BLAS-based Block Memory Accessors with Applications to Mixed-Precision Sparse Direct Solvers

Antoine JEGO

Sorbonne University, France

Abstract

Mixed-precision algorithms can take advantage of the lower precision of operands to enhance performance. In various contexts, it is beneficial to decouple the storage precision from the compute precision: the data is stored and accessed in low precision, but the computations are kept in high precision. This "memory accessor" approach benefits from reduced data accesses and improved accuracy, and can simplify the programming of mixed precision software packages. In this work, we develop such a memory accessor and investigate how it can accelerate sparse linear solvers. In particular, we assess its impact on custom floating-point datatypes unsupported by hardware and on structures such as Block Low-Rank formats. When considering BLAS-2 memory-bound operations like trsv we observe that the storage cost adequately matches the performance of the operation, in multiple parallel settings, provided that the conversion from storage to compute precision is efficient. For custom datatypes, we may take advantage of the recent AVX512-VBMI instruction set to reach the required efficiency. We present performance experiments using the sparse solver MUMPS.