Efficient Triangular Matrix Vector Multiplication on the GPU

Takahiro Inoue, Hiroki Tokura, Koji Nakano, Yasuaki Ito
Department of Information Engineering, Hiroshima University
Higashi-Hiroshima, Japan
{takahiro, tokura, nakano, yasuaki}@cs.hirohsima-u.ac.jp

The main purpose of this paper is to present a very efficient GPU implementation to compute the product of a triangular matrix and a vector (trmv). Usually, developers use cuBLAS, a linear algebra library optimized for each of various generations of GPUs, to compute the trmv. To attain better performance than cuBLAS, our GPU implementation the trmv uses various acceleration technique for latest GPUs. More specifically, in our GPU implementation, (1) only one kernel is called (2) maximum number of threads are invoked, (3) all memory access to the global memory is coalesced, (4) all memory access to the shared memory has no bank conflict, and (5) shared memory access is minimized by a warp shuffle function. Experimental results for five generations of NVIDIA GPUs show that for matrices of sizes from 32×32 to $16\text{K} \times 16\text{K}$ show that our GPU implementation is faster than cuBLAS and muBLAS for almost all sizes and GPU generations.

Keywords: Matrix multiplication, trmv, Parallel algorithm, GPGPU.