ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

CAAM: Compressor-Based Adaptive Approximate Multiplier for Neural Network Applications

[1]B. Satish, [2] DR.Ch. Ravi Kumar.

[1]P.G Student, Dept ELECTRONICS AND COMMUNICATION ENGINEERING, Prakasam Engineering College(Autonomous), kandukur, India.

[2]Professor, Dept ELECTRONICS AND COMMUNICATION ENGINEERING, Prakasam Engineering College(Autonomous), kandukur, India.

ABSTRACT

The rapid growth of artificial intelligence and neural network applications has increased the demand for energy-efficient hardware accelerators. Multipliers are among the most resource-intensive components in neural network processors, consuming significant power and area. This project proposes a Compressor-Based Adaptive Approximate Multiplier (CAAM) that improves computational efficiency by utilizing adaptive error-tolerant compressors. The architecture reduces switching activity and critical path delay while maintaining acceptable computational accuracy.

Approximate computing techniques enable a flexible trade-off between power consumption and output precision. The proposed design significantly lowers hardware complexity and area utilization compared to exact multipliers. Experimental evaluation demonstrates considerable energy savings with minimal degradation in neural network accuracy. The adaptive approximation mechanism allows scalability across different workloads and precision requirements. The design is modeled and verified using SystemVerilog to ensure reliability and functionality. CAAM provides an efficient and practical solution for next-generation edge AI and neural network accelerator systems.

Keywords: Approximate Computing, Neural Networks, Adaptive Multiplier, Compressor Architecture, Energy Efficiency, SystemVerilog, Hardware Accelerator, Edge AI, Low Power VLSI, Artificial Intelligence.

INTRODUCTION

Artificial Intelligence and Machine Learning technologies are rapidly transforming modern computing systems. Neural networks serve as the core computational engine behind applications such as image recognition, speech processing, autonomous vehicles, and natural language understanding. These applications require extensive arithmetic operations, particularly multiplication, which contributes significantly to power consumption and hardware complexity. Traditional exact multipliers provide high computational accuracy but consume large amounts of area and energy. In energy-constrained environments such as edge devices and embedded systems, such resource requirements become a major challenge.

Approximate computing has emerged as an effective technique to reduce computational overhead by allowing small acceptable errors in calculations. The proposed Compressor-Based Adaptive Approximate Multiplier (CAAM) utilizes error-tolerant compressors to achieve significant power savings and performance improvements. The architecture minimizes switching activity and shortens critical path delay while maintaining high output quality. Its adaptive nature allows designers to control approximation levels based on application requirements. The design supports multiple neural network workloads and varying precision demands.

SystemVerilog is used for accurate hardware modeling, simulation, and verification of the architecture. The proposed approach enhances energy efficiency without severely affecting neural network accuracy. Furthermore, the architecture reduces hardware cost through lower area utilization. The flexibility and scalability of CAAM make it suitable for future AI hardware accelerators. Thus, CAAM effectively bridges the gap between computational efficiency and performance in modern AI systems.

LITERATURE SURVEY

Several researchers have explored approximate multiplier architectures to improve energy efficiency in digital systems. In 2020, Uppugunduru Anil Kumar and Syed Ershad Ahmed proposed compressor-based approximate multipliers for media processing applications, achieving notable reductions in area and power consumption. In 2021, researchers developed an ultra-compact imprecise 4:2 compressor that significantly reduced delay, transistor count, and energy consumption using deep nanoscale technologies.

Another study introduced constant-carry approximate compressors for low-power and high-speed multipliers suitable for error-resilient applications such as image processing and DCT operations. In 2022, the HEAM architecture optimized approximate multipliers specifically for deep neural networks by minimizing average computation errors. This approach achieved substantial reductions in area, power consumption, and delay while maintaining neural network accuracy. Researchers have also emphasized workload-aware optimization techniques that adapt multiplier behavior according to application requirements. A comprehensive survey published in 2023 reviewed various approximate multiplier designs ranging from algorithm-level techniques to circuit-level implementations. The survey highlighted the importance of approximate computing in machine learning and edge computing systems. Existing studies demonstrate that approximation techniques can effectively balance accuracy and hardware efficiency.

However, many existing solutions lack adaptability and dynamic precision control. Most architectures focus on fixed approximation strategies that cannot respond to varying workloads.

EXISTING SYSTEM

The existing multiplier architectures used in neural network hardware primarily depend on exact computation techniques such as Array Multipliers, Wallace Tree Multipliers, and Booth Multipliers. These architectures provide high computational accuracy but require significant hardware resources. Large area utilization and high switching activity increase overall power consumption. As neural networks become more complex, these multipliers contribute substantially to energy overhead.

Several approximate multipliers have been introduced to address these challenges by sacrificing small amounts of accuracy. However, most existing approximate designs use fixed approximation methods and cannot dynamically adjust their behavior. They often fail to balance power efficiency, accuracy, and delay across diverse neural network workloads. Furthermore, static architectures limit flexibility for different application requirements. High hardware complexity also increases implementation costs. Therefore, existing systems require a more adaptive and energy-efficient multiplier architecture for modern AI applications.

PROPOSED SYSTEM

The proposed system introduces a Compressor-Based Adaptive Approximate Multiplier (CAAM) designed specifically

for neural network applications. The architecture divides the multiplication process into Exact, Approximate, and Truncated regions. The Exact region accurately computes critical higher-order bits to maintain output quality. The Approximate region employs adaptive compressors that simplify hardware and reduce power consumption.

A correction logic module is incorporated to minimize error accumulation and improve reliability. The Truncated region removes selected least significant partial products to further reduce area and energy usage. Multi-level compression techniques efficiently reduce partial products before final addition. The adaptive approximation mechanism enables dynamic trade-offs between accuracy and efficiency. The design significantly lowers switching activity and critical path delay. As a result, CAAM achieves improved performance, lower power consumption, and reduced hardware cost for AI accelerators.

SYSTEM ARCHITECTURE

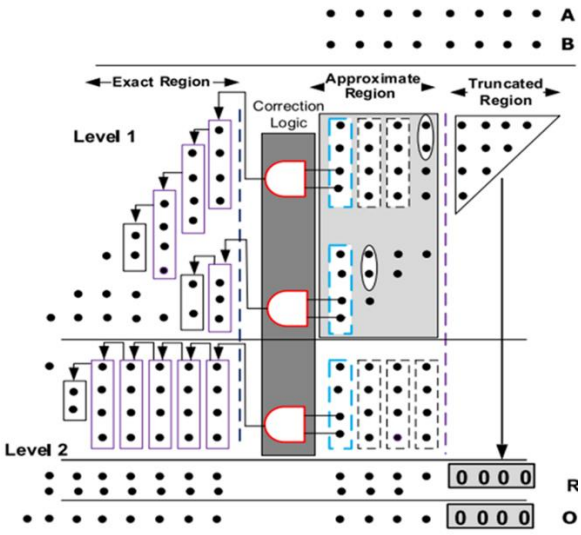

Fig:1 System Architecture

The CAAM architecture consists of several functional blocks working together to perform efficient multiplication. Initially, partial products are generated from the input operands A and B. These partial products are distributed across Exact, Approximate, and Truncated regions. The Exact region computes significant bits accurately to preserve output precision.

The Approximate region utilizes compressor-based circuits that reduce hardware complexity and delay. A correction logic unit is placed between exact and approximate sections to compensate for approximation errors. Multiple compression stages progressively reduce partial products into two final rows. These rows are processed using a fast adder to generate the final product. The Truncated region removes low-significance partial products to save power and area. This

hierarchical architecture effectively balances accuracy, performance, and hardware efficiency for neural network processing.

METHODOLOGY DESCRIPTION

The methodology begins with the design of an adaptive approximate multiplier architecture using SystemVerilog. The multiplication operation starts by generating partial products from two binary input operands. These partial products are divided into exact, approximate, and truncated regions based on significance levels. Exact computation is performed for higher-order bits to preserve output quality. Adaptive compressors are employed in intermediate regions to reduce hardware complexity and power consumption.

The approximation level can be adjusted according to application requirements and workload characteristics. Correction logic is incorporated to control error propagation and improve computational reliability. The architecture is modeled at the Register Transfer Level (RTL) using SystemVerilog constructs. Functional verification is carried out using comprehensive testbenches and simulation environments. Compilation and simulation are performed using QuestaSIM tools. Various test cases are applied to evaluate functionality under different input conditions. Performance metrics such as delay, area utilization, power consumption, and accuracy are analyzed.

Code coverage techniques ensure complete verification of all modules and signal transitions. Results are compared against conventional exact multipliers to quantify improvements. Finally, the architecture is validated for neural network applications, demonstrating its suitability for energy-efficient AI hardware accelerators.

RESULTS AND DISCUSSION

Simulation and evaluation results demonstrate the effectiveness of the proposed CAAM architecture. The design achieves significant reductions in power consumption by minimizing switching activity through adaptive compressors. Critical path delay is reduced, resulting in higher operational speed. Area utilization is also decreased due to simplified compressor structures and partial product truncation. Experimental analysis indicates that only minor accuracy degradation occurs despite approximation techniques. The adaptive approximation mechanism allows optimization based on workload requirements.

Neural network inference tasks maintain competitive classification accuracy while benefiting from energy savings. Correction logic effectively controls error

accumulation and enhances output reliability. Compared to conventional exact multipliers, CAAM provides a superior balance between accuracy and efficiency. These results confirm that the proposed architecture is highly suitable for edge AI and low-power neural network accelerators.

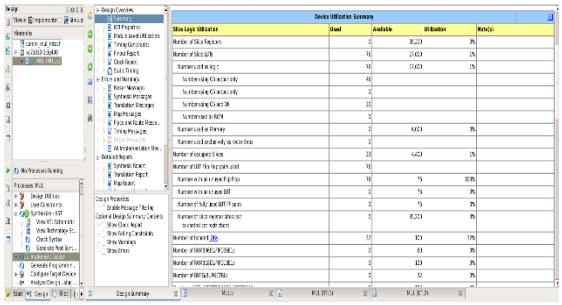

Fig 2: Area

The proposed CAAM architecture significantly reduces hardware area by replacing exact compressors with approximate compressors and truncating less significant partial products. This reduction in logic complexity lowers silicon area utilization and decreases overall hardware implementation cost.

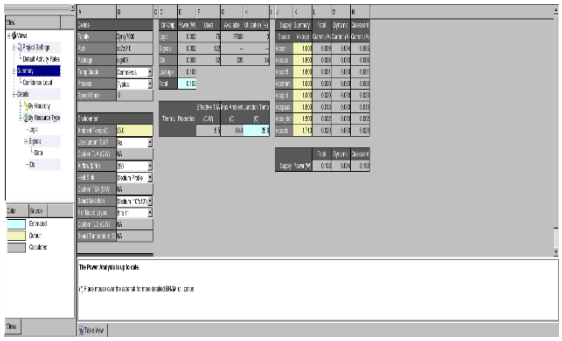

Fig 3: Power Consumption

CAAM achieves lower power consumption by minimizing switching activity within the multiplier circuit through adaptive approximate computation. The reduced number of logic gates and simplified compressor structures contribute to substantial energy savings, making it suitable for low-power AI applications.

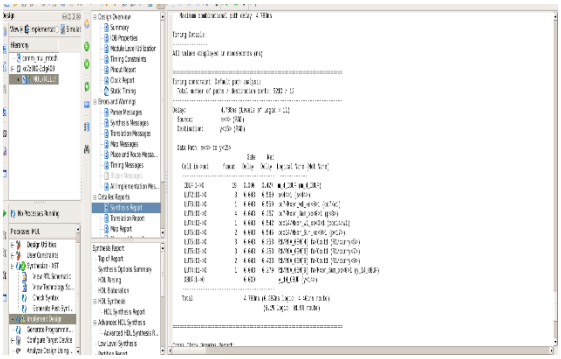

Fig 4: Delay

The architecture reduces critical path delay by using simplified approximate compressors instead of complex exact arithmetic units. As a result, multiplication operations are completed faster, improving overall system speed and throughput for neural network accelerators.

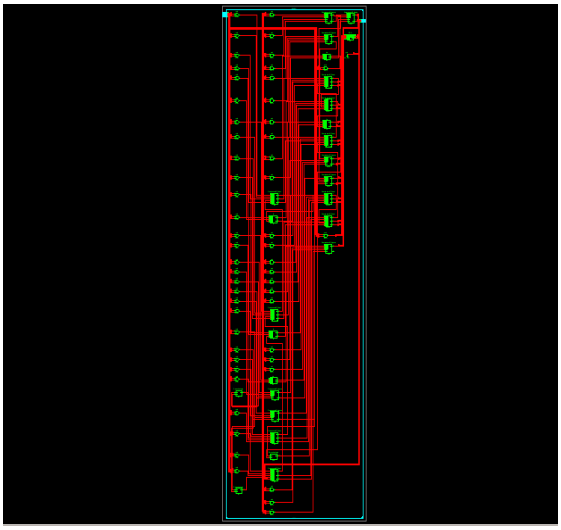

Fig 5: RTL view

The RTL View represents the internal hardware architecture of the CAAM multiplier, showing the flow of data between partial product generation, compression, correction, and final addition stages.

CONCLUSION

The Compressor-Based Adaptive Approximate Multiplier (CAAM) provides an efficient hardware solution for modern neural network applications. By employing compressor-based approximation techniques, the architecture significantly reduces power consumption and hardware complexity. The adaptive design allows flexible control of approximation levels according to application requirements. Reduced switching activity and shorter critical paths improve overall performance. The inclusion of correction logic ensures acceptable computational accuracy despite approximation.

Comprehensive verification using SystemVerilog and code coverage techniques confirms design reliability. Experimental results demonstrate substantial energy savings with minimal impact on neural network performance. The architecture successfully balances power, area, speed, and accuracy requirements. Its scalability makes it suitable for various AI workloads and hardware platforms. Therefore, CAAM represents a promising next-generation multiplier architecture for energy-efficient neural network accelerators.

REFERENCES

1. Shafique, M. et al., An Energy-Efficient Approximate Multiplier for Error-Tolerant Applications, IEEE Transactions on VLSI Systems, 2021.
2. Jaiswal, A. and Raghuvanshi, R., Design of Low Power Approximate 4:2 Compressor for Multiplier, IJCET, 2019.
3. Venkatachalam, N. and Shahana, S., An Area and Power-Efficient Approximate Multiplier Using Novel 4:2 Compressor, IJIRCCE, 2020.
4. Hashemi, S. et al., DRUM: A Dynamic Range Unbiased Multiplier for Approximate Applications, DATE Conference, 2015.
5. Suresh, K. and Devi, V. S., Power-Efficient Adaptive Approximate Multiplier for Deep Neural Networks, Springer Journal, 2022.
6. Adaptive Approximate Multiplier Using Dual Accuracy, Integration Elsevier, 2022 – Uses 32nm technology to balance accuracy and power but involves complex control logic.

7. Approximate Multipliers for Edge AI, IEEE Transactions on Circuits & Systems, 2021 – Employs 16nm arithmetic for edge inference with slight accuracy loss.

8. Reconfigurable Approximate Multipliers, Springer - JETTA, 2020 – Offers runtime accuracy-performance trade-offs in 45nm but needs extra area for reconfigurability.