

## Performance Analysis and Optimization for Diagonal Sparse Matrix-Vector Multiplication on Machine Learning Unit

**Authors :** Qiuyu Dai, Haochong Zhang, Xiangrong Liu

**Abstract :** Diagonal sparse matrix-vector multiplication is a well-studied topic in the fields of scientific computing and big data processing. However, when diagonal sparse matrices are stored in DIA format, there can be a significant number of padded zero elements and scattered points, which can lead to a degradation in the performance of the current DIA kernel. This can also lead to excessive consumption of computational and memory resources. In order to address these issues, the authors propose the DIA-Adaptive scheme and its kernel, which leverages the parallel instruction sets on MLU. The researchers analyze the effect of allocating a varying number of threads, clusters, and hardware architectures on the performance of SpMV using different formats. The experimental results indicate that the proposed DIA-Adaptive scheme performs well and offers excellent parallelism.

**Keywords :** adaptive method, DIA, diagonal sparse matrices, MLU, sparse matrix-vector multiplication

**Conference Title :** ICS 2023 : International Conference on Supercomputing

**Conference Location :** Paris, France

**Conference Dates :** June 22-23, 2023