Reproducible and Accurate Matrix Multiplication for High-Performance Computing

Sylvain Collange, David Defour, Stef Graillat, and Roman Iakymchuk

INRIA – Centre de recherche Rennes – Bretagne Atlantique
Campus de Beaulieu, F-35042 Rennes Cedex, France
sylvain.collange@inria.fr
DALI–LIRMM, Université de Perpignan
52 avenue Paul Alduy, F-66860 Perpignan, France
david.defour@univ-perp.fr
Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6
F-75005 Paris, France
CNRS, UMR 7606, LIP6, F-75005 Paris, France
Sorbonne Universités, UPMC Univ Paris 06, ICS
F-75005 Paris, France
{stef.graillat, roman.iakymchuk}@lip6.fr

Keywords: Matrix multiplication, reproducibility, accuracy, long accumulator, multi-precision, multi- and many-core architectures.

The increasing power of current computers enables one to solve more and more complex problems. This, therefore, requires to perform a high number of floating-point operations, each one leading to a round-off error. Because of round-off error propagation, some problems must be solved with a longer floating-point format.

As Exascale computing (10^{18} operations per second) is likely to be reached within a decade, getting accurate results in floating-point arithmetic on such computers will be a challenge. However, another challenge will be the reproducibility of the results – meaning getting a bitwise identical floating-point result from multiple runs of the same code – due to non-associativity of floating-point operations and dynamic scheduling on parallel computers.

Reproducibility is becoming so important that Intel proposed a "Conditional Numerical Reproducibility" (CNR) in its MKL (Math Kernel Library). However, CNR is slow and does not give any guarantee concerning the accuracy of the result. Recently, Demmel and Nguyen [1] proposed an algorithm for reproducible summation. Even though their algorithm is fast, no information is given on the accuracy.

More recently, we introduced [2] an approach to compute deterministic sums of floating-point numbers efficiently and with the best possible accuracy. Our multi-level algorithm consists of two main stages: filtering that relies upon fast vectorized floating-point expansions; accumulation which is based on superaccumulators in a high-radix carry-save representation. We presented implementations on recent Intel desktop and server processors, on Intel Xeon Phi accelerator, and on both AMD and NVIDIA GPUs. We showed that the numerical reproducibility and bit-perfect accuracy can be achieved at no additional cost for large sums that have dynamic ranges of up to 90 orders of magnitude by leveraging arithmetic units that are left underused by standard reduction algorithms.

In this talk, we will present a reproducible and accurate (rounding to the nearest) algorithm for the product of two floating-point matrices in parallel environments like GPU and Xeon Phi. This algorithm is based on the DGEMM implementation. We will show that the performance of our algorithm is comparable with the classic DGEMM.

References

[1] J. Demmel, H. D. Nguyen, Fast Reproducible Floating-Point Summation, *Proceeding of the 21st IEEE Symposium on Computer Arithmetic*, Austin, Texas, USA (2013), pp. 163-172.

[2] S. Collange, D. Defour, S. Graillat, R. Iakymchuk, Full-Speed Deterministic Bit-Accurate Parallel Floating-Point Summation on Multi- and Many-Core Architectures, *Research Report*. HAL ID: hal-00949355. February 2014.