Reduced/Extended/Mixed-precision matrix computations on FPGAs and GPGPUs)

Imamura Toshiyuki
RIKEN Center for Computational Science

１. Project Purpose

This study aims to explore the basic linear algebra operations using reduced/ extended-precision formats such as 16-bit, 24-bit, 96-bit, and 128-bit on floating-point and fixed-point operations. Specifically, there are the following topics to be studied.

(1) Data representation: we explore arbitrary precision representation methods that can be handled efficiently and easily on both general processors (CPU/GPU) and FPGA and are also oriented toward mixed-precision computations.

(2) Arithmetic and matrix operations: we will study the IPs of basic floating-point operations with reduced/extended precision on FPGAs, and then develop FPGA-based acceleration on matrix operations using the IPs. At the same time, we will also investigate high performance emulation methods on CPUs/GPUs and compare them with FPGAs to explore the advantages of each hardware platform. Specifically for matrix multiplication on CPUs/GPUs, we will extend the Ozaki scheme-based approach that we developed in FY2021.

(3) Memory-bound and sparse operations: These operations have completely different properties from dense matrix multiplication. This year, we will study the data representation in sparse operation and the performance of arbitrary/mixed-precision Conjugate Gradient solvers on a CPU, a GPU, and an FPGA by extending the solvers we developed in FY2020. And their performance will be evaluated and compared.

２. Results

This year, we have conducted various enhancements of high-accuracy matrix product routines based on the Ozaki scheme and realized a preliminary study to ensure the reproducibility of the CG method. We have achieved high performance by leveraging the capabilities of GPUs; in particular, we have demonstrated an affinity with the TensorCore engine. The results were presented at major conferences such as [1][2][3][4]. Also, experiments on the implementation of BLAS with low-precision and integer arithmetic, which is the basis of mixed-precision computing, have progressed, and applications using the SYMV kernel on a Nvidia GPU are also in progress [5].

3. Roles of the MCRP and its significance

MCRP provides the atop hardware system Cygnus, which becomes a platform that allows us to combine the FPGA technology we require with conventional technologies such as CPUs and GPUs. It is also a place to find opportunities for future collaborative applications between other fields to become a significant example of the cross-field science tightened by computer science. These significances are different from those of other supercomputer public offering projects and can be highly appreciated.

4. Future plan

As pointed out in the FY2021 results, several themes could not be finalized. Thus, we would like to correspond to the pending issues. However, due to a revision of the technical substance and the relocation of the members of the project, FY2022 will be temporarily difficult to implement; the proposal will be suspended for FY2022, and the project will be restarted in a new form in FY2023 or later. We will then work to enhance the computational concepts that will be accelerated by FPGA technology.

5. Publications and conference presentations

(1) Journal papers

[1] Accurate Matrix Multiplication on Binary128 Format Accelerated by Ozaki Scheme, Proc. The 50th International Conference on Parallel Processing (ICPP-2021), Vol. 78, 2022/1/11, https://doi.org/10.1145/3472456.3472493, 2021, Daichi Mukunoki, Katsuhisa Ozaki, Takeshi Ogita, Toshiyuki Imamura

(2) Presentations

[2] A Fast Infinite Precision Inner Product using Ozaki Scheme and Dot2, and Its Application to Reproducible Conjugate Gradient Solvers, ISC High Performance (ISC 2022), 2022, Hamburg, Daichi Mukunoki, Katsuhisa Ozaki, Takeshi Ogita, Toshiyuki Imamura

[3] Accurate Matrix Multiplication on Binary128 using Ozaki Scheme, ISC High Performance (ISC 2021), 2021, Online, Daichi Mukunoki, Katsuhisa Ozaki, Takeshi Ogita, Toshiyuki Imamura

[4] Impact and Contribution of Ozaki scheme in High Performance Computing, International Workshop on Reliable Computing and Computer-Assisted Proofs

(ReCAP 2022), 2022/03/15, Online, Daichi Mukunoki, Katsuhisa Ozaki, Takeshi Ogita, Toshiyuki Imamura, Roman Iakymchuk

[5] CP-ALS についての HPC からの考察と試み, 日本応用数理学会 2021 年度年会, 2021/09/09, Online, 今村俊幸

(3) Others

Supercomputer	Use	Allocated resources*	
		Initial resources	Additional resources
Cygnus	Yes	10000	0
Oakforest-PACS	No	0	0
*in units of node-hour product			