Leveraging GPU Parallelism for Efficient CSR Sparse Matrix-Vector Multiplication

Antonio Gelain
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy
antonio.gelain@studenti.unitn.it

***Abstract*—Sparse Matrix-Vector multiplication (SpMV) is a fundamental operation in fields such as scientific computing, graph analysis and machine learning, which can often becomes a performance bottleneck due to the large amount of data which needs to be processed. This paper will cover various implementations and optimization strategies used to achieve faster computation leveraging both the Graphics Processing Units (GPUs) and the Compressed Sparse Row (CSR) matrix format.**

Although the CSR format is space efficient and widely used, its row-based structure presents difficulties for parallel execution on GPUs. Performance evaluations on a diverse set of sparse arrays drawn from real-world applications demonstrate significant speed gains over basic GPU implementations and competitive performance with more advanced libraries. The results highlight the feasibility of CSR-based SpMV for high-performance computing on GPUs, especially when combined with architecture-aware optimizations.

I. Introduction

II. Problem Statement

Sparse Matrix-Vector multiplication is a well-known problem whose purpose is to calculate the product between a sparse matrix, that is, whose values are mostly zeros and therefore do not contribute to the final result, and a vector.

III. Experimental setup

IV. Conclusion