Enrique S. Quintana-Orti. **Opportunities for Approximate vs Transprecision Computing in Sparse Linear Solvers for GPUs**

Abstract

The convention in scientific computing is to employ IEEE double-precision (64-bit) arithmetic for all computations involving floating-point data. Nonetheless, appealing benefits from the adoption of mixed precision schemes have been reported for the solution of dense and sparse linear systems on graphics processing units (GPUs) via iterative refinement.

In this talk, we will illustrate the benefits of a generalization of the mixed precision strategy, known as Transprecision Computing (TC), in terms of execution time and energy efficiency. For this purpose, we will employ several case studies arising in the iterative solution of sparse linear systems on GPUs, with codes currently integrated the Ginkgo library (https://ginkgo-project.github.io). In some detail, this research effort exploits the fact that, for sparse linear algebra operations, the cost is dominated by the memory accesses while the arithmetic is largely irrelevant. To leverage this property, the Ginkgo solvers store certain parts of the data in reduced precision in memory, but operate in "full" 64-bit precision in order to bound the accumulation of rounding errors. Reduced-precision storage can be leveraged to maintain approximation operators, such as a preconditioner, or in a solver that gradually augments the precision of the operands as the iteration converges to the solution.