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

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

***Abstract*—The increasing computational demand of modern applications, especially in machine learning and digital signal processing, has driven the adoption of hardware accelerators to improve performance and energy efficiency. In this context, Field-Programmable Gate Arrays (FPGAs) enable the implementation of highly parallel and customizable architectures.**

This paper presents the design and verification 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 is connected to a soft-core processor through an Avalon Memory-Mapped interface using a dedicated wrapper. Functional verification and performance analysis were conducted via RTL simulation, demonstrating low latency and high throughput.

***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 consist of repetitive and highly parallel operations, which are not efficiently exploited by conventional CPU architectures.

Hardware accelerators have emerged as a viable solution by offloading specific computational tasks to dedicated processing units. FPGAs play a key role in this scenario due to their ability to implement customized parallel architectures with reduced development time when 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 multiply-and-accumulate (MAC) intensive workloads, commonly found in machine learning applications [4].

This work presents the design of an INT8 systolic array accelerator and its integration into an FPGA-based System-on-Chip using a standard on-chip bus interface.

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. According to Hennessy and Patterson [1], this approach significantly reduces instruction overhead and improves throughput.

Despite the area overhead of FPGA implementations when compared to ASICs, their flexibility and reconfigurability make them attractive for research and prototyping purposes [3].

B. Systolic Array Architecture

The systolic array architecture, introduced by Kung and Leiserson [4], consists of a network of processing elements that operate synchronously while data flows rhythmically between them. This approach minimizes memory access and maximizes data locality.

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

C. INT8 Fixed-Point Arithmetic

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

III. Accelerator Design

The proposed accelerator was fully implemented in VHDL using a modular design approach. Each processing element performs INT8 multiplication followed by accumulation in a 32-bit register, preventing overflow during successive operations.

Multiple processing elements are interconnected to form a one-dimensional systolic array. Input data propagates through the array while weights remain stationary, enabling efficient reuse and high parallelism.

The design is fully synchronous and supports pipelined operation, allowing higher operating frequencies.

IV. System-on-Chip Integration

To integrate the accelerator into a complete system, a wrapper compatible with the Avalon Memory-Mapped interface was developed, following Intel's specifications [6]. This wrapper encapsulates the internal accelerator logic and exposes control and data registers to the processor.

The complete System-on-Chip was assembled using Intel Platform Designer [7], including:

- Soft-core processor

- On-chip memory
- Clock and reset controllers
- Custom INT8 accelerator

Figure 1 illustrates the overall system architecture.

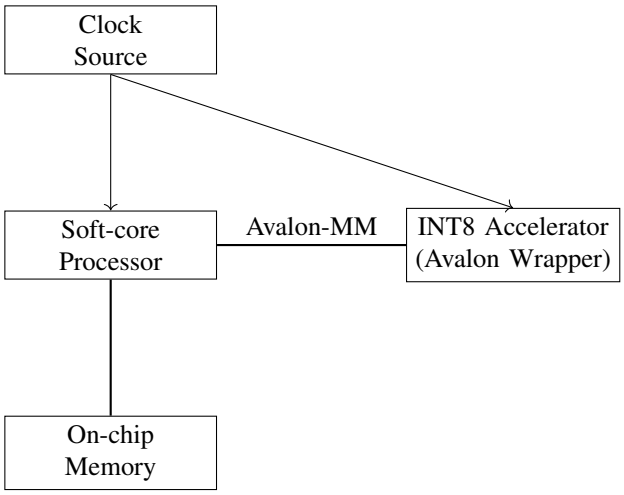

Fig. 1. FPGA-based System-on-Chip architecture with integrated INT8 accelerator

V. Verification and Performance Analysis

Functional verification was performed at RTL level using ModelSim. A dedicated VHDL testbench was developed to emulate software behavior, including register writes, accelerator activation, and result acquisition.

The system operated with a 10 ns clock period (100 MHz). Simulation results showed that the accelerator produces a valid output three clock cycles after the start signal assertion, corresponding to a latency of approximately 30 ns.

After pipeline filling, new results were generated every two clock cycles, demonstrating high throughput and efficient utilization of the systolic array architecture.

VI. Conclusion

This paper presented the design, integration, and verification of an INT8 systolic array accelerator implemented on an FPGA-based System-on-Chip. The results demonstrate correct functional behavior, low latency, and high throughput, validating the effectiveness of systolic array architectures for hardware acceleration.

Future work includes quantitative comparison against software execution on embedded processors and expansion of the accelerator to support larger arrays and additional operations.

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 Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 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.