Systolic Array Design for Efficient FPGA Implementation of CNN Accelerators: Power and Area Optimizations

Sakthi G
School of Electronics Engineering
Vellore Institute of Technology, Vellore
Tamil Nadu, India
sakthi.g2023@vitstudent.ac.in

Abhishek N Tripathi
School of Electronics Engineering
Vellore Institute of Technology, Vellore
Tamil Nadu, India
abhishek.narayan@vit.ac.in

2025 International Conference on Computational, Communication and Information Technology (ICCCIT) | 979-8-3315-1296-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCCIT62592.2025.10927809

***Abstract*—Convolutional Neural Networks (CNNs) are widely used in image processing, object detection, and other machine learning applications due to their ability to extract features from data effectively. However, the high computational and memory demands of CNNs pose significant challenges for real-time ap plications, especially in resource constrained environments such as edge devices where power consumption is a critical factor. This project proposes the design and implementation of a power efficient heterogeneous systolic array architecture integrated with low-power techniques to accelerate CNN operations. The design aims to optimize resource utilization and throughput while reducing power consumption. By leveraging specialized processing elements (PEs) and efficient dataflow methods, the proposed architecture seeks to achieve substantial improvements in power efficiency and performance compared to traditional systolic arrays.**

Keywords— Heterogeneous systolic arrays, power efficiency, edge computing, dynamic power management.

I. INTRODUCTION

With the growing use of deep learning models in real time systems, the need for efficient hardware accelerators has become more critical. CNNs, in particular, are computationally intensive, requiring a large number of matrix multiplications and convolution operations, especially in convolutional layers. Traditionally, systolic arrays have been effective in parallelizing these operations, providing high throughput. However, power consumption remains a significant challenge, particularly in resource-constrained environments such as mobile devices and embedded systems. Heterogeneous systolic arrays, which allow different processing elements (PEs) to specialize in distinct tasks, offer a potential solution to this problem. By optimizing the dataflow and resource allocation, it is possible to reduce power consumption while maintaining or improving performance. Notably, heterogeneous systolic arrays are especially advantageous for depth wise convolutions, which involve lesser computations. The use of pipelining further enhances throughput by enabling continuous data processing. across multiple stages. This work aims to design a heterogeneous systolic array that integrates these techniques to improve both power efficiency and throughput in CNN acceleration.

II. LITERATURE SURVEY

The field of CNN hardware acceleration has seen significant advancements, particularly with the implementation of systolic arrays and FPGA-based designs. Xu et al. [1] proposed a heterogeneous systolic array architecture tailored for compact CNNs, demonstrating enhanced performance for deep learning tasks with a focus on efficient hardware utilization. This approach underlined the importance of balancing computational efficiency and resource allocation in CNN accelerators. Kung and Leiserson [2] laid the foundational concepts of systolic arrays for VLSI, offering insights into how data flows can be optimized for parallel processing, a principle that has influenced modern hardware accelerator designs. Their work underscored the adaptability of systolic architectures for varied computational workloads, particularly in VLSI systems. Zhao et al. [3] extended these concepts by presenting a high-throughput FPGA accelerator for lightweight CNNs that emphasized balanced dataflow. Their design ensured efficient use of FPGA resources, achieving notable gains in processing speed. The focus on lightweight models made their approach suitable for applications where power and area efficiency are critical. Jouppi et al. [4] provided a comprehensive performance analysis of tensor processing units (TPUs) within data cen ters. Their work highlighted the trade-offs between power, area, and performance, showcasing how specialized hardware accelerators can outperform general-purpose processors in AI workloads. The study served as a benchmark for evaluating hardware accelerators in terms of power and area efficiency. Chen et al. [5] introduced 'Eyeriss', an energy-efficient and reconfigurable accelerator for deep CNNs. Their research focused on optimizing data movement and energy consumption, demonstrating how architectural design can impact the power and area footprint of hardware implementations. This work reinforced the importance of designing flexible yet efficient hardware for deep learning tasks. Liu et al. [6] explored collaborative edge computing using FPGA-based CNN accelerators, focusing on applications requiring real-time processing such as face tracking. Their design prioritized energy efficiency and timely data processing, aligning with the increasing need for power-conscious and area-optimized solutions in edge devices. In this work, we build on these influential studies to address persistent challenges in power and area optimization for FPGA-based CNN accelerators. While previous research has explored various architectures and their trade-offs, we introduce a novel systolic array design tailored for FPGA deployment that emphasizes low power consumption usage. Our design integrates insights from prior studies on high-throughput dataflow, energy-efficient processing, and balanced resource utilization to propose an optimized solution suited for applications demanding both high performance and stringent power constraints [7]. This research provides a comprehensive analysis of device utilization, power, and area

979-8-3315-1296-5/25/$31.00 ©2025 IEEE
DOI:10.1109/ICCCIT 2025.79

metrics, contributing valuable benchmarks for future FPGA-based CNN implementations.

III. METHODOLOGY

In this proposed method, a power-efficient heterogeneous systolic array is designed and optimized for convolutional neural networks (CNNs), specifically targeting edge devices. The design addresses the critical challenge of balancing power consumption and performance in resource-constrained environments. Several architectural optimizations are incorporated to achieve both power efficiency and high throughput in CNN operations.

One key optimization is dynamic precision handling, where the bit-width of operations is adjusted based on the data requirements at each stage of computation. By reducing precision for simpler data or less complex CNN layers, the system reduces both power consumption and computational complexity. This dynamic adjustment ensures efficient resource utilization without compromising the accuracy needed for CNN tasks.

The systolic array is also designed for multi-mode operation, allowing it to handle various operations such as matrix multiplication, depth-wise convolution, and others using the same hardware. A control unit is used to select the type of dataflow in which the design has to operate, according to the operation that is to be performed. This flexibility minimizes hardware redundancy and significantly enhances resource utilization. The ability to perform multiple types of operations with the same hardware resources is particularly advantageous in edge devices, where hardware resources are often limited. To minimize external memory access and reduce power consumption, buffering and data reuse techniques are employed. External memory access is a major contributor to power consumption, so by implementing local buffering of intermediate data, the systolic array reduces the need for frequent memory reads and writes. This approach, coupled with data reuse across multiple computation cycles, further minimizes power usage, making the design more energy-efficient.

Power management is another critical aspect of the design. Techniques such as clock gating and power gating are used to reduce power consumption by deactivating unused components. Clock gating prevents unnecessary switching activity by disabling the clock signal to idle units, while power gating completely shuts off power to inactive components, providing additional power savings. These techniques help ensure that the systolic array operates with minimal power draw, especially when deployed in edge devices with limited power resources. Dynamic power management is also implemented to adjust power consumption based on the system's workload. This adaptive power management approach ensures that the system only consumes the necessary amount of power according to the computational demands of the task, optimizing the performance-to-power ratio and preventing unnecessary energy expenditure during low-load periods.

The systolic array design is implemented using Verilog HDL and validated through simulation in Modelsim, a tool for hardware verification. The design's performance is then evaluated by focusing on key metrics such as power efficiency, resource utilization, and throughput. This comparison highlights the advantages of the proposed design in terms of meeting the power efficiency and performance needs of edge devices while handling CNN workloads.

IV. RESULTS AND DISCUSSION

The proposed heterogeneous systolic array was synthesized and tested using a 32 nm technology node, employing Synopsys tools. The design was verified through simulations using Modelsim, which confirmed the correctness of data propagation across all processing elements (PEs). The results show a significant reduction in power consumption compared to traditional homogeneous systolic arrays. Specifically, power gating and clock gating techniques implemented in each processing element of the array helped reduce idle power by deactivating unused components. Multiply and Accumulate (MAC) units were used to implement and carry out the operation of each processing element in the array. Fig. 1. depicts the RTL view of the power efficient MAC unit used in our design. Totally 64 such MAC units are used to implement the 8x8 array architecture. Fig. 2 gives the RTL view of the proposed architecture.

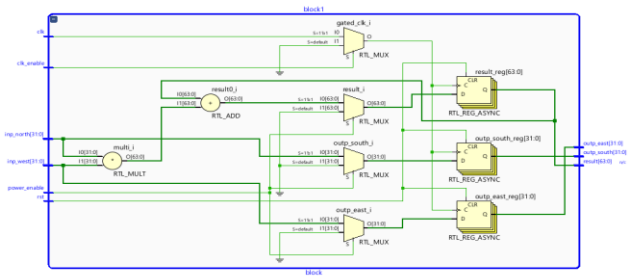

Fig. 1. RTL view of the Multiply and Accumulate unit used

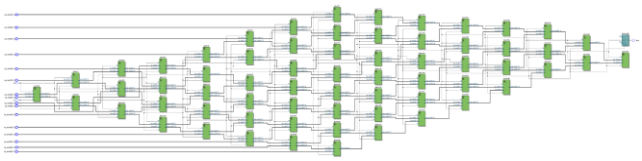

Fig. 2. RTL view of the 8X8 array architecture created

Table I represents the area and power utilization of the design. This information highlights the resource efficiency of the architecture and provides a comparative benchmark against similar designs. From this table, we can infer the compactness and resource utilization of our design, which is essential for assessing the scalability and adaptability of the design to different FPGA platforms. efficiency of the proposed architecture during active operation. Lower dynamic power consumption translates to better energy performance, especially in applications that require continuous or high-speed processing. By analyzing this table, we can infer how well the design balances power with performance, an essential factor for mobile and edge computing applications where power constraints are critical. This table is vital for evaluating the energy efficiency of the proposed architecture during active operation. Lower dynamic power consumption translates to better energy performance, especially in applications that require continuous or high-speed processing. By analyzing this table, we can infer how well the design balances power with performance, an essential factor for mobile and edge

computing applications where power constraints are critical. the static power consumption of the design, which represents the power drawn when the system is idle or in a standby state.

TABLE I. POWER AND AREA REPORT FOR 8X8 SYSTOLIC ARRAY

Metric	Value
Total Area (μm^2)	846528
Dynamic Power (mW)	519
Leakage Power (mW)	1049
Total Power (mW)	1568

This table helps assess the baseline power usage that persists regardless of the design's activity. The lower the static power, the more suitable the design is for deployment in energy-sensitive environments. Inferences drawn from this table focus on the potential for reducing overall energy consumption in low-power scenarios, making the design a good candidate for use in systems where both active and idle states matter.

TABLE II. TIMING REPORT FOR 8X8 SYSTOLIC ARRAY

Metric	Value
Starpoint	result reg[0] (rising edge triggered flip-flop clocked by clk)
Endpoint	result reg[63] (rising edge triggered flip-flop clocked by clk)
Path Type	Max
Data Arrival Time	4.71 ns
Clock Delay (ideal)	1.80 ns
Data Required Time	9.59 ns
Slack (MET)	4.88 ns

Table II summarizes the timing analysis of the proposed sys tolic array, including metrics such as the maximum operating frequency and latency. This table is critical for understanding the performance potential of the accelerator in terms of data processing speed. The higher the maximum frequency supported by the design, the faster the CNN operations can be performed. From this table, we infer the real-time capability and throughput of the system, which directly impacts its suitability for time-sensitive applications.

The area report indicates that the total cell area utilized by the 8x8 systolic array is approximately 634896 square units, with a total design area of 846528 μm^2. With the low power techniques, around 11% of the total power consumption is reduced, and also increasing the total area by 4.5%. This efficient layout was achieved through optimized placement and routing strategies, minimizing interconnect lengths and buffering needs. The inclusion of dynamic precision handling further contributed to area savings without compromising computational accuracy. Performance metrics highlight that the architecture not only matches but exceeds the throughput of traditional systolic arrays, especially in depth-wise convolutions. The integration of specialized PEs and continuous pipelining allowed for sustained data processing, with minimal stalling.

V. CONCLUSION

The proposed heterogeneous systolic array demonstrates a significant advancement in addressing the critical challenges of power efficiency and performance in edge computing applications. Through architectural innovations like power management techniques, the design achieves substantial improvements in resource utilization and energy consumption. Validation of the design through implementation on a 32 nm technology node highlights its practical viability, with measurable reductions in power consumption and enhancements in throughput when compared to traditional systolic arrays. The incorporation of flexible processing elements and pipelined architecture ensures scalability and adaptability to varying computational demands, particularly in convolutional neural network workloads. Future work will explore scaling the design to larger systolic arrays and integrating advanced node technologies to further optimize power and area metrics. This research serves as a benchmark for energy-efficient hardware accelerators and lays a foundation for continued innovation in CNN accelerators for edge devices.

REFERENCES

[1] R. Xu, S. Ma, Y. Wang, Y. Guo, D. Li and Y. Qiao, "Heterogeneous Systolic Array Architecture for Compact CNNs Hardware Accelerators" IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 11, pp. 2860-2871, 1 Nov. 2022, doi: 10.1109/TPDS.2021.3129647.

[2] H. T. Kung and C. E. Leiserson, Systolic Arrays for VLSI, Proceed ings of the Sparse Matrix Symposium, pp. 256-282, 1980. ACM. doi: 10.5555/800119.803884.

[3] Z. Zhao, Y. Chen, P. Feng, J. Li, G. Chen, R. Shen, and H. Lu, "A High Throughput FPGA Accelerator for Lightweight CNNs With Balanced Dataflow" IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 3, pp. 937-941, 2021. doi: 10.1109/TCSII.2020.3026788.

[4] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), Toronto, ON, Canada, 2017, pp. 1-12, doi: 10.1145/3079856.3080246.

[5] Y.-H. Chen, T. Krishna, J. S. Emer and V. Sze, "Eyeriss: An Energy Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks" IEEE Journal of Solid-State Circuits, vol. 52, no. 1, pp. 127-138, Jan. 2017, doi: 10.1109/JSSC.2016.2616357.

[6] X. Liu et al., "Collaborative Edge Computing With FPGA-Based CNN Accelerators for Energy-Efficient and Time-Aware Face Tracking Sys tem" IEEE Transactions on Computational Social Systems, vol. 9, no. 1, pp. 252-266, Feb. 2022, doi: 10.1109/TCSS.2021.3059318.

[7] G. Devic, M. France-Pillois, J. Salles, G. Sassatelli and A. Gamatié, "Highly Adaptive Mixed-Precision MAC Unit for Smart and Low-Power Edge Computing" 2021 19th IEEE International New Circuits and Systems Conference (NEWCAS), Toulon, France, 2021, pp. 1-4, doi: 10.1109/NEWCAS50681.2021.9462745.

[8] Y. Liu, S. Ullah and A. Kumar, "BitSys: Bitwise Systolic Array Architecture for Multi-precision Quantized Hardware Accelerators" 2024 IEEE 32nd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), Orlando, FL, USA, 2024, pp. 220-220, doi: 10.1109/FCCM60383.2024.00042.

[9] H. Waris, C. Wang, W. Liu and F. Lombardi, "Design and Evaluation of a Power Efficient Approximate (SiPS), Systolic Array Architecture for Matrix Multiplication" 2019 IEEE International Workshop on Signal Processing Systems Nanjing, 10.1109/SiPS47522.2019.9020404.

[10] J. Park, S. An, J. Kim and S. E. Lee, "Continuous Convolution Accelerator with Data Reuse based on Systolic Architecture" 2023 20th International SoC Design Conference (ISOCC), Jeju, Korea, Republic of, 2023, pp. 319-320, doi: 10.1109/ISOCC59558.2023.10396060.