12.03.2014 17:50	Karl Rupp	0.154-115 Cauerstraße 7/9, Room 00.154

Achieving Portable High Performance for Iterative Solvers on Accelerators

Karl Rupp (TU Wien), Philippe Tillet (National Chiao Tung University, Taiwan), Ansgar Jüngel (TU Wien), Tibor Grasser (TU Wien)

Many supercomputers, clusters, and workstations today are equipped with accelerators such as graphics processing units (GPUs) and Intel's many-integrated core architecture (MIC). While their highly parallel architectures are very efficient for dense linear algebra operations, particularly those which are compute-bound rather than limited by memory bandwidth, their use for iterative solvers such as the conjugate gradient (CG) method or the generalized minimum residual (GMRES) method requires additional tricks in order to obtain best performance. Rather than minimizing the number of floating point operations, the key to best performance is a minimization of synchronization and memory transfers as well as maximizing the saturation of the memory channels. Because these optimizations are ultimately device-specific, parametric implementations are required in order to obtain portable performance.

In this talk we summarize our experiences from optimizing the performance of CG and GMRES methods. We start with simple, text-book-like implementations typically used for sequential processing and incrementally remove algorithmic bottlenecks for highly parallel architectures through kernel fusion and pipelining. Then, the choice of sparse matrix formats for computing matrix-vector products and their implication on performance depending on the underlying hardware is discussed. Finally, the remaining vector operations are investigated, where we demonstrate additional gains of our best, device-specific implementations over straight-forward, device-agnostic parallel implementations. We conclude with a summary of our optimizations for different GPUs from NVIDIA and AMD as well as for the Intel Xeon Phi.