Richard W. Vuduc

Model-driven autotuning of sparse matrix-vector multiply on GPUs

Georgia Institute of Technology
College of Computing
Computational Science and Engineering Division
266 Ferst Drive
Atlanta
GA 30332-0765
USA
richie@cc.gatech.edu
Jee Whan Choi
Amik Singh

We present a performance model-driven framework for automated performance tuning (autotuning) of sparse matrix-vector multiply (SpMV) on systems accelerated by graphics processing units (GPU). Our study consists of two parts.

First, we describe several carefully hand-tuned SpMV implementations for GPUs, identifying key GPU-specific performance limitations, enhancements, and tuning opportunities. These implementations, which include variants on classical blocked compressed sparse row (BCSR) and blocked ELLPACK (BELLPACK) storage formats, match or exceed state-of-the-art implementations. For instance, our best BELLPACK implementation achieves up to 29.0 Gflop/s in single-precision and 15.7 Gflop/s in double-precision on the NVIDIA T10P multiprocessor (C1060), enhancing prior state-of-the-art unblocked implementations (Bell and Garland, 2009) by up to 1.8× and 1.5× for single- and double-precision respectively.

However, achieving this level of performance requires input matrix-dependent parameter tuning. Thus, in the second part of this study, we develop a performance model that can guide tuning. Like prior autotuning models for CPUs (e.g., Im, Yelick, and Vuduc, 2004), this model requires offline measurements and run-time estimation, but more directly models the structure of multithreaded vector processors like GPUs. We show that our model can identify the implementations that achieve within 15% of those found through exhaustive search.

This paper appeared in the 15th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP), January 2010.