Performance Evaluation of Accurate Matrix-matrix Multiplications on GPU Using Sparse Matrix Multiplications

Fumiya Ishiguro
Graduate School of Informatics
Nagoya University
Japan
ishiguro@hpc.itc.nagoya-u.ac.jp

Takahiro Katagiri
Information Technology Center
Nagoya University
Japan

Satoshi Ohshima
Information Technology Center
Nagoya University
Japan

Toru Nagai
Information Technology Center
Nagoya University
Japan

Basic Linear Algebra Subprograms (BLAS) is a frequently used numerical library for linear algebra computations. However, it places little emphasis on computational accuracy, especially with respect to the accuracy assurance of the results. Therefore, ensuring the accuracy of the computational results of BLAS operations is a crucial challenge. Studies for ensuring computational accuracy are based on a rounding concept called faithful rounding [1]. Faithful rounding algorithms perform rounding based on the upper or lower bound as the rounding point for accurate results [1]. Using the concept of faithful rounding, we can achieve "assured" matrix-matrix multiplication (MMM), which was proposed by Ozaki et al.[2][3]. Hereafter, we refer to such assured MMM as the Ozaki method.

We need to consider high-performance implementations of the Ozaki method to adapt it to advanced computational environments, including supercomputers. Current computers are based on parallel computing, especially the use of multi-core CPUs. Hence, thread-level parallelization should be taken into account. Thus far, few studies have attempted to adapt thread-level parallelization of the Ozaki method. Therefore, we have proposed a threaded implementation of the Ozaki method in our previous work [4]. We utilized the various implementations of the Ozaki method on a GPU in this study.

In this study, we contribute the following two points: (1) We evaluate the Sparse Matrix-Dense Matrix multiplication (SpMxDM) implemented on GPU with the property of allowing dense matrices to be transformed into sparse matrices during the algorithm; (2) We evaluate Sparse Matrix-Matrix Multiplications (SpMM) with Sparse Matrix-Sparse Matrix Multiplications (SpMxSpM).

Results with the Reedbush-H supercomputer system at the Information Technology Center at the University of Tokyo indicated that when the input matrix is large, the efficiency of computations by the Ozaki method can be improved by the GPU environment. In particular, (1) the implementation of SpMV with the CRS format achieved a 3.24-times speedup and the ELL format achieved a 2.03-times speedup on the GPU compared with a CPU. (2) The implementation of SpMxSpM achieved a maximum of 8.44-times speedup compared to SpMM.

The results of the SpMM and SpMxSpM were compared with the case of the Ozaki method calculated with dense **dgemm** when the matrix B was a zero matrix with only one element inserted. When the size was smaller than N = 500, SpMM provided a shorter execution time than **dgemm**. When the size was larger than N = 100, SpMxSpM provided a shorter execution time than **dgemm**.

ACKNOWLEDGMENTS

This research was supported by MEXT under "Exploratory Issue on Supercomputer Fugaku" (Development of Verified Numerical Computations and Super High-performance Computing Environment for Extreme Researches) and "Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures" in Japan (Project ID: jh170011-DAJ). This work was also supported by JSPS KAKENHI Grant Number JP18K19782.

REFERENCES

[1] S. M. Rump, T. Ogita and S. Oishi. Accurate floating-point summation Part I: Faithful Rounding. SIAM J. Sci. Comput., 31(1), (2008), 189–224.

[2] K. Ozaki, T. Ogita and S. Oishi. Tight and efficient enclosure of matrix multiplication by using optimized BLAS. Numer. Linear Algebra Appl., 18(2), (2011), 237–248.

[3] K. Ozaki, T. Ogita, S. Oishi and S. M. Rump (2012). Error-free transformations of matrix multiplication by using fast routines of matrix multiplication and its applications. Numer. Algorithms, 59(1), 95–118.

[4] S. Ichimura, T. Katagiri, K. Ozaki, T. Ogita and T. Nagai. Threaded Accurate Matrix-Matrix Multiplications with Sparse Matrix-Vector Multiplications. In Proc. of the Parallel and Distributed Processing Symposium Workshop (IPDPSW2018), IEEE International, (2018), 1093–1102.