Accurate and Reproducible BLAS Routines with Ozaki Scheme for Many-core Architectures

Daichi Mukunoki[1], Takeshi Ogita[2], Katsuhisa Ozaki[3]
[1]RIKEN Center for Computational Science, Hyogo, Japan
[2]Tokyo Woman's Christian University, Tokyo, Japan
[1]Shibaura Institute of Technology, Saitama, Japan
daichi.mukunoki@riken.jp

As floating-point computations suffer from rounding errors, the result may be inaccurate and not identical (non-reproducible). In particular, heterogeneous computing has many factors to disturb reproducibility. The loss of accuracy and reproducibility may become a crucial issue on the debugging of complex codes as well as the reliability of computations. This paper proposes high-performance implementations of reproducible Basic Linear Algebra Subprograms (BLAS) routines with tunable accuracy for many-core architectures. Our approach is based on an accurate matrix-multiplication method, Ozaki scheme, which can be constructed on existing level-3 BLAS which performs standard floating-point operations. In this paper, we show the performance of three routines, inner product (DOT), matrix-vector multiplication (GEMV), and matrix-multiplication (GEMM) on a Volta GPU through the comparison with the standard routines provided by the vendor. Besides, we demonstrate the reproducibility between CPU and GPU and accuracy.

Keywords: accurate, reproducible, BLAS, many-core, GPU.