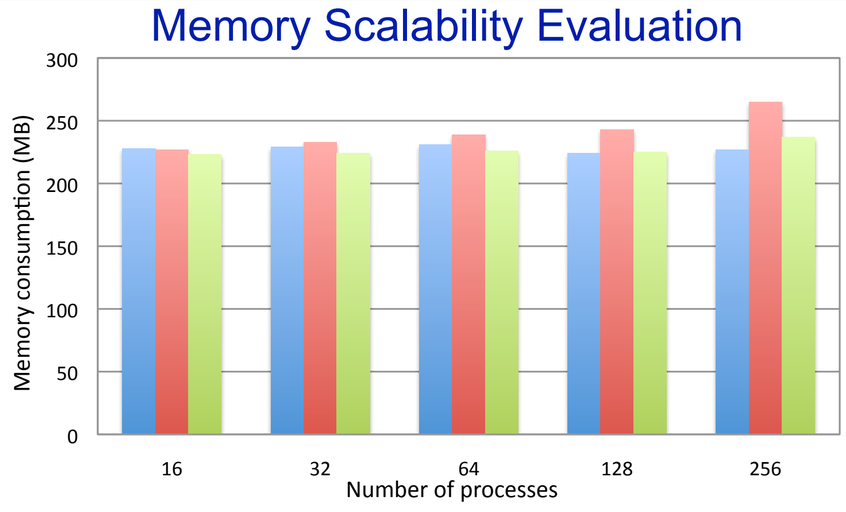
Microbenchmark - upc_memget



Evaluation using NAS Benchmarks



Memory Scalability Evaluation



- GASNet-INCR uses on-demand connection establishment, reducing strict memory requirements.
- GASNet-INCR best scalability due to inherent Aptus design

Conclusions

- GASNet-INCR performs identically with GASNet-IBV in microbenchmarks
- Integrated Communication Runtime (INCR): supports MPI and UPC simultaneously
- Promising: MPI communication not harmed and UPC communication not penalized
- No need for programmer to barrier between UPC and MPI modes, as is current practice
- Pure UPC NAS: 10% improvement CG (B, 128), 23% improvement MG (B, 128)
- MPI+UPC FT: 34% improvement for FT (C, 128)

Publication:

- J. Jose, M. Luo, S. Sur and D. K. Panda, Unifying UPC and MPI Runtimes: Experience with MVAPICH, Fourth Conference on Partitioned Global Address Space Programming Model (PGAS10), Oct. 2010.

Acknowledgements

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

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