Reduced/extended/Mixed-precision matrix computations on FPGAs and GPGPUs

Toshiyuki Imamura
RIKEN Center for Computational Science

１．Project Purpose

This research aims to investigate the matrix computation based on the reduced/extended-precision arithmetic such as 8-bit, 16-bit, 24-bit, 48-bit, 96-bit, and 128-bit for the fixed-point and the floating-point numbers, that include non-builtin data format, and where CPUs/GPUs/FPGAs supports hardware-oriented variable numerical precision, flexibly.

Following kickoff in 2019, we have started the project from a typical linear algebra operation such as matrix multiplication and then extend it to solve other higher-level matrix problems. Since we have kicked off the project, our research objectives are shown further as follows.

(1) Data formats of reduced/extended-precision on FPGA and GPU. Different data representation formats of the reduced/extended-precision are studied, and their performance in matrix multiplication is evaluated on single or multiple or hybrid of FPGAs and GPUs, concerning hardware resource utilization, data throughput, and so on.

(2) Exploration of design space. The related algorithms and performance tuning techniques of matrix computation with the reduced/extended-precision will be researched. The corresponding systems are implemented by using FPGAs and GPUs and evaluated on the supercomputer Cygnus. Through comparisons of their performance, we explore the design space in the Cygnus, including the coordination of FPGA and GPU, data communication, mixed programming on FPGA.

(3) Building kernels and IP libraries of matrix computation for GPGPU and FPGA are conducted, and they are optimized based on the Cygnus.

２．Results

Although this year was supposed to be the year to overcome the shortcomings of the project, it was not easy to devote enough time to the project due to the impact of Covid19. For FPGAs, we could not proceed with implementing floating-point operations as planned, but we have continued to promote the verification of mixed-precision operations using GPUs as the central platform, as we did in FY2019. As the results, we

presented FP32 and FP64 emulation using a V100 Tensor Cores, and FP128 operations at ISC20 [1] and HPC Asia21 [2]. In addition, we have achieved a fixed-point implementation in the SYMV kernel[3].

３．Roles of the MCRP and its significance

Like the last year, 2019, MCRP provides the atop hardware system Cygnus, a platform that allows us to combine the FPGA technology we require with conventional technologies such as CPUs and GPUs. It is also an excellent place to share collaborative applications to become a significant example of the cross-field science tighten by computer science. The significance is unique and highly appreciated.

４．Future plan

Due to some restrictions by Covid19, we could not complete some of the planned topics in FY2020. These are anticipated to be relieved via our one-year-experiences. We must overcome these circumstances. Untouched items, for example, implementation and evaluation of flexible FP-units on an FPGA, are our short-ranged milestones during FY2020-2022, and must be conducted in FY2021. We are going to enhance our computational concepts accelerated by FPGA technologies.

５．Publications and conference presentations

(1) Journal papers

[1] Mukunoki D., Ozaki K., Ogita T., Imamura T., DGEMM Using Tensor Cores, and Its Accurate and Reproducible Versions, Lecture Notes in Computer Science 12151, pp. 230-248, doi:10.1007/978-3-030-50743-5_12, 2020

[2] Daichi Mukunoki, Katsuhisa Ozaki, Takeshi Ogita, Roman Iakymchuk, Conjugate Gradient Solvers with High Accuracy and Bit-wise Reproducibility between CPU and GPU using Ozaki scheme, The proceedings of HPC Asia 2021, The International Conference on High Performance Computing in Asia-Pacific Region, pp. 100-109, doi:10.1145/3432261.3432270, 2021

(2) Presentations

None

(3) Others

[3] Ope Source publication, ASPEN.K2 version 1.8, newly supports integer SYMV kernels; i128symv, i64symv, i32symv, and i16symv. https://www.r-

ccs.riken.jp/labs/lpnctrt/projects/aspen-k2/index.html

Supercomputer	Use	Allocated resources*	
		Initial resources	Additional resources
Cygnus	Yes	5000	0
Oakforest-PACS	No	0	0
*in units of node-hour product			