

D3.13 - Analysis of algorithmic characteristics and hardware architecture templates - Final

In the evolving landscape of data-driven applications, the demand for efficient, high-performance computing has never been more critical. Machine Learning (ML) algorithms, the driving force behind advancements in pattern recognition, image classification, and natural language processing, have outgrown the capabilities of traditional processors like central processing units (CPUs) and general-purpose graphics processing units (GPUs). This work highlights the inherent limitations of conventional processors. While designed for flexibility across various applications, CPUs and GPUs incur significant energy and latency overhead during instruction fetching and decoding, making them less ideal for specific applications such as Algebraic Machine Learning(AML). To address this challenge, this work explores the design of a dedicated hardware architecture specialized for AML. By eliminating unnecessary overheads and dedicating all resources to AML computations, the goal is to utilize massive parallelization to boost AML computation performance. However, an obstacle arises when dealing with the substantial huge data size in AML, and facing memory bandwidth limitations. This memory bottleneck leads to suboptimal resource utilization and longer execution times, a problem known as the "memory wall." To address this issue and enhance throughput, this work proposed several hardware optimization techniques. The optimizations are performed in different levels of abstraction, such as data format level, architecture level, and system level. Also, this work emphasizes the important role of data compression techniques. It notes that AML tasks often involve large, sparsely populated bit arrays, rendering the conventional data formats inefficient. In this project, we explore various conventional data compression strategies and finally propose a new algorithm, to optimize data size while preserving essential information.

Finally, we have deployed an FPGA platform and implemented our proposed hardware architecture on real hardware to assess the performance. As shown in deliverable D3.13, In comparison with the CPU implementation of AML, our hardware architecture shows 1.6X better throughput while having 3X lower power consumption. The proposed hardware architecture is scalable, and based on the chosen platform, the throughput can be further increased.