Co-designing and Accelerating HPC Applications with a Hybrid MPI+PGAS Models

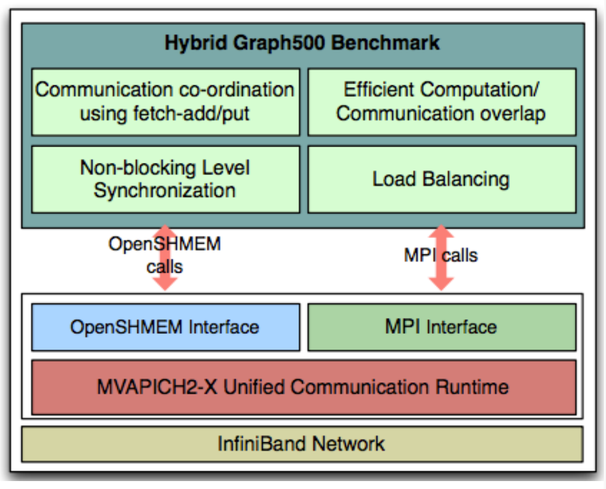
Khaled Hamidouche, Mingzhe Li, and D.K. Panda – The Ohio State University
{hamidouc, limin, panda}@cse.ohio-state.edu

Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models

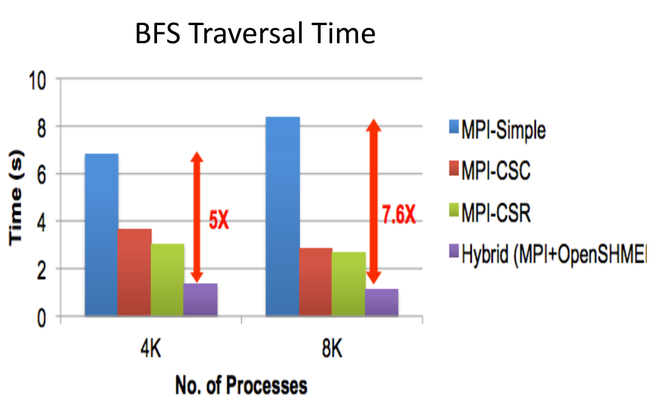
Graph500 represents data intensive and irregular applications that use graph algorithm-based processing methods. It exhibits highly irregular and dynamic communication pattern.

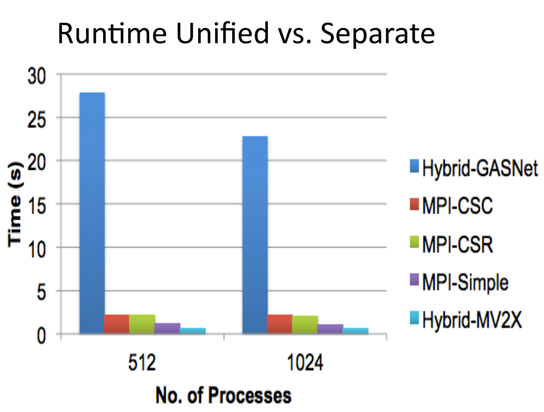
Detailed Design

- One-sided and fetch-add atomic operations for communication and co-ordination
- Buffer Structure for efficient computation-communication overlap
- Non-blocking barrier for level synchronization
- Load balancing



Evalution Results





Publications:

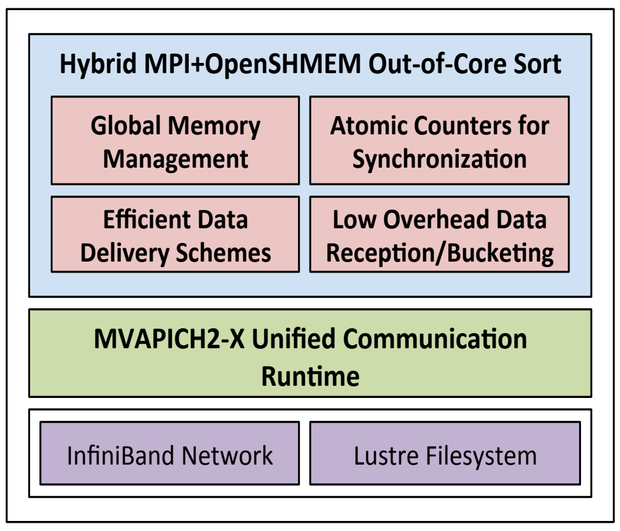
- J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing a Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models, Int'l Super Computing Conference (ISC '13), June 2013

Designing Scalable Out-of-Core Sorting with Hybrid MPI+OpenSHMEM Programming Models

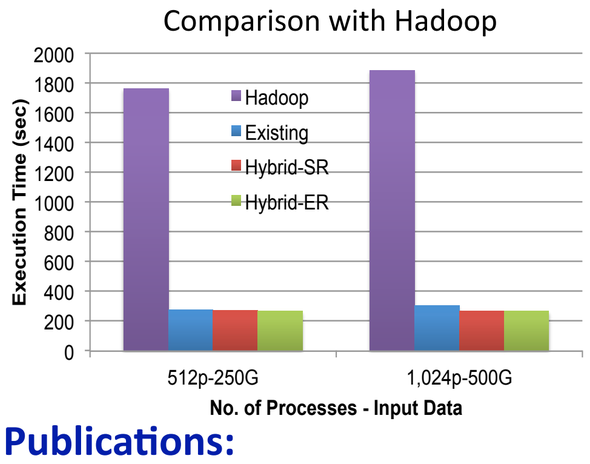
Sort: One of the most common algorithms in data analytics. It reads data from a global filesystem, sort it and write back to global filesystem

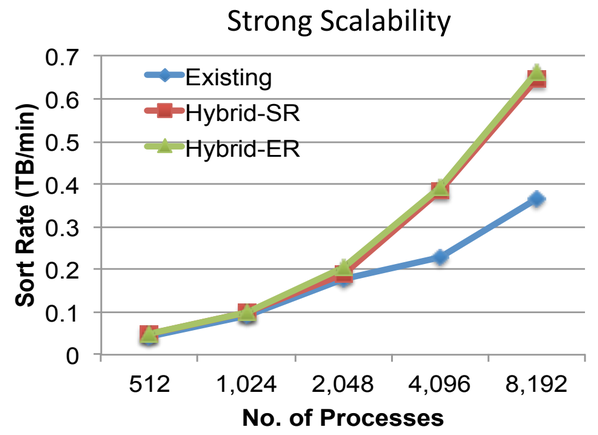
Detailed Design

- Efficient Global Memory View and Memory Management
- Synchronization and Destination Selection using Atomic Counters
- Efficient remote buffer co-ordination and data-delivery using compare-swap and put+notify
- Low overhead Data Reception/ Bucketing in Sort Group



Evalution Results





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+OpenSHMEM Programming Models, Int'l Conference on Partitioned Global Address Space Programming Model (PGAS '14), October.2014

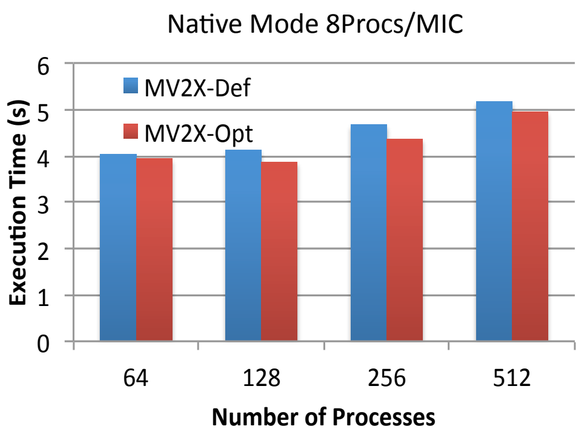
High Performance OpenSHMEM for Xeon Phi Clusters: Extensions, Runtime Designs and Application Co-design

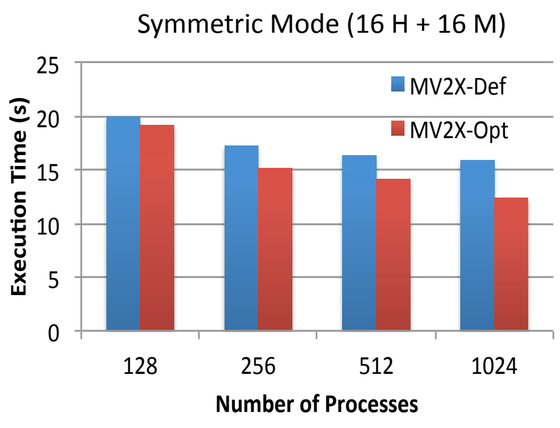
Enhance OpenSHMEM symmetric memory allocation to provide non-uniform memory allocations. Re-design Graph500 to make use of the non-uniform memory allocation and efficiently utilize Xeon Phi Clusters

Redesigning Graph500

- MIC Cores prepare the NewQueue while host processes the CurrQueue
- Newly discovered vertices are placed at owner process' MIC memory
- MIC cores require memory only for holding new vertices
- At the end of each level, NewQueue is transferred to host memory

Evalution Results





Publications:

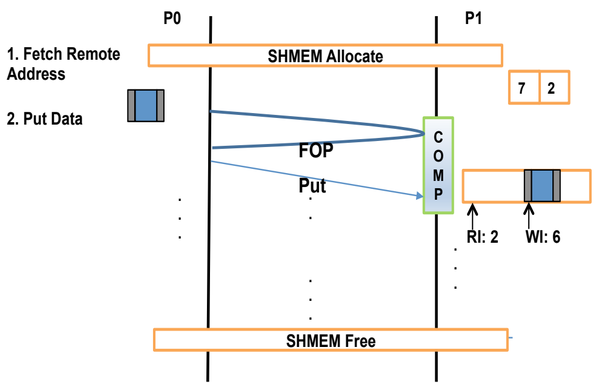
- J. Jose, K. Hamidouche, X. Lu, S. Potluri, J. Zhang, K. Tomko and D. K. Panda, High Performance OpenSHMEM for Intel MIC Clusters: Extensions, Runtime Designs and Application Co-Design, IEEE Int'l Conference on Cluster Computing (CLUSTER '14), September 2014

Scalable MiniMD Design with Hybrid MPI and OpenSHMEM

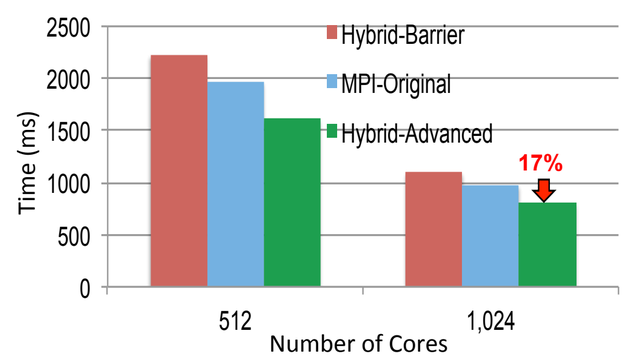
MiniMD is a Molecular Dynamics (MD) mini-application in the Mantevo project at Sandia National Laboratories

Hybrid-Advanced Design

- Better buffer management than Hybrid-Barrier design and remove barrier
- Communication
  Replace MPI Isend/Irecv with one-sided Put communication
- Synchronization
  Use Atomic operation to get target address from target process
  Target process polls on local receive buffer to ensure arrival of completion data



Evalution Results



Publications:

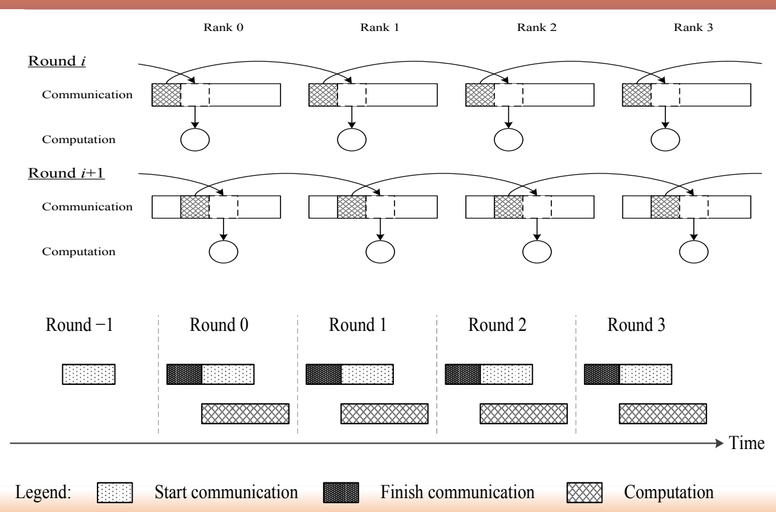
- M. Li, J. Lin, X. Lu, K. Hamidouche, K. Tomko and D. K. Panda, Scalable MiniMD Design with Hybrid MPI and OpenSHMEM, OpenSHMEM User Group(OUG 14), held in conjunction with Eighth Conference on Partitioned Global Address Space Programming Model (PGAS '14), October 2014

Accelerating MaTEx K-NN with Hybrid MPI and OpenSHMEM

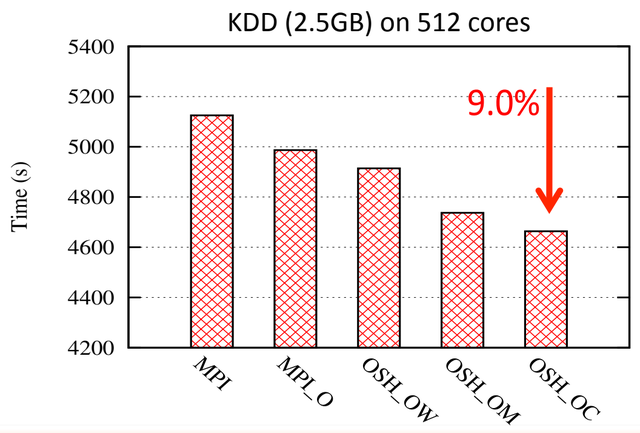
K-Nearest Neighbors (k-NN) algorithm is a popular supervised machine learning algorithm for classification

Redesigning k-NN in MaTEx

- Overlapped Data Flow
- One-sided Data Transfer
- Circular-buffer Structure



Evalution Results



Publications:

- J. Lin, K. Hamidouche, J. Zhang, X. Lu, A. Vishnu, D. Panda. Accelerating k-NN Algorithm with Hybrid MPI and OpenSHMEM, (OpenSHMEM '15) 2015

Acknowledgements

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

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.