Design towards Modern High Performance Numerical LA Library Enabling Heterogeneity and Flexible Data Formats

Toshiyuki IMAMURA a,1 Daichi MUKUNOKI a Yusuke HIROTA a
Susumu YAMADA b Masahiko MACHIDA b

a RIKEN, Advanced Institute for Computational Science
b CCSE Japan Atomic Energy Agency

Abstract. This work introduces a design proposal towards modernization of high performance numerical library enabling various parallel execution, such as offloading, many-core concurrent execution, heterogeneous hybrid execution. A prototype implementation of Eigen-G2 employs a parallel task model supported by OpenMP, and exclusive control of offloading a GPU device. Also, multiple data formats are available by taking advantage of a metaprogramming support of C++ language. Eigen-G2 exhibits good parallel performance scalability when we use both multi-core CPUs and a GPU at the same time.

Keywords. modernization of numerical library, task parallel model, OpenMP 4.x, management of multiple data format, eigensolver

Introduction

Recent advancement of computer architecture yields a tremendous diversity and leads to emerging demands of a variety of numerical libraries such for distributed memory, thread parallelism, offloading to GPUs and FPGA accelerator boards. Furthermore, programming style and execution model expose the needs of flexible numerical kernels, such as the batched-style BLAS kernels, which were recently proposed to enable concurrent fine-grain execution for multiple sequences of small to large problems. Therefore, kernel developers need to develop a machine-specific version for each hardware architecture due to the practical reason of performance tuning and run-time environment. This results in complex library packages, and many discussions about such topics are being repeated, for example, [1]. As single precision floating-point format becomes popular due to its higher computational performance, the movement to shorter data formats, FP16 and INT8, is stimulated by AI fields, especially deep-learning empowered by GPU. Although calculation with mixed-precision data formats demonstrates good performance...