Single Level Fast Multipole Method on GPU cluster for Electromagnetic Problems

Vinh Dang*, Quang Nguyen, Ozlem Kilic, and Esam El-Araby,

Department of Electrical Engineering and Computer Sciences
The Catholic University of America, Washington, DC 20064, U.S.A.
13dang@cardinalmail.cua.edu , 93nguyen@cardinalmail.cua.edu , kilic@cua.edu, and aly@cua.edu

Fast Multipole Method (FMM) is a mathematical technique which was developed to seek rapid solutions to integral equations of scattering for Helmholtz problems. For scattering problems, the integral equation is discretized into a matrix equation by the method of moments (MoM). The resultant equation is then typically solved by the direct LU, or an iterative method which requires $O(N^3)$ or $O(N^2)$ floating point operations respectively. However, if FMM is implemented, the complexity is reduced to $O(N^{3/2})$. Moreover, the multilevel fast multipole algorithm (MLFMA) which is a multistage FMM can further reduce the complexity to $O(N\log N)$. These methods are promising for providing a path to large scale computing in electromagnetics.

Due to the rapid advancements of massively parallel architectures in graphics processing units (GPUs) along with their powerful computational capabilities over the past few years, GPUs have been increasingly used to accelerate intensively computational tasks in variety of areas. Furthermore, a recent trend is the application of High Performance Computing (HPC) clusters for large scale problems because they can overcome the memory limitation of GPUs, and take advantage of the conventional CPU-based cluster to achieve more acceleration.

Recently, many authors have investigated the parallelization of FMM and MLFMA approaches on CPU cluster to solve real-life problems which lead to matrix equation of millions of unknowns. However, to the best of our knowledge, FMM and MLFMA have not been studied much for electromagnetic scattering problems on GPU clusters. Only a few authors have applied single level FMM to solve acoustic scattering on multi-GPU systems. While MLFMA has a lower complexity, it reveals a poor scalability compared with FMM in distributed memory computer system because of its requirement of a sophisticated load balancing scheme. Consequently, FMM becomes more appropriate for parallel.

In this paper, the exploitation of an HPC GPU cluster to implement single level fast multipole method for large scale scattering problems is investigated. The NVIDIA's Compute Unified Device Architecture **(**CUDA) and MVAPICH2 programming environments are utilized on a 13-node cluster equipped with GPUs and interconnected through a high bandwidth, low latency Infiniband network. A load-balancing strategy is carefully taken into consideration to distribute computational tasks such as near-field interactions, aggregation, translation and disaggregation into computing nodes, and to minimize communication overheads. A combined solution of MVAPICH2 and Pthreads models is proposed to maximize the system resource utilization and to accommodate for higher data sets. The experimental results of canonical problems as well as real-life problems involving complicated shapes demonstrate the acceleration achieved by the HPC GPU cluster in comparison with a conventional CPU cluster. The results also show the accuracy and favorable scalability features of the proposed GPU cluster parallelization of single level FMM.