Design and Implementation of an INT8 Systolic Array Accelerator on FPGA Integrated into a System-on-Chip

Thiago Fernandes
University of Vale do Itajaí (UNIVALI)
Itajaí, Brazil
thifj@edu.univali.br

***Abstract*—The increasing computational demand of modern applications, especially in digital signal processing and machine learning, has motivated the development of hardware accelerators capable of delivering high performance and energy efficiency. Field-Programmable Gate Arrays (FPGAs) provide an attractive platform for such accelerators due to their flexibility and inherent parallelism.**

This paper presents the design and implementation of an INT8 arithmetic accelerator based on a systolic array architecture, fully described in VHDL and integrated into a System-on-Chip (SoC). The accelerator is connected to a softcore processor through an Avalon Memory-Mapped interface using a dedicated wrapper module. The proposed architecture emphasizes modularity, scalability, and efficient data reuse.

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

I. Introduction

The rapid growth of computational workloads in applications such as machine learning, computer vision, and digital signal processing has intensified the need for specialized computing architectures. Many of these applications rely heavily on repetitive arithmetic operations that can benefit significantly from parallel execution.

Hardware accelerators provide a promising solution by offloading computationally intensive tasks from general-purpose processors to dedicated hardware units. Among available platforms, FPGAs stand out due to their capability to implement custom parallel architectures while maintaining flexibility.

Systolic array architectures have been widely adopted in modern accelerators due to their regular structure, scalability, and efficient data reuse. In particular, INT8 arithmetic has gained prominence as it offers reduced area and power consumption while maintaining acceptable accuracy for many applications.

This work presents the design of an INT8 systolic array accelerator and its integration into an FPGA-based SoC, focusing on architectural design, hardware modularization, and hardware/software interaction.

II. Background

A. FPGA-Based Hardware Accelerators

Hardware accelerators are specialized circuits designed to perform specific tasks more efficiently than general-purpose processors. On FPGAs, accelerators can exploit spatial and temporal parallelism, enabling deep pipelining and high throughput designs.

B. Systolic Array Architecture

A systolic array consists of an array of processing elements that rhythmically compute and pass data through the system. Each processing element performs a simple operation, while the overall architecture achieves high performance through parallelism and local communication.

This structure is particularly effective for matrix operations and multiply-accumulate workloads commonly found in machine learning applications.

C. INT8 Fixed-Point Arithmetic

INT8 fixed-point arithmetic is widely used in hardware accelerators due to its reduced resource usage and lower power consumption. Accumulation is typically performed in higher precision, such as 32-bit registers, to avoid overflow and preserve numerical accuracy.

III. Methodology

The development process followed an incremental approach. First, a basic processing element (PE) capable of performing INT8 multiplication and accumulation was designed in VHDL. The PE includes pipelined registers to support high operating frequencies.

Next, multiple PEs were interconnected to form a one-dimensional systolic array, allowing data to propagate sequentially while weights remain stationary. After validating the accelerator at the RTL level, a wrapper compatible with the Avalon Memory-Mapped interface was developed.

Finally, the accelerator was integrated into a System-on-Chip using Intel Platform Designer, alongside a softcore processor, on-chip memory, clock source, and reset controller.

IV. Accelerator Design

The proposed accelerator is fully implemented in VHDL and follows a modular design strategy. Each processing element receives an input value and a weight, performs an INT8 multiplication, and accumulates the result into a 32-bit register.

The systolic array is formed by chaining multiple processing elements, enabling efficient data flow and minimizing memory

access overhead. The final accumulated result is exposed through the wrapper for processor access.

This modular structure allows scalability by adjusting the number of processing elements without altering the overall architecture.

V. System-on-Chip Architecture

The SoC architecture follows a bus-based model in which a softcore processor orchestrates system execution and peripheral control. The INT8 accelerator is integrated as a custom peripheral connected through the Avalon Memory-Mapped interface.

Fig. 1 illustrates the overall system architecture.

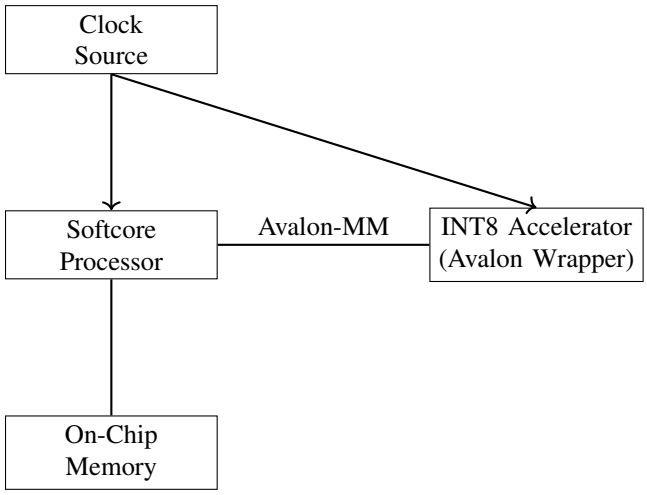

Fig. 1. System-on-Chip architecture with integrated INT8 accelerator

A. Communication Model

Communication between the processor and the accelerator occurs through memory-mapped registers. The processor writes input data and control signals to the accelerator and reads the computed results once processing is complete.

This approach ensures loose coupling between software and hardware, improving system modularity and reusability.

B. Data Flow

The software initiates computation by configuring the accelerator registers and asserting a start signal. Internally, the wrapper translates these commands into control signals for the systolic array. Data flows sequentially through the processing elements, while weights remain stationary, enabling efficient reuse.

VI. Conclusion

This work demonstrated the successful design and integration of an INT8 systolic array accelerator on FPGA within a System-on-Chip architecture. The proposed solution highlights the advantages of systolic arrays for arithmetic-intensive workloads and demonstrates efficient hardware/software integration using the Avalon interface.

Future work may explore multi-dimensional systolic arrays, support for additional data types, and performance evaluation using real-world workloads.

References

[1] Intel Corporation, *Avalon Interface Specifications*, Intel FPGA Documentation.

[2] Intel Corporation, *Platform Designer User Guide*, Intel FPGA Documentation.

[3] H. T. Kung, "Why Systolic Architectures?" *IEEE Computer*, vol. 15, no. 1, pp. 37–46, 1982.