Faster Compressed Sparse Row (CSR)-based Sparse Matrix-Vector Multiplication using CUDA

Yongchao Liu, Jorge González-Domínguez, Bertil Schmidt
Institute of Computer Science, University of Mainz, Germany
Emails: {yliu, j.gonzalez, bertil.schmidt}@uni-mainz.de

Abstract

LightSpMV [1] is a novel CUDA-compatible sparse matrix-vector multiplication (SpMV) algorithm using the standard compressed sparse row (CSR) storage format. It advances SpMV speedup by benefiting from the fine-grained dynamic distribution of matrix rows over vectors, where a warp is virtualized as a single instruction multiple data (SIMD) vector and can be further split into a set of equal-sized smaller vectors for fine-grained processing.

In LightSpMV, we have investigated two dynamic row distribution approaches at the vector and warp levels with atomic operations and warp shuffle functions as the fundamental building blocks. We have evaluated LightSpMV using various sparse matrices and further compared it to the CSR-based SpMV subprograms in the state-of-the-art CSR [3] and cuSPARSE [9] libraries. Performance evaluation reveals that on a single Tesla K40c GPU, LightSpMV is superior to both CSR and cuSPARSE, with a speedup of up to 2.60 and 2.63 over CSR and up to 1.93 and 1.79 over cuSPARSE for single and double precision, respectively. The source code of LightSpMV is available at http://lightspmv.sourceforge.net.

Compressed Sparse Row (CSR) Format

- A frequently used format for sparse matrix storage in CPU-centric software
- Efficient compression of structured and unstructured sparse matrices
- Good amenability to efficient algorithms designed for CPUs
- Enables good SpMV performance on CPUs, but shows a relatively low performance on GPUs
- Good amenability to efficient algorithms designed for CPUs
- Precision, respectively. The source code of LightSpMV is available at http://lightspmv.sourceforge.net

Sparse Matrix-Vector Multiplication

- 4. N. Bell and M. Garland: CUDA-based Sparse Matrix-Vector Multiplication on CUDA-enabled GPUs. 29th IEEE International Conference on Application-specific Systems, Architectures and Processors (v0.4), 2013

References