Design, Integration, and Performance Analysis of an INT8 Systolic Array Accelerator on an FPGA-Based System-on-Chip

Thiago Fernandes
Department of Computer Engineering
University of Vale do Itajaí (UNIVALI)
Itajaí, Brazil

***Abstract*—The growing computational demands of modern applications, particularly in machine learning and digital signal processing, have increased the need for hardware accelerators capable of delivering high performance with reduced energy consumption. Field-Programmable Gate Arrays (FPGAs) enable the implementation of highly parallel and customizable architectures suitable for such workloads.**

This paper presents the design, integration, and performance evaluation of an INT8 hardware accelerator based on a systolic array architecture, fully implemented in VHDL and integrated into an FPGA-based System-on-Chip (SoC). The accelerator communicates with a soft-core processor through an Avalon Memory-Mapped interface using a dedicated wrapper. RTL simulation results demonstrate low latency and substantial throughput improvements compared to a general-purpose processor executing the same workload in software.

***Index Terms*—FPGA, Hardware Accelerator, Systolic Array, INT8 Arithmetic, System-on-Chip.**

I. Introduction

The rapid growth of computational workloads in areas such as computer vision, signal processing, and machine learning has exposed the limitations of general-purpose processors in terms of performance and energy efficiency [1]. Many of these workloads are dominated by repetitive and highly parallel operations, such as multiply-and-accumulate (MAC), which are inefficiently handled by sequential processor architectures.

Hardware accelerators have emerged as a viable solution by offloading computationally intensive tasks to dedicated processing units. FPGAs play a key role in this scenario due to their ability to implement customized parallel architectures while maintaining flexibility and reduced development cost compared to ASICs [2].

Among several architectural approaches, systolic arrays stand out for their regular structure, efficient data reuse, and scalability. These characteristics make them particularly suitable for MAC-intensive workloads commonly found in machine learning applications [4].

This work presents the design and integration of an INT8 systolic array accelerator on an FPGA-based System-on-Chip and provides an analytical performance comparison against a general-purpose processor model.

II. Background and Related Work

A. Hardware Acceleration on FPGAs

Hardware accelerators are specialized processing units designed to execute specific functions more efficiently than general-purpose processors. As discussed by Hennessy and Patterson [1], this approach reduces instruction overhead and enables higher throughput by exploiting parallelism.

Although FPGA-based accelerators typically exhibit higher area overhead compared to ASICs, their reconfigurability and rapid prototyping capabilities make them suitable for research and embedded systems development [3].

B. Systolic Array Architecture

The systolic array architecture, originally proposed by Kung and Leiserson [4], consists of a network of processing elements operating synchronously while data flows rhythmically between them. This model minimizes memory accesses and improves data locality.

Systolic arrays are widely used in matrix multiplication, convolution, and neural network inference due to their high computational efficiency and predictable timing behavior.

C. INT8 Fixed-Point Arithmetic

The adoption of INT8 fixed-point arithmetic has become common in modern accelerators due to reduced area, power consumption, and latency. Prior work demonstrates that, with proper quantization techniques, INT8 arithmetic can achieve competitive accuracy for inference workloads [5].

III. Accelerator Architecture

The proposed accelerator was fully implemented in VHDL using a modular design approach. Each processing element performs an INT8 multiplication followed by accumulation in a 32-bit register, ensuring safe accumulation without overflow.

Multiple processing elements are interconnected to form a one-dimensional systolic array. Input data propagates through the array while weights remain stationary, enabling efficient data reuse and parallel computation. The design is fully synchronous and supports pipelined execution.

IV. System-on-Chip Integration

To integrate the accelerator into a complete system, a wrapper compatible with the Avalon Memory-Mapped interface was developed according to Intel specifications [6]. This wrapper exposes control and data registers to the processor, enabling seamless software control.

The System-on-Chip was assembled using Intel Platform Designer [7] and includes a soft-core processor, on-chip memory, clock and reset controllers, and the custom INT8 accelerator.

V. Performance Evaluation and Comparison

Performance evaluation was conducted using an analytical cycle-count model derived from RTL simulation results. The accelerator operates at 100 MHz and produces a valid result three clock cycles after activation. After pipeline filling, new results are generated every two clock cycles.

For comparison, a simple in-order processor executing the same MAC operation in software was considered, following assumptions commonly adopted in architectural studies [1].

A. Latency Comparison

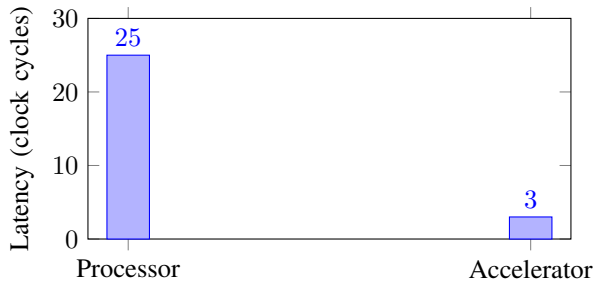

Fig. 1. Latency comparison between processor and accelerator

B. Throughput Comparison

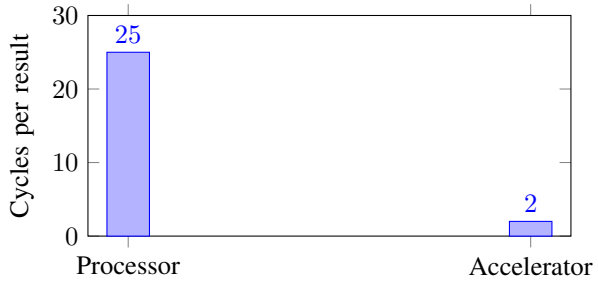

Fig. 2. Throughput comparison expressed as cycles per result

Lower values indicate higher throughput.

VI. Speedup Analysis

Speedup is defined as the ratio between the execution time of a baseline processor and that of the accelerator [1]. Two complementary metrics are considered.

A. Latency Speedup

$$\text{Speedup}_{\text{latency}} = \frac{25}{3} \approx 8.33\times \tag{1}$$

B. Throughput Speedup

$$\text{Speedup}_{\text{throughput}} = \frac{25}{2} = 12.5\times \tag{2}$$

TABLE I
Speedup Comparison Between Processor and Accelerator

Metric	Speedup
Latency Speedup	8.33×
Throughput Speedup	12.5×

VII. Conclusion

This paper presented the design, integration, and performance analysis of an INT8 systolic array accelerator implemented on an FPGA-based System-on-Chip. RTL simulation results demonstrate that the proposed accelerator achieves substantially lower latency and higher throughput compared to a general-purpose processor executing the same workload in software.

The results confirm that systolic array architectures combined with INT8 arithmetic are well suited for MAC-intensive applications, providing significant performance gains through parallelism and pipelining. Future work includes scaling the architecture to larger arrays, supporting additional operations, and validating performance on physical FPGA platforms.

References

[1] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 6th ed., Morgan Kaufmann, 2019.

[2] I. Kuon, R. Tessier, and J. Rose, "FPGA architecture: Survey and challenges," *Foundations and Trends in Electronic Design Automation*, vol. 2, no. 2, pp. 135–253, 2008.

[3] I. Kuon and J. Rose, "Measuring the gap between FPGAs and ASICs," *IEEE Trans. Computer-Aided Design*, vol. 26, no. 2, pp. 203–215, 2007.

[4] H. T. Kung and C. E. Leiserson, "Systolic arrays for VLSI," in *Sparse Matrix Proceedings*, Academic Press, 1979.

[5] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. CVPR*, 2018.

[6] Intel Corporation, *Avalon Interface Specifications*, 2024.

[7] Intel Corporation, *Platform Designer User Guide*, 2024.