Motivation

- Achieves scalability
- Network resources shared among MPI and PGAS (UPC/OpenSHMEM)
- MPI Performance not harmed and UPC/OpenSHMEM/CAF performance
- Consumes lesser network resources

Need for a Unified Runtime

- Dealloc 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!!!

Our Approach - MVAPICHE2-X

- Enables Hybrid (MPI+PGAS) Programming
- Available in MVAPICHE2-X 2.0b Release!

Unified Communication Runtime for MPI and PGAS

- Unified Communication Runtime (UCR) extends MVAPICHE2 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

Experimental Results

Evaluation using UPC-NAS Benchmarks

- NAS MG - UPC Benchmark
- NAS FT Hybrid (UPC-MPI)

Evaluation using CAF-NAS Benchmarks

- NAS CAF Benchmark (Class C)
- NAS CAF Benchmark (Class D)

Application Evaluation

- Broadcast (256 processes) Latency
- 2D FFT
- Out-of-Core Sort

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

Resources shared between MPI and PGAS

- Connections, buffers, memory registrations
- Schemes for establishing connections (fixed, on-demand)
- RDMA for large AMs and for PUT, GET

Conclusions

- Promising: MPI communication not harmed; Better performance for UPC/OpenSHMEM/CAF
- Hybrid MPI+OpenSHMEM/CAF 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, M. Luo, S. Sur and D.K. Panda, Unifying MPI and MPI Runtimes: Experience with MVAPICHE, Partitioned Global Address Space Programming Model (PGAS ’10)  

Acknowledgements

This research is supported in part by National Science Foundation grants OCI-0926601, OCI-1148372 and OCI-1213594.

MVAPICHE2/MVAPICHE2-X: MPI/PGAS over Infiniband,10GE/iWarp & RoCE

http://mvapich.cse.ohio-state.edu/