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, where CPUs/GPUs/FPGAs supports hardware-oriented variable numerical precision, flexibly. We will start from a typical linear algebra operation such as matrix multiplication, and then extend to solve other higher-level matrix problems. The research objectives are shown as follows in detail.

(1) Data formats of reduced/extended-precision on FPGA and GPU. Different data representation formats of the reduced/extended-precision will be studied, and their performance in matrix multiplication will be evaluated on FPGA and GPU, such as 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 will be implemented by using FPGAs and GPUs, and evaluated on the supercomputer Cygnus. Through comparisons of their performance, the design space in the Cygnus is explored, including the coordination of FPGA and GPU, data communication, mixed programming on FPGA, and so on.

(3) Building kernels and IP libraries of matrix computation. The kernels and IP libraries of matrix computation for GPGPU and FPGA will be developed and optimized based on the Cygnus.

2. Results

Since this year was the initial phase of our project, we spent much time on the startup process and learned a new FPGA environment. Mixed-precision computing is one of the main-waves in computer and computation science as well as the artificial intelligence field. This year, we have proposed two significant computational concepts of *'minimal precision computing'* and *'weak-numerical*

reproducibility.' They are initiated under the collaboration between this MCRP project group and other researchers from RIKEN CCS and LIP6 Sorbonne University (Paris, France). These technologies can be realized by variable-precision floating-point arithmetic and randomness of the routing mode by reconfigurable devices such as an FPGA. We have sought first to overcome these demands with FPGAs, but unfortunately, this year, we were not able to utilize them fully. On the other hand, we have realized various implementations that can be connected in the future, such as high-precision calculation by built-in data type supported by CPU and emulation of Double Precision calculation by using Tensor Cores on a V100.

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 tighten 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 in the result of the initial year, we could not initiate some of the themes. Thus, we would like to cope with the untouched items listed upon our initial plan. Then, we will enhance our computational concepts accelerated by FPGA technologies.

5. Publications and conference presentations

(1) Journal papers

[1] DGEMM using Tensor Cores and Its Accurate and Reproducible Versions, Lecture Notes in Computer Science (in printing), Daichi Mukunoki, Katsuhisa Ozaki, Takashi Ogita, and Toshiyuki Imamura

(2) Presentations

[1] Numerical Reproducibility based on Minimal-Precision Validation, Computational Reproducibility at Exascale Workshop (CRE2019), in cooperation with SC19, 2019, Nov. 17, Denver, USA, Toshiyuki Imamura, Daichi Mukunoki, Fabienne Jézéquel, Stef Graillat, Roman Iakymchuk

[2] Minimal-Precision Computing for High-Performance, Energy-Efficient, and Reliable Computations, SIAM Conference on Parallel Processing for Scientific

Computing (PP19), 2020 Feb. 15, Seattle, USA, Daichi Mukunoki

[3] Precision Tuning of the Arithmetic Units in Matrix Multiplication on FPGA, SIAM Conference on Parallel Processing for Scientific Computing (PP20) 2020, Feb. 15, Seattle, USA, Toshiyuki Imamura and Yiyu Tan

[4] Reduced and Extended-precision Computations on FPGAs and GPUs,the 11th Symposium on Discovery, Fusion, Creation of New Knowledge by Multidisciplinary Computational Sciences2019, Oct. 15, Tsukuba, Japan, Yiyu Tan, Daichi Mukunoki, Toshiyuki Imamura, Norihisa Fujita, Taisuke Boku

[5] Minimal-Precision Computing for High-Performance, Energy-Efficient, and Reliable Computations, France-Japan-Germany trilateral workshop: Convergence of HPC and Data Science for Future Extreme Scale Intelligent Applications, 2019, Nov. 7, Tokyo, Japana, Daichi Mukunoki, Toshiyuki Imamura, Yiyu Tan, Atsushi Koshiba, Jens Huthmann, Kentaro Sano, Fabienne Jézéquel, Stef Graillat, Roman Iakymchuk, Norihisa Fujita, Taisuke Boku

[6] Minimal-Precision Computing for High-Performance, Energy-Efficient, and Reliable Computations, SC19, The International Conference for High Performance Computing, Networking, Storage, and Analysis, 2019, November 17-22, Denver, USA, Daichi Mukunoki, Toshiyuki Imamura, Yiyu Tan, Atsushi Koshiba, Jens Huthmann, Kentaro Sano, Fabienne Jézéquel, Stef Graillat, Roman Iakymchuk, Norihisa Fujita, Taisuke Boku

[7]Minimal-Precision Computing for High-Performance, Energy-Efficient, and Reliable Computations, 2nd R-CCS International Symposium, 2020, Feb. 17-18, Nichii Gakkan Kobe Port Island Center & RIKEN R-CCS, Kobe, Japan, Toshiyuki Imamura, Daichi Mukunoki, Yiyu Tan, Atsushi Koshiba, Jens Huthmann, Kentaro Sano, Fabienne Jézéquel, Stef Graillat, Roman Iakymchuk, Norihisa Fujita and Taisuke Boku

[8] Overview of minimal-precision computing and (weak)-numerical reproducibility, Workshop on Largescale Parallel Numerical Computing Technology (LSPANC 2020 January), 2020, Jan. 30, RIKEN R-CCS, Kobe, Toshiyuki Imamura

[9] Precision Tuning of the Arithmetic Units in Matrix Multiplication on FPGA, Workshop on Largescale Parallel Numerical Computing Technology (LSPANC 2020 January) 2020, Jan. 30, RIKEN R-CCS, Kobe, Yiyu Tan

(3) Others

Supercomputer	Use	Allocated resources*	
		Initial resources	Additional resources
Cygnus	Yes	800	0
Oakforest-PACS	No	0	0
*in units of node-hour product			