Using Mixed Precision in Numerical Computations to Speedup Linear Algebra Solvers

Jack Dongarra, University of Tennessee/Oak Ridge National Lab

Low-precision floating-point arithmetic is a powerful tool for accelerating scientific computing applications, especially those in artificial intelligence. Here, we present an investigation showing that other high-performance computing (HPC) applications can also harness this power. Specifically, we use the general HPC problem, Ax=b, where A is a large dense matrix, and a double precision (FP64) solution is needed for accuracy. Our approach is based on mixed-precision (FP16 and FP64) iterative refinement, and we generalize and extend prior advances into a framework, for which we develop architecture-specific algorithms and highly tuned implementations. These new methods show how using half-precision Tensor Cores (FP16-TC) for the arithmetic can provide up to 4×speedup. This is due to the performance boost that the FP16-TC provide as well as to the improved accuracy over the classical FP16 arithmetic that is obtained because the GEMM accumulation occurs in FP32 arithmetic.