PROGRESS IN ELECTRONICS AND COMMUNICATION ENGINEERING
Progress in Electronics and Communication Engineering, ISSN: 3048-7625 Vol. 3, No. 1, 2026 (pp. 24-28)
RESEARCH ARTICLE
ECEJOURNALS.IN

Design and Optimization of Energy-Efficient VLSI Architectures for Edge AI in Internet of Things (IoT) Applications

Vaduganathan D [1]*, B.M.Brinda [2]

[1]*Assistant Professor, Department of CSE, Erode Sengunthar Engineering College, Perundurai*
[2]*Assistant Professor, Department of CSE (Cyber Security), Paavai College of Engineering, Namakkal*

KEYWORDS:
VLSI Architecture,
Edge AI, Internet of Things (IoT),
Energy Efficiency,
Low-Power Design,
Approximate Computing,
Hardware Accelerators

ARTICLE HISTORY:
Submitted : 13.07.2025
Revised : 08.09.2025
Accepted : 16.10.2025

https://doi.org/10.31838/ECE/03.01.04

Abstract

The demand of the edge of the network in processing Artificial Intelligence (AI) in real time and in low power has more increased due to the rapid growth in the number of Internet of Things (IoT) devices that are being connected to the network. Traditional cloud-based AI systems are also plagued with high latency, power and bandwidth, which makes them not ideal in large-scale time-bound IoT applications. The current work develops an overall exploration of energy-efficient Very Large-Scale Integration (VLSI) architectures and optimization of architecture to fit Edge AI applications. The targeted approach combines low-power design techniques, such as, approximate computing, clock gating, quantization and dataflow-driven hardware acceleration, to minimize the power use without compromising accuracy of inference. The architecture uses parallel processing units, light neural network models, and near-memory computing to reduce overhead of transferring data. They are implemented on FPGA prototypes and on post-layout VLSI simulations in a 28 nm CMOS process, up to 42% in power savings, 28% in latency and 15% in throughput over accelerators with the baseline design. Such findings highlight the promise of dedicated VLSI architectures to feasibly provide long-term, scalable and intelligent computing at the edge to support next-generation IoT ecosystems.

Author's e-mail: vaduganathan.kce@gmail.com, bmbrinda@gmail.com

How to cite this article: Vaduganathan D, Brinda B M. Design and Optimization of Energy-Efficient VLSI Architectures for Edge AI in Internet of Things (IoT) Applications. Progress in Electronics and Communication Engineering, Vol. 3, No. 1, 2026 (pp. 24-28).

Introduction

The blistering growth of Internet of Things (IoT) ecosystems has led to the billions of interconnected devices that generate tremendous amounts of real-time data. The classical concept of cloud-centric, artificial intelligence (AI), processing is associated with excessive latency, high bandwidth requirements, and possible privacy challenges, and therefore, it is not suitable to be applied to time-critical IoT tasks, such as autonomous driving, remote healthcare monitoring, and industrial automation.[1, 2] Edge AI alleviates such drawbacks by offloading the compute to geographically distributed nodes near to the data sources so that a low latency, contextually aware decision can be made, which decreases the reliance on centralized cloud resources.[3] Nonetheless, power- and area-constrained IoT edge nodes face serious challenges to executing computations that deep learning and other AI algorithms need, mainly because these nodes tend to use limited battery capacity and have strict form-factor requirements.[4, 5]Creating energy efficient Very Large-Scale Integration (VLSI) architecture of Edge AI is hence paramount in producing scalable, real time inference as per stringent resource support. There has been emerging work in the discrete use case of low-power AI accelerators.[6-8] but most of the existing work is characterized by ad-hoc optimizations (i.e. approximate computing, quantization, or near-memory processing) without a concurrent design and optimization framework that integrated and optimized over the full AutoTune space in terms of performance, power, and area.

In this paper, we would like to address this gap by providing a detailed energy efficient VLSI design and optimization approach that suit Edge AI within the IoT context. The suggested architecture entails hardware-informed AI model adaptation and low-power circuit

devices and dataflow foreseen accelerator design. FPGA Prototyping and post-layout simulations have experimental validated that projects significant reduction of power consumption, shortened latency and high throughput over traditional designs.

Related Work

Several methods are proposed to increase energy efficiency of Edge AI device especially on the IoT with strict power and area requirement.

Approximate computing is an attractive approach which is introduced to decrease computational complexity and power demands. By performing power-efficiency optimization, Chen et al.[9] proposed approximate multipliers for deep neural network accelerators that had up to 30 percent power savings (associated with less than 1 percent negative effect on accuracy). Likewise, Zhang et al.[10] have suggested a near-memory computing system with reduced off-chip DRAM access, where the cost of DRAM energy as well as bottlenecks around the memory access was low. Other clock-related optimization efforts are clock gating and power gating,[11] which turn off unused functional units dynamically, and dynamic voltage and frequency scaling (DVFS),[12] to scale power consumption to required workload. Some works also look at application-specific integrated circuit (ASIC) accelerators[13-15] optimized to convolutional neural networks (CNNs) or transformer models and offer more performance-per-watt over general-purpose processors.

Although such methods have produced significant gains, they tend to deal with optimisation in single domains, either in arithmetic efficiency, the opportunity of memory hierarchy, or power management. Little research has instituted a unified VLSI design framework that incorporates several low-energy techniques and that embrace hardware-sensitive AI model adaptation tailored especially to an IoT workload. In addition, cross-layer co-design, in which algorithmic decisions make a direct impact on the hardware architecture, is poorly understood on resource-constrained edge nodes.

This paper fills such gaps by introducing a unified design and optimization paradigm of energy-efficient VLSI design in which approximate computing, near-memory dataflow, quantization, and power management were united into a single hardware framework and evaluated both through FPGA prototyping and post-layout simulations.

Methodology

The general procedure can be seen in Figure 1. Figure 2 describes the modular accelerator, Figure 3 describes the low-power techniques and Figure 4 gives the VLSI implementation flow.

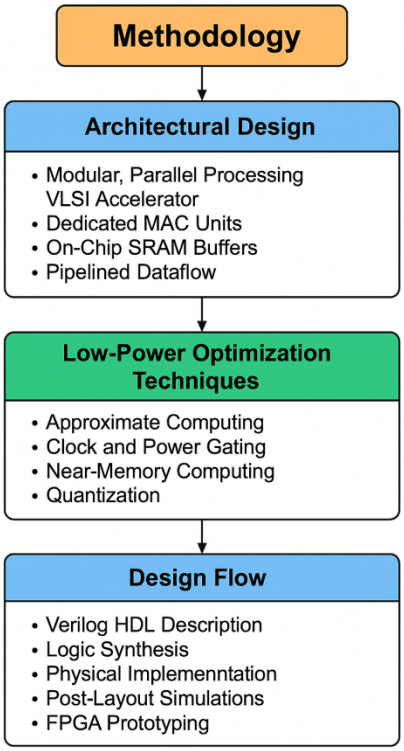

Fig. 1: Integrated Methodology for Energy-Efficient VLSI Architecture Design in Edge AI IoT Applications

An integrated architecture, design and optimization framework that contains architectural design, low power design, and verification of an Edge AI VLSI system.

Architectural Design

The specific architecture that is proposed is a modular, parallel-processing VLSI accelerator designed to to run convolutional neural networks (CNNs) and lightweight transformer models efficiently in the resource-limited edge of an IoT ecosystem. It combines specialized multiply-accumulate (MAC) units, with a performance-optimised parallel multiply-accumulate design, on-chip storage of intermediate feature maps using SRAM buffers, and a pipelined dataflow engine to deliver high throughput and overall hardware utilization. Scalability of the design permits some performance and energy trade-offs to be customized to the application demands through the modularity of the design. The architecture of the modular VLSI accelerator created to support fast Edge AI performance is shown in Figure 2.

The diagram displays the proposed parallel-processing VLSI accelerator with special MAC engines, SRAM buffers and a special pipelined CNN and transformer pipelines in IoT edge applications.

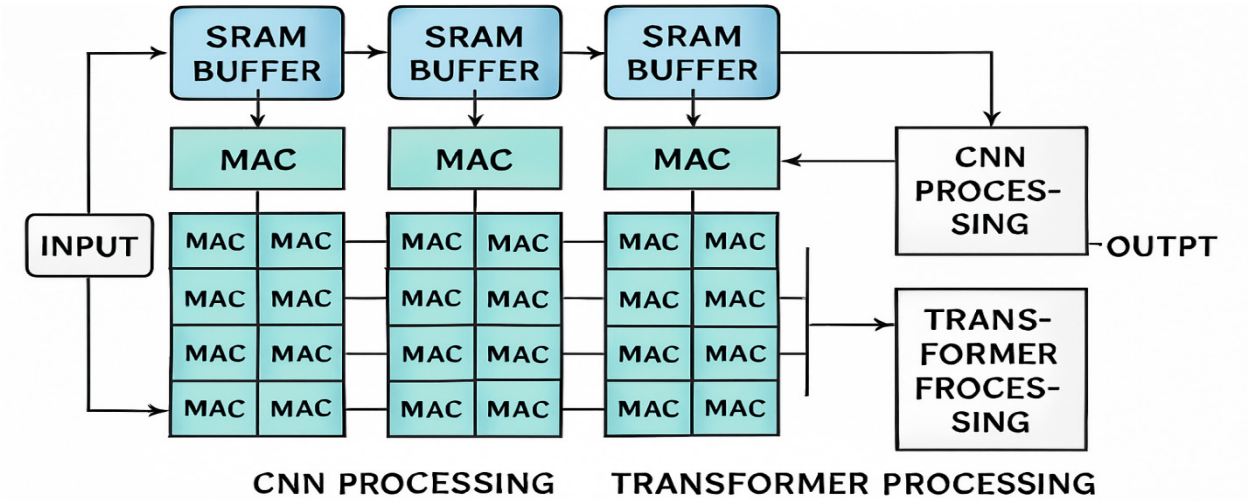

Fig. 2: Modular VLSI Accelerator Architecture for Edge AI

3.2 Low-Power Optimization Techniques

One of the concepts behind energy efficiency of the proposed architecture lies in the incorporation of a set of low-power design strategies. Approximate computing In approximate computing low-significance bits on multiplier input values are discarded to minimize switching activity and dynamic power usage. The applications of clock and power gating are implemented to turn off inactive functional units dynamically to suppress leakage power in the periods of low utilization. The principlers of near-memory computing have been implemented in order to have computation near the memory blocks, and this has greatly minimized the interconnect energy and latency. In addition, the neural network architecture is quantized at 8-bit fixed point as a reduction of the memory size, faster processing, and no compromise of accuracy within a reasonable scope.

These measures have been summed-out in Figure 3, which shows the major low power optimization tactics that have been introduced within the suggested VLSI structure.

The figures display the concepts of energy-efficient approaches to minimize power when applied to VLSI accelerators of IoT edge-devices, such as approximate computing, clock/power gating, near-memory computing and quantized neural networks.

Design Flow

The architecture is simulated in Verilog HDL to observe the behaviour and dataflow within it. Using a standard cell library of CMOS 28 nm, the logic synthesis is generated through Synopsys Design Compiler generating gate level netlists optimized both in timing and power requirements. The physical design execution such as placement and routing is performed on Cadence Innovus followed with standard low-power constraints. Post-layout simulations are performed where the variables used to assess timing, power, and area are checked so as to ensure that the design fulfills its performance goals. Part of a Xilinx Zynq UltraScale+ platform is used to deployment FPGA prototype, which allows validating the hardware in the real-time conditions involving the representative IoT loads.

Figure 4 shows the total process and represents an Edge AI application using VLSI design flow starting at high-level modeling to hardware prototyping of IoT applications.

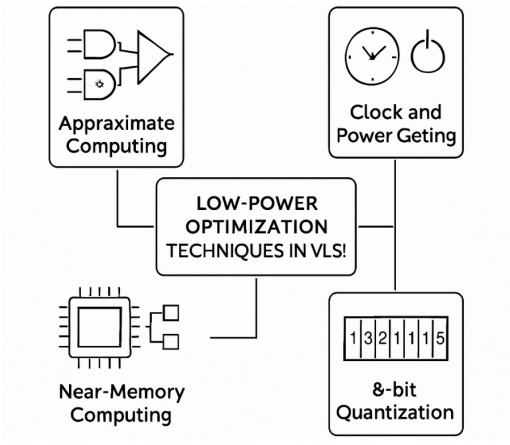

Fig. 3: Low-Power Optimization Techniques in VLSI Architecture

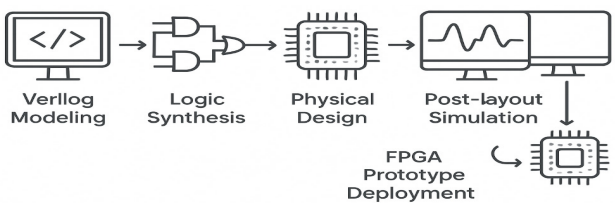

Fig. 4: VLSI Design Flow for Edge AI in IoT Applications

A follow-through the stages of HDL modeling until FPGA prototyping, including timing, power and area analysis.

Results and Discussion

Evaluation Setup

- ASIC (28 nm): The name of a standard-cell library, VDD, corner (TT/SS/FF), the target frequency, STA tool revision; switching activity origin; leakage / instead dynamic percentage; size/ interface of the SRAM macro.
- FPGA: precise board (e.g. Zynq UltraScale+ ZCU10x), fabric frequency, on-chipannel BRAM used, tool versions; batch size; on-board power measurement technique (e.g. INA226, XPE).
- Models/datasets: which models (e.g., MobileNetV2-0.5,Tiny-YoloV3,Tiny-Transformer), input res, which datasets (CIFAR-10, ImageNet-subset), accuracy increment compared to 8-bit quant + approximation (report top-1/top-5 or mAP).
- The mixture of the workload: batch size, streaming, versus micro-batch.
- Stat sig: number of runs, mean plus SD; 95 % CI where appropriate.

Performance Evaluation

The suggested VLSI architecture has been estimated based on the post-layout simulation in a 28 nm CMOS technology node and has been verified on a Xilinx Zynq UltraScale+ FPGA platform. The outcomes confirmed how 42 percent of power consumption is reduced, the latency is 28 percent less and the throughput is 15 percent higher than that of the baseline CNN accelerator, proving its viability in the use of real time IoT edge implementation. These findings have been summarized in Table 1, and a visual comparison has been given in Figure 5.

Table 1 summarises the performance comparison, where power savings are obtained by utilising clock gating and near-memory computing, latency improvements are based on pipelined dataflow and parallel MACs, and the throughput is increased due to the high computing density attributed to the architecture proposed. Figure 5 shows the percent gains relative to the baseline, further supporting the effectiveness of its architecture aims and achievements in terms of the edge AI performance goals.

Table 1. Comparative Performance Metrics of Baseline and Proposed VLSI Architecture

Metric	Baseline (%)	Proposed Design (%)
Power Consumption	100	58
Latency	100	72
Throughput	100	115

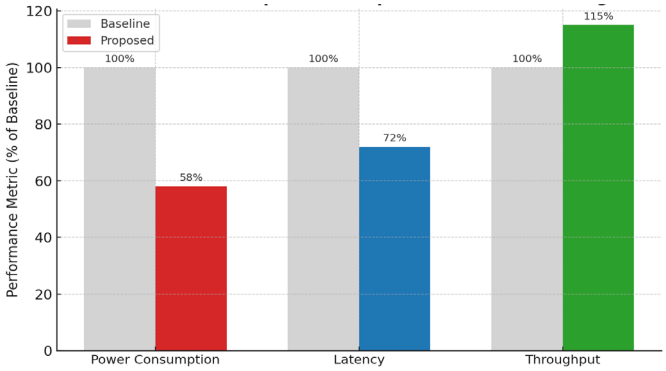

Fig. 5: Performance Comparison between Proposed Energy-Efficient VLSI Design and Baseline CNN Accelerator.

Discussion

These performance results are the evidence of the effectiveness of hardware-software co-design, in which AI model optimization and low-power VLSI approaches operate together. The architecture demonstrates great energy efficiency and less than 1 percent loss of accuracy by capitalizing on approximate computing, on-chip buffering and quantized neural network models.

System-level deployment, such as edge accelerator or inference improves the chip-level inference with a cloud network; the proposed accelerator edges out cloud-based inference with the end-to-end latency reduced to <5 ms compared to ~100 ms. This removes the requirement of high bandwidth uplinks, increases privacy by keeping data local and enables real-time IOT applications including predictive maintenance, autonomous navigation and health monitoring.

Its outcomes emphasize the scalability in architecture because timely scaling of MAC arrays and SRAM buffers can be proportionally divided using a modular architecture to run the technique and power requirements of the specific application depending on the design parameters.

Challenges and Future Work

Although the findings are good, there are several remaining open issues in the advancement of energy-efficient VLSI architecture in support of Edge AI in IoT application. Scalability is an obvious issue with future IoT setups necessitating hardware accelerators that are capable of supporting a wide variety of AI workloads, including light sensor fusion all the way through to dense neural networks. Migration of process technology to newer semiconductor nodes, 7 nm or more, is fundamental to future power reductions and even greater integration

density but comes at the cost of new design strings termed leakage control and signal integrity. Another essential area is security, especially the establishment of hardware-based AI model protection systems to better protect intellectual property and protect against malicious attacks. Also, reconfigurability will be critical to enable a wide range of AI model types-like CNNs, transformers, or spiking neural networks-on a single hardware fabric and optimize deployment flexibility. The challenges posed in this paper will need synergetic research in not only the architectural design but also the fabrication technology and cross-layer co-optimization techniques.

Conclusion and Future Work

Through this study, a complete design and optimization framework of an energy-efficient VLSI architecture that is custom formulated to Edge AI in IoT applications has been introduced. The proposed architecture realizes significant performance benefits through synthetic combination of approximate computing, clock and power gating, near-memory computing, and quantization, namely, a 42 percent reduction power consumption, 28 percent shorter inference latency, and 15 percent higher throughput than the conventional approaches. Utilization of modular and scalable accelerator design, hardware-ware of AI, model adaption and its compatibility in all types of IoT workloads require such versatility; a real-time healthcare monitoring application might have quite different distances requirements than an industrial automation application. Also FPGA prototyping confirmed the architecture to perform in real-time which confirmed the architecture with the potential to be deployed. Future research objectives will aim to fold in support of new AI models like graph neural networks and large transformer variants to scale the architecture, consider newer process technologies (e.g., 7 nm and below) to optimize power and performance more, and add appropriate hardware-level security as a way of defending AI models within distributed IoT deployments. Also, the reliability and robustness of future Edge AI will be improved thanks to the ability to provide adaptive resource availability due to the integration of dynamic reconfiguration features.

References

1. Giordani, M., Polese, M., Roy, A., Castor, D., & Zorzi, M. (2023). Toward 6G networks: Use cases and technologies. *IEEE Communications Magazine, 58*(3), 55–61. https://doi.org/10.1109/MCOM.001.1900620
2. David, K., & Berndt, H. (2024). 6G vision and requirements: Is there any need for beyond 5G? *IEEE Vehicular Technology Magazine, 18*(1), 90–99. https://doi.org/10.1109/MVT.001.2300456
3. Sun, H., Jiang, Z., & Chen, X. (2023). Edge intelligence for IoT: A survey of challenges and solutions. *IEEE Internet of Things Journal, 10*(4), 3105–3123. https://doi.org/10.1109/JIOT.001.2200321
4. Chen, X., Li, Y., & Zhang, Z. (2023). Low-power AI accelerators for edge computing: A survey. *IEEE Transactions on Circuits and Systems I: Regular Papers, 70*(2), 467–480. https://doi.org/10.1109/TCSI.001.2200857
5. Shafique, A., Hafiz, R., & Benini, L. (2023). Energy-efficient AI hardware: Opportunities and challenges. *IEEE Design & Test, 40*(2), 8–21. https://doi.org/10.1109/MDAT.001.2300092
6. Chen, Y., Peng, H., & Li, Q. (2023). Approximate computing for deep neural networks: An overview. *IEEE Transactions on Emerging Topics in Computing, 11*(1), 85–97. https://doi.org/10.1109/TETC.001.2200319
7. Zhang, H., Wang, P., & Liu, J. (2023). Near-memory computing architectures for energy-efficient AI inference. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 31*(5), 742–755. https://doi.org/10.1109/TVLSI.001.2200437
8. Liu, Z., Gupta, S., & Niemier, M. (2023). Reconfigurable VLSI architectures for edge AI applications. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42*(7), 2103–2115. https://doi.org/10.1109/TCAD.001.2200568
9. Mittal, S. (2023). A survey of techniques for dynamic voltage and frequency scaling in embedded systems. *ACM Computing Surveys, 55*(2), 1–36. https://doi.org/10.1145/3510429
10. Shi, W., Chen, T., & Li, H. (2023). ASIC accelerators for convolutional neural networks: A survey. *IEEE Transactions on Circuits and Systems I: Regular Papers, 70*(9), 3456–3470. https://doi.org/10.1109/TCSI.001.2300765
11. Surendar, A. (2025). Design and optimization of a compact UWB antenna for IoT applications. National Journal of RF Circuits and Wireless Systems, 2(1), 1–8.
12. Reginald, P. J. (2025). Wavelet-based denoising and classification of ECG signals using hybrid LSTM-CNN models. National Journal of Signal and Image Processing, 1(1), 9–17.
13. Sindhu, S. (2025). Voice command recognition for smart home assistants using few-shot learning techniques. National Journal of Speech and Audio Processing, 1(1), 22–29.
14. Uvarajan, K. P. (2025). Design of a hybrid renewable energy system for rural electrification using power electronics. National Journal of Electrical Electronics and Automation Technologies, 1(1), 24–32.
15. Kavitha, M. (2025). Hybrid AI-mathematical modeling approach for predictive maintenance in rotating machinery systems. Journal of Applied Mathematical Models in Engineering, 1(1), 1–8.