An Efficient Sparse Matrix Storage Format for Sparse Matrix-Vector Multiplication and Sparse Matrix-Transpose-Vector Multiplication on GPUs

Ryohei Izawa (Inoguchi Lab.)

The utilization of sparse matrix storage formats is widespread across various fields, including scientific computing, machine learning, and statistics. Within these domains, there is a need to perform Sparse Matrix-Vector Multiplication (SpMV) and Sparse Matrix-Transpose-Vector Multiplication (SpMVT) iteratively within a single application. However, executing SpMV and SpMVT on GPUs using existing sparse matrix storage formats presents challenges related to memory usage, load balancing, and memory access efficiency.

In our research, we propose a novel sparse matrix storage format named GCSB, specifically designed for efficient SpMV and SpMVT operations on GPUs, leveraging high memory compression. Initially, we adapt CSB, a sparse matrix storage format compatible with CPU-based SpMV and SpMVT, for GPU use in a straightforward manner, referred to as CSB-baseline. Subsequently, we extend the CSB-baseline to propose GCSB, which enables faster execution of SpMV and SpMVT than CSR through load balancing and efficient utilization of L1 cache, while maintaining theoretical memory usage equivalent to that of CSR.

Through experiments, we demonstrate that GCSB achieves SpMV and SpMVT with theoretical memory usage equivalent to CSR while outperforming CSR in terms of execution speed on several matrices from the University of Florida Sparse Matrix Collection. GCSB achieves up to 1.47× speedup on TITAN RTX and 2.75× on A100. Additionally, we show that GCSB reduces L1 cache miss counts compared to CSB-baseline. Furthermore, we qualitatively evaluate that GCSB demonstrates its superior performance under conditions where non-zero elements are broadly distributed throughout the matrix, the matrix size is considerable, and the proportion of non-zero elements within the matrix is relatively high.