Presented at International Supercomputing Conference (ISC), June 2011

MVAPICH2-GPU: Optimized GPU to GPU Communication for InfiniBand Clusters

H. Wang, S. Potluri, M. Luo, A. K. Singh, S. Sur, D. K. Panda

Department of Computer Science and Engineering

The Ohio State University

{wangh, potluri, luom, singhas, ssur, panda}@cse.ohio-state.edu

Abstract

Data parallel architectures, such as General Purpose Graphics Units (GPGPUs) have seen a tremendous rise in their application to the eld of supercomputing. While the GPU architecture provides very high peak ops, communication latencies and moving data in/out of GPUs remain the biggest hurdle to overall performance and programmer productivity. The Message Passing Interface (MPI), is the most popular parallel programming model in use today. Applications executing on a cluster with GPUs have to manage data movement via CUDA library in addition to MPI. Currently, data movement by CUDA and MPI libraries is not integrated, as a result efciencies are lost due to explicit staging of memory buffers. In addition, it is currently not possible to use MPI-2 one sided communication for memory windows in GPU memory. In this paper, we propose a novel MPI design that integrates the CUDA data movement transparently with MPI communication. The result is that the programmer is presented with one MPI interface that can communicate to and from GPU memory. GPU data movement and RDMA data transfer are overlapped with each other inside the MPI library. The proposed design is incorporated into the popular MVAPICH2 library. To the best of our knowledge, this is the rst work of its kind to enable advanced MPI features and optimized pipelining in a widely used MPI library. Up to 45% improvement in one-way latency is observed for message size of 4MB. Further, we show that collective communication performance can be improved signicantly through our novel integration of MPI and CUDA libraries: 32%, 37% and 30% improvement for Scatter, Gather and Allotall collective operations, respectively. In addition, we also enable MPI-2 one sided communication to operate on memory windows on GPUs. MVAPICH2-GPU achieves 45% improvement for Put and Get operations for message size of 4MB.