Robust, and Energy-Efficient Computations

In numerical computations, the precision of floating-point calculations is a key factor to determine the performance (speed and energy-efficiency) as well as the reliability (accuracy and reproducibility). However, the precision generally plays a contrary role for both. Therefore, the ultimate concept for maximizing both at the same time is the optimized/reduced precision computation through precision-tuning, which adjusts the minimal precision for each operation and data. Several studies have been already conducted for it so far, but the scope of those studies is limited to the precision-tuning alone. Instead, we propose a more broad concept of the optimization/minimal precision and robust (numerically reliable) computing with precision-tuning, involving both hardware and software stack.

Our Proposal

Minimal-Precision Computing

Minimal-precision computing is both reliable (aka robust and sustainable) and energy-efficient. It enables us to compute the required accuracy of the result as well as energy-efficiency.

High-performance

Performance can be improved through the minimal-precision as well as fast numerical libraries and accelerators.

Energy-Efficient

Through use of minimal-precision as well as the energy-efficient hardware acceleration with FPGA and GPU.

Robust (Numerically Reliable)

To enable the minimal-precision computing is processed based on numerical validation, guaranteeing also reproducibility.

General

Our scheme is applicable for any floating-point computations. It contributes to low development cost and sustainability (easy maintenance and system portability).

Comprehensive

We propose a total system from the precision-tuning to the execution of the tuned code, combining heterogeneous hardware and hierarchical software stack.

Realistic

Our system can be realized by combining available in-house technologies.

Our Contributions

Stochastic Arithmetic Tools

Discrete Stochastic Arithmetic (DSA) [14][15] enables us to perform statistical arithmetic [16] i.e., the number of correct significant digits (SCDs), by executing the code as probability computation on FPGA and GPU. As a result, DSA is a general scheme applicable for any floating-point computations on heterogeneous accelerators and no code modification is necessary. This paradigm consists of several advantages, including: (1) new accuracy evaluation in performance, scalability, and development cost compared to the other numerical validation / validation methods.

FPGA as an Arbitrary-Precision Computing Platform

FPGA enables us to implement arbitrary precision on hardware high-level synthesis (VLSI) enables to program it in (C++) language. However, to improve the accuracy and efficiency of very high-precision computation is still challenging. Heterogeneous computing with FPGA & GPUs is also a challenge.

Fast and Accurate Numerical Libraries

Arbitrary precision arithmetic is performed using MPFR or CRUN, but the performance is very low. To overcome it, we are developing a new numerical library: the Fast and Accurate Numerical Libraries (FAN). The library is based on double-precision (Doub-precision) arithmetic, and it is combined with the optimized BLAS (QPBLAS). They support distributed environments with High Performance/Low-Cost (HPC/LC) schemes.

Conclusion & Future Work

We proposed a new systematic approach for minimal-precision computations. This approach is robust, general, comprehensive, high-performance, and realistic. Although the proposed system is still in development, it can be constructed by combining already available (developed) in-house technologies and extending them. Our ongoing step is to demonstrate the system on a proxy application.