Presented at IEEE Cluster, September 2011

Optimized Non-contiguous MPI Datatype Communication for GPU Clusters: Design, Implementation and Evaluation with MVAPICH2

H. Wang, S. Potluri, M. Luo, A, K. Singh, X. Ouyang
S. Sur, D. K. Panda
Department of Computer Science and Engineering
The Ohio State University
{wangh, potluri, luom, singhas, ouyangx, 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 for High End Computing. However, data movement in and out of GPGPUs remains the biggest hurdle to overall performance and programmer productivity. Real scientific applications utilize multi-dimensional data. Data in higher dimensions may not be contiguous in memory. In order to improve programmer productivity and to enable communication libraries to optimize non-contiguous data communication, the MPI interface provides MPI datatypes. Currently, state of the art MPI libraries do not provide native datatype support for data that resides in GPU memory. The management of non-contiguous GPU data is a source of productivity and performance loss, because GPU application developers have to manually move the data out of and in to GPUs. In this paper, we present our design for enabling highperformance communication support between GPUs for noncontiguous datatypes. We describe our innovative approach to improve performance by offloading datatype packing and unpacking on to a GPU device, and pipelining all data transfer stages between two GPUs. Our design is integrated into the popular MVAPICH2 MPI library for InfiniBand, iWARP and RoCE clusters. We perform a detailed evaluation of our design on a GPU cluster with the latest NVIDIA Fermi GPU adapters. The evaluation reveals that the proposed designs can achieve up to 88% latency improvement for vector datatype at 4 MB size with micro benchmarks. For Stencil2D application from the SHOC benchmark suite, our design can simplify the data communication in its main loop, reducing the lines of code by 36%. Further, our method can improve the performance of Stencil2D by up to 42% for single precision data set, and 39% for double precision data set. To the best of our knowledge, this is the first such design, implementation and evaluation of non-contiguous MPI data communication for GPU clusters.