Piyush K Sao

Scalable Sparse Direct Solver for Hybrid Architecture

RM1343
266 ferst Dr NW
Atlanta GA 30318
piyush3@gatech.edu
Xiaoye S. Li
Richard W. Vuduc

We present new algorithmic techniques for improving the scalability and efficiency of sparse direct solvers, specifically for distributed memory systems comprised of hybrid multicore CPU systems and hardware co-processors. This work extends the algorithm in SUPERLU_DIST, which is both right-looking and statically pivoted.

First, we present a novel algorithm for exploiting intra-node co-processors, which we call the HALO (Highly Asynchronous Lazy Offload) algorithm. HALO combines highly aggressive use of asynchrony with accelerated offload, lazy updates, and data shadowing (a la halo or ghost zones). These techniques hide and reduce communication, whether to local memory, across the network, or over PCIe. We evaluated our implementation on both NVIDIA GPUs and Intel Xeon-Phi accelerated systems. We use various realistic test problems for both single-node and multi-node configurations, finding that our implementation achieves speedups of up to 2.5x over an already efficient multicore CPU implementation.

Secondly, we present new algorithmic techniques to improve the scalability of sparse direct solvers at high core counts. Our techniques aggressively exploit elimination tree parallelism and replicate data to reduce per process communication volume. The net effect is to overlap communication with computation. We give preliminary experimental results that show the efficacy of these techniques.

Beyond sparse direct solvers, we believe our proposed techniques can apply to wide range of other sparse linear algebra computations.