Date - 07.09.2023

Topic - Accelerating CPU-Based Sparse General Matrix Multiplication With Binary Row Merging

Authors - ZHAOYANG DU, (Graduate Student Member, IEEE), YIJIN GUAN , TIANCHAN GUAN , DIMIN NIU , HONGZHONG ZHENG , YUAN XIE , (Fellow, IEEE)

Date of Publication - 26.07.2022

Pdf Link- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9840346

Summary-

Sparse matrix is a matrix which contains very few non-zero elements. Sparse general matrix multiplication (SpGEMM) is a fundamental operation in various real-world applications, including algebraic multigrid solvers, graph algorithms, and simulations. This research paper focuses on memory access efficiency when optimising SpGEMM for multi-core CPU architectures.

This paper highlights two critical memory access patterns in SpGEMM. First, there are vast and irregular memory accesses to the B matrix due to the row-wise dataflow. These memory accesses are challenging for modern CPU memory systems. Second, during the accumulation phase, memory accesses involve merging numerous intermediate lists of variable lengths.

Existing SpGEMM libraries prioritize computational complexity. But they often neglect memory access efficiency. This leads to suboptimal performance on CPU architectures. This paper introduces a new addition method called BRMerge. It follows the row-wise dataflow. BRMerge accesses the B matrix efficiently and generates intermediate lists. It stores them efficiently using a ping-pong buffer. It then joins these lists into a tree-like hierarchy.

The architectural benefits of BRMerge are streaming access patterns, minimised TLB cache misses, and reasonably high cache hit rates. These result in low access latency and high bandwidth utilization. This study presents two SpGEMM libraries based on BRMerge that use distinct allocation algorithms. Performance tests on two Intel Xeon CPUs show that these libraries outperform SpGEMM libraries.

Opinion-

This research paper addresses an important issue in high-performance computing- optimizing sparse general matrix multiplication on multi-core CPU architectures. It introduces BRMerge, a new addition method that efficiently handles memory access patterns during SpGEMM. It considerably improves speed over existing libraries.

This paper's emphasis on memory access efficiency, streaming access patterns, and cache utilization is critical for achieving high SpGEMM performance on modern CPU architectures. The research outcomes are promising for substantial performance speedups over existing libraries. I think this will have a significant impact on applications relying on SpGEMM, making them more efficient and practical on CPU-based systems.