International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

ENERGY-EFFICIENTAPPROXIMATE CIRCUITS THAT CAN BE ADJUSTED FOR SELF-POWERED AI AND AUTONOMOUS EDGE COMPUTING SYSTEMS

[1]K. SWAPNA, [2]Dr.V.T. VENKATESWARLU, [3]Y. VEERNA, [4]M. REKHA, [5]T. AMULYA
[2]Professor, ECE Dept, RISEKrishna Sai Prakasam Group of Institutions, Valluru, AP
[1,3,4,5]B. Tech final year students,ECE Dept, RISEKrishna Sai Prakasam Group of Institutions, Valluru, AP
[1]kuppakulaswapna15gmail.com,[2]venki.vt@gmail.com,[3]veernayasarapuveerna@gmail.com,[4]mancharekha15@gmail.com,[5]amulyathota903@gmail.com

ABSTRACT

The growing reliance of edge artificial intelligence (AI) systems on self-powered or energy-constrained hardware platforms necessitates extremely efficient processing units without sacrificing functional precision. Optimizing multiply-and-accumulate (MAC) operations is crucial for sustainable edge intelligence since they account for the majority of energy consumption in contemporary AI workloads. In order to enable dynamically adjustable accuracy–energy trade-offs for MAC architectures, this work presents unique tunable approximation compressor designs. Low-complexity approximation techniques and tenability knobs that enable real-time configuration based on workload precision needs or available energy budgets are features of the suggested compressors. The designs maintain sufficient accuracy for error-resilient AI workloads while achieving significant savings in power consumption, latency ,and silicon area when incorporated into MAC units. The designs' effectiveness and viability are validated by hardware implementation and assessment using Xilinx Vivado, indicating their appropriateness for energy-adaptive computation in next- generation edge AI and autonomous sensing platforms.

Index Terms— Approximate Computing, Tunable Approximation, Compressor Design, multiply–Accumulate Units, EdgeAI, Low-Power Hardware, Energy-Constrained Systems, Vivado Implementation

Received: 06-01-2026 Accepted: 13-02-2026 Published: 20-02-2026

I INTRODUCTION

The rapid advancement of Edge Artificial Intelligence (Edge AI) has enabled intelligent data processing directly on resource-constrained edge devices such as IoT nodes, wearable electronics, autonomous sensors, and mobile platforms. Unlike cloud-based systems, edge devices must operate under strict power, area, and thermal constraints while delivering real-time performance. Consequently, the development of energy-efficient hardware architectures has become a fundamental requirement for deploying AI workloads at the edge.

At the core of most AI and signal-processing applications are multiply–accumulate (MAC) operations, which dominate both execution time and energy consumption in convolutional neural networks (CNNs), deep neural networks (DNNs), and filtering systems. A MAC unit primarily consists of a multiplier followed by an adder, where the multiplier is the most computationally intensive component. As model complexity and data precision increase, conventional exact arithmetic implementations become increasingly inefficient for edge platforms due to their high switching activity and logic depth.

Approximate computing has emerged as a promising paradigm to address these limitations by exploiting the inherent error tolerance of AI inference workloads. Many machine learning applications can tolerate minor numerical inaccuracies without significantly affecting overall system accuracy or perceptual quality. By relaxing strict correctness constraints, approximate arithmetic circuits can achieve substantial reductions in power consumption, silicon area, and propagation delay. Among various arithmetic components, approximate adders and multipliers have received considerable research attention. In particular, the partial product reduction stage in multipliers contributes

significantly to switching activity and critical path delay. Optimizing this stage therefore offers substantial potential for improving overall energy efficiency in Edge AI hardware accelerators.

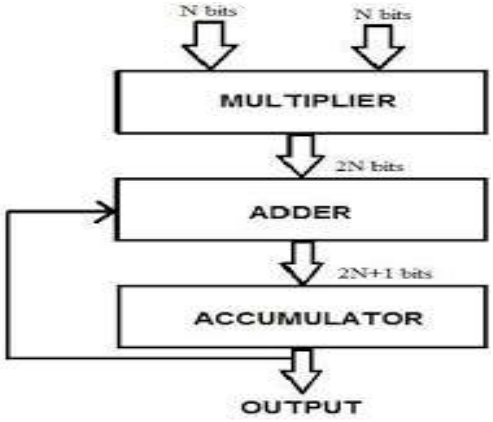


Fig.1 MAC Block Diagram

II EXISTING METHOD

In conventional approximate computing systems, the approximation level is fixed at design time. These circuits (such as approximate adders, multipliers, and logic blocks) are optimized for reduced power consumption and area, but they lack adaptability. This makes them suitable for error-tolerant applications but less effective in self-powered AI and autonomous edge computing systems, where energy availability fluctuates and workload accuracy requirements vary.

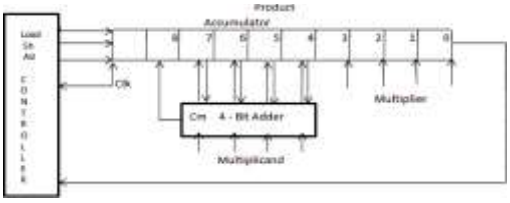


Fig.2 Multiplier Block Diagram

Conventional Multiplier Operation:

Consider the multiplication of two binary numbers:
101 x 10011

The process involves the following steps:

- Shifter Stage:
  - Input: 101
  - Operation: Multiple shift operations are performed.
- Adder/Buffer Stage:
- Partial products are generated and accumulated using adders and buffers.
- Example intermediate results:
- 101 → Shifter
- 1111 → Adder, Buffer
- 101 → Shifter
- 1011111 → Adder, Buffer
- This sequence demonstrates that multiple shifters and adders are required to complete the multiplication.

Draw backs in existing System:

- Conventional MAC units require more hardware resources, which increases the overall chip area.
- Performing fully accurate computations consumes higher power, reducing battery life in edge devices.
- Long carry propagation paths in arithmetic circuits cause more delay, affecting system speed and efficiency.

III PROPOSED METHOD

To overcome the drawbacks of existing MAC designs, we use a compressor-based MAC architecture, which reduces area, power consumption, and computational delay.

This paper provides tunable approximate compressor topologies featuring 2:1, 3:2, 4:2, and 5:3 compressors that may dynamically switch between two approximation modes (AM-1 and AM-2). By altering the critical logic path and combining several logic operations into a single circuit that is managed by a choose signal, the tunability is attained. The compressor alternates between a hardware-optimized simplified logic function and an intermediate logic function by turning on or off particular logic blocks. This method reduces the partial product creation latency while enabling a real-time trade-off between accuracy and energy economy. The suggested compressors successfully lower complexity and power consumption while preserving adaptability for energy-efficient MAC and multiplier designs in Edge-AI applications.

PROPOSEDTUNABLEAPPROXIMATE2:1COMPRESSOR(PTAXC21)

The PTAxC21 integrates two distinct logic functions corresponding to different approximation levels, enabling real-time switching between modes. This design allows the circuit to adapt dynamically to energy and accuracy requirements.

$$X = \begin{cases} O_0^{t-21} = p_0 + p_1, & \text{when } S = 1 \\ O_0^{t-21} = p_0, & \text{when } S = 0 \end{cases} \quad (1)$$

PROPOSED TUNABLE APPROXIMATE 3:2 COMPRESSOR (PTAXC32)

International Journal of Engineering Research and Science & Technology

The PTAxC32 integrates two logic functions corresponding to different approximation levels, enabling dynamic switching between accuracy and energy-optimized modes. The design modifies only the critical path associated with output O_{t-32}^1, while maintaining a stable secondary output O_{t-32}^0.

$$Y = \begin{cases} O_0^{t-32} = p_0 + p_1 \\ O_1^{t-32} = p_0p_1 + p_2, & \text{when } S = 1 \\ O_0^{t-32} = p_0 + p_1 \\ O_1^{t-32} = p_2, & \text{when } S = 0 \end{cases}$$

(2)

PROPOSED TUNABLEAPPROXIMATE 4:2 COMPRESSOR (PTAXC42)

The PTAxC42 integrates two logic functions corresponding to different approximation levels, enabling dynamic switching between accuracy and energy-optimized modes. The design exploits the common logic function (p_0+p_1) to construct an integrated logic path for the output O_{t-42}^1.

$$Z = \begin{cases} O_0^{t-42} = (p_0p_1) + (p_2 + p_3) \\ O_1^{t-42} = (p_2p_3) + (p_0 + p_1), & \text{when } S = 1 \\ O_0^{t-42} = (p_0p_1) + (p_2 + p_3) \\ O_1^{t-42} = (p_0 + p_1), & \text{when } S = 0 \end{cases}$$

(3)

PROPOSED TUNABLE APPROXIMATE 5:3 COMPRESSOR (PTAXC53)

The PTAxC53 integrates two logic functions corresponding to different approximation levels, enabling dynamic switching between accuracy and energy-optimized modes. The design leverages the common logic function block (p_2+p_3) to construct an integrated logic path for the output O_{t-53}^0.

$$W = \begin{cases} O_0^{t-53} = p_0p_1 + p_2 + p_3 \\ O_1^{t-53} = (p_2p_3) + p_4, & \text{when } S = 1 \\ O_2^{t-53} = p_0 + p_1 \\ O_0^{t-53} = p_2 + p_3 \\ O_1^{t-53} = (p_2p_3) + p_4, & \text{when } S = 0 \\ O_2^{t-53} = p_0 + p_1 \end{cases}$$

(4)

Approximate Multiplier and MAC Unit

Multiplication is a fundamental arithmetic operation in deep neural networks (DNNs), yet conventional multipliers suffer from high power consumption, latency, and area overhead, making them unsuitable for low-power edge devices. To overcome these limitations, we propose an approximate multiplier (PAxC MUL) and a tunable approximate multiplier (PTAxC MUL), both of which enhance energy efficiency, resource utilization, and performance. The design methodology introduces approximation in the partial product reduction stage of an N\times N multiplier, where the partial product matrix (PPM) is divided into three regions: an approximate section reduced using proposed compressors, a hybrid section reduced using both approximate and exact adders to balance carry propagation, and an exact section reduced using conventional adders. This structured approach achieves significant energy savings and latency reduction while maintaining acceptable accuracy, and the resulting fixed and tunable multipliers are further extended to design an approximate MAC unit suitable for self-powered AI and autonomous edge computing systems.

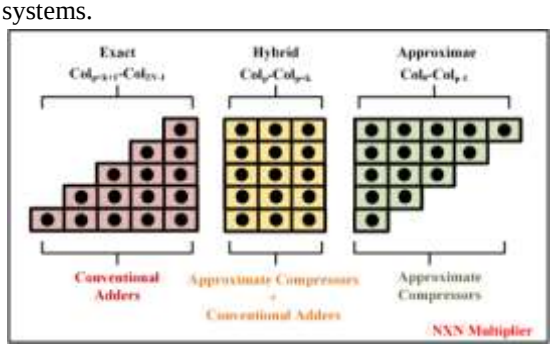


Fig.3 Exact, Hybrid and Approximate sections of a Partial Product Matrix(PPM) in an N×N multiplier.

In the proposed approximate multiplier design, the partial product matrix (PPM) is divided into three distinct regions—approximate, hybrid, and exact—to balance energy efficiency with computational accuracy. In the approximate section, the reduction of partial products is carried out using the proposed approximate compressors, while in the exact section conventional full adders and half adders are employed to ensure precision. Between these two lies the hybrid section, which combines approximate compressors and exact adders to maintain proper carry flow during the transition from approximate to exact computation. The PPM is progressively reduced to two rows through multiple reduction stages, with approximation introduced only in the first and second stages to minimize its impact on the final result. Finally, the last two rows are reduced using a ripple carry adder, completing the multiplication process. This methodology is demonstrated through the

International Journal of Engineering Research and Science & Technology

implementation of a 16×16 approximate multiplier (PAxC MUL-16×16) and its tunable counterpart (PTAxC MUL-16×16), where the PPM is generated by dot operations between input operands and reduced stage by stage, achieving substantial energy savings while maintaining acceptable accuracy levels.

IV SIMULATIONMETHODOLOGYANDRESULTS

In this section, we present a comprehensive analysis of the proposed approximate compressors (PAxC21, PAxC32, PAxC42, PAxC53), tunable approximate compressors (PTAxC21, PTAxC32, PTAxC42, PTAxC53), approximate multiplier (PAxCMUL16×16), tunable approximate multiplier (PTAxC MUL 16 × 16), approximate MAC unit (PAxC MAC-16) and tunable approximate MAC (PTAxC MAC16)

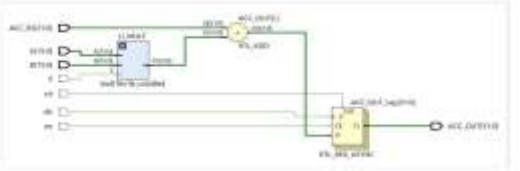


Fig.4 MAC ELABORATION DESIGN

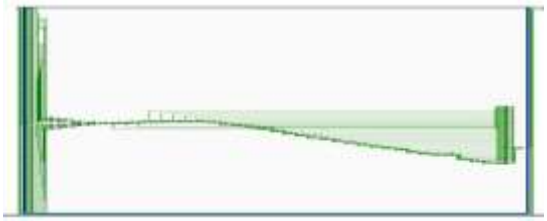


Fig.5 MAC ELABORATION DESIGN



Fig.6 MULTIPLIER SIMULATION OUTPUT



Fig.7 MULTIPLIER SIMULATION OUTPUT



Fig.8 MAC SIMULATION OUTPUT



Fig.9 MAC SIMULATION OUTPUT

Utilization | Post-Synthesis | Post-Implementation

Graph | Table

Resource	Utilization	Available	Utilization %
LUT	374	41000	0.91
FF	32	82000	0.04
IO	97	300	32.33
BUFG	1	32	3.13

Fig.10 AREA OF MAC

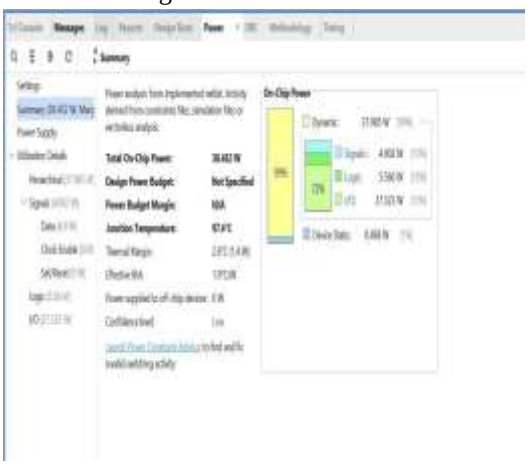


Fig.11 MAC POWER

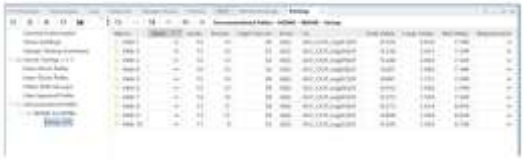


Fig.12 MAC DEALY

PERFORMANCE COMPARISONAND DISCUSSION

Below Table compares the proposed design with an existing architecture in terms of area, power, and delay.

Parameter	Proposed Design	Existing Design
Area(LUTs)	374	700
Power(Watts)	38.452	55.508
Delay(ns)	9.5	24.8

COMPRESSION OF EXISTING AND PROPOSED MAC UNITS

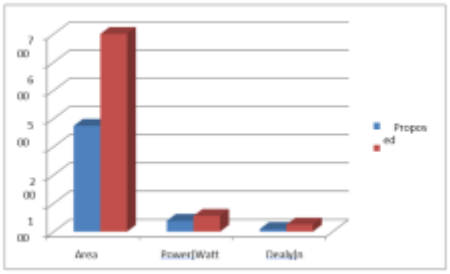


Fig.13 PERFORMANCE COMPARISON AND DISCUSSION

The proposed architecture achieves a 46.6% reduction in silicon area through simplified compressor logic and minimized carry propagation paths. Power consumption is reduced by approximately 30.7%, primarily due to lower switching activity and decreased logic depth within the partial product reduction stage. Furthermore, the critical path delay is reduced by 61.7%, resulting in significantly faster computation and enhanced throughput performance.

These improvements clearly demonstrate that the introduction of tunable approximation at the compressor level substantially enhances the efficiency of the MAC unit. By strategically relaxing arithmetic precision in non-critical computation paths, the design achieves an optimal balance between accuracy and hardware efficiency. Consequently, the proposed architecture is highly suitable for energy-constrained and performance-critical Edge AI systems, where power efficiency, reduced silicon footprint, and low-latency computation are essential design requirements.

V CONCLUSION

This work presented the design and evaluation of novel tunable approximate compressor architectures for energy-constrained multiply–accumulate (MAC) operations targeting Edge AI applications. The proposed approach introduces controlled approximation at the compressor level to significantly reduce hardware complexity, power consumption, and critical path delay while maintaining acceptable computational accuracy for error-tolerant workloads. The architecture was implemented and compared with an existing conventional design in terms of area, power, and delay to validate its efficiency.

Experimental results demonstrate substantial improvements across all key VLSI performance metrics. The proposed architecture requires only 374 LUTs, achieving a 46.6% reduction in area compared to the existing design, which occupies 700 LUTs. This area reduction is primarily attributed to simplified compressor logic and minimized carry propagation paths. In terms of power consumption, the proposed design consumes 38.452 W, representing a 30.7% reduction compared to the existing implementation, which consumes 55.508 W. The lower power dissipation confirms the effectiveness of approximate computation in reducing switching activity and logic depth—an essential requirement for battery-operated and energy-constrained edge devices.

Furthermore, the proposed architecture achieves a significantly lower delay of 9.5 ns compared to 24.8 ns for the conventional design, resulting in a 61.7% improvement in speed. The reduced critical path delay highlights the suitability of the proposed compressors for high-throughput MAC operations required in real-time Edge AI workloads.

REFERENCES

[1] J. Han and M. Orshansky, “Approximate computing: An emerging paradigm for energy-efficient design,” in *Proc. IEEE European Test Symposium (ETS)*, 2013, pp. 1–6.

[2] O. Akbari, M. Kamal, A. Afzali-Kusha, and M. Pedram, “Dual-quality 4:2 compressors for utilizing in dynamic accuracy configurable multipliers,” *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 25, no. 4, pp. 1352–1361, Apr. 2017.

[3] R. Zindani, M. Kamal, A. Afzali-Kusha, and M. Pedram, “ROBA multiplier: A rounding-based approximate multiplier for high-speed yet energy-efficient digital signal processing,” *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 24, no. 2, pp. 393–404, Feb. 2016.

[4] S. Narayanamoorthy, H. A. Moghaddam, Z. Liu, T. Park, and N. S. Kim, “Energy-efficient approximate multiplication for digital signal processing and classification

applications," *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 23, no. 6, pp. 1180–1184, Jun. 2015.

[5] T. Raimon, K. Ling Subramanian, and S. Bhanja, "Probabilistic error modeling for nano-domain logic circuits," *IEEE Transactions on Nanotechnology*, vol. 8, no. 4, pp. 508–516, Jul. 2009.

[6] Vikram, S. (2025). Modernizing Data Infrastructure: How AI and ML are Transforming SQL and NoSQL Usage in Distributed Manufacturing.

[7] Charan Nandigama, N. (2020). An Integrated Data Engineering and Data Science Architecture for Scalable Analytical Warehousing and Real-Time Decision Systems. International Journal of Business Analytics and Research (IJBAR).

[8] Todupunuri, A. (2024). Develop Machine Learning Models to Predict Customer Lifetime Value for Banking Customers, Helping Banks Optimize Services. International Journal of All Research Education & Scientific Methods, 12(10), 1254–1259. https://doi.org/10.56025/ijaresm.2024.12102412 54

[9] Ganji, M. (2025). Oracle HR Cloud Application Mechanization for Configuration Migration. INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH, 13(2). https://doi.org/10.56975/ijedr.v13i2.301303

[10] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys*, vol. 48, no. 4, Article 62, Mar. 2016.

[11] Bajarang Bhagwat, V. (2023). Optimizing Payroll to General Ledger Reconciliation: Identifying Discrepancies and Enhancing Financial Accuracy. JOURNAL OF ADVANCE AND FUTURE RESEARCH, 1(4). https://doi.org/10.56975/jaafr.v1i4.501636

[12] Mallick, P. (2022). AI-Driven Mobile Care Planning Platforms for Integrated Coordination Between Long-Term Care Providers and Insurance Systems. Available at SSRN 6066586.

[13] Y. Yazdanbakhsh et al., "ALWANN: Adaptive layer-wise approximation for neural network accelerators," *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 27, no. 4, pp. 834–847, Apr. 2019.

[14] A. Momeni, J. Han, P. Montuschi, and F. Lombardi, "Design and analysis of approximate compressors for multiplication," *IEEE Transactions on Computers*, vol. 64, no. 4, pp. 984–994, Apr. 2015.

[15] C.-H. Chang, J. Gu, and M. Zhang, "Ultra-low-voltage low-power CMOS 4–2 and 5–2 compressors for fast arithmetic circuits," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 10, pp. 1985–1997, Oct. 2004.