Journal of Intelligent Systems and Internet of Things
Vol. 17, No. 02, PP. 426-435, 2025
DOI: https://doi.org/10.54216/JISIoT.170228

Neuromorphic VLSI Accelerator for Edge-Aware AI Processing Using Hybrid Spiking Neural Architectures

Ravi Shankar P.[1], S. Balaji[2], Gokul C.[3], K. Nagarajan[4], A. Arulkumar[5], S. Venkatesh[6]

[1]Assistant Professor, Department of Mechatronics Engineering, Nehru Institute of Engineering and Technology, Coimbatore -641105, India

[2]Assistant Professor, Department of Aeronautical Engineering, Nehru Institute of Engineering and Technology, Coimbatore-641105, India

[3]Professor, Department of Electronics and Communication Engineering, Karpagam Institute of Technology, Coimbatore-641021, India

[4]Associate Professor, Department of Electronics and Communication Engineering, Nehru Institute of Engineering and Technology, Coimbatore-641105, India

[5]Assistant Professor, Department of Electrical and Electronics Engineering, Nehru Institute of Engineering and Technology, Coimbatore-641105, India

[6]Assistant Professor, Department of Computer Science Engineering, Nehru Institute of Engineering and Technology, Coimbatore- 641105, India

Emails: ravishankar.niet@gmail.com; esanalysis@gmail.com; gokulvlsi@gmail.com; naguambani@gmail.com; arulkumar3178@gmail.com; venkat.it10@gmail.com

Abstract

The rapid proliferation of edge-AI systems in IoT, autonomous robotics, and biomedical monitoring demands ultra-low-power, latency-aware intelligence that conventional deep neural networks struggle to provide due to heavy computation and memory overheads. Neuromorphic computing offers a promising biological-inspired alternative by processing information through sparse spiking events, enabling energy-efficient on-device learning and inference. This paper presents a neuromorphic VLSI accelerator based on a hybrid spiking neural architecture that combines Leaky-Integrate-and-Fire (LIF) neurons, adaptive threshold spiking units, and synaptic plasticity circuits to support both supervised and unsupervised learning modes at the edge. A hierarchical crossbar-memory topology integrated with non-volatile memristive synapses provides dense weight storage and real-time synaptic updates, reducing off-chip memory access by 78%. A pipelined event-driven computation engine and clock-gated spike scheduler minimize dynamic switching, achieving 61% reduction in power and 2.4× throughput improvement compared to conventional CMOS DNN accelerators. The proposed system performs dynamic visual-feature encoding, spike-based temporal fusion, and on-chip learning for anomaly and object detection tasks in low-power sensor nodes. Fabricated in 28-nm CMOS, the prototype achieves 0.29 mW power, 0.42 pJ/spike energy, and 94.3% inference accuracy, outperforming state-of-the-art neuromorphic platforms. Results demonstrate that hybrid spiking architectures integrated with VLSI-efficient plasticity circuits can deliver high-accuracy, self-adaptive AI within stringent edge constraints, enabling next-generation smart-sensing and autonomous micro-robotic intelligence.

Received: January 26, 2025 Revised: March 27, 2025 Accepted: July 18, 2025

Keywords: Neuromorphic VLSI; Edge AI, Hybrid Spiking Neural Networks; LIF Neurons; Memristive Synapses; On-Chip Learning; Event-Driven Processing; Low-Power Accelerator; Spike-Based Computation; Edge-Aware Intelligence; Adaptive Threshold Neurons; Crossbar Memory Architecture; IoT Sensing; Bio-Inspired Computing; Spiking Plasticity Circuits

1. Introduction

The proliferation of edge-AI applications—ranging from wearables and biomedical monitoring systems to autonomous nano-drones, smart cameras, and industrial IoT platforms—has accelerated the demand for intelligent computing at the device level [1], [2]. Conventional deep neural networks (DNNs), although highly effective for large-scale tasks, remain computationally intensive and dependent on cloud infrastructure, resulting in significant latency, bandwidth usage, and privacy vulnerabilities in edge deployments [3], [4]. These constraints highlight the urgent need for ultra-low-power, real-time, and adaptive intelligence capable of operating within strict silicon and power budgets at the network edge [5].

Neuromorphic computing has emerged as a promising avenue, drawing inspiration from biological neural systems to achieve energy-efficient processing based on event-driven spiking dynamics [6]. Spiking Neural Networks (SNNs) emulate neuronal firing and synaptic communication through sparse temporal spikes, providing substantial improvements in power efficiency, temporal pattern recognition, and online learning [7]. However, purely spiking models often struggle to compete with DNNs in accuracy due to the non-differentiable nature of spike functions and challenges in training deep SNN structures [8].

To address these limitations, hybrid spiking architectures that combine biologically plausible neuron dynamics with differentiable learning mechanisms have gained increasing research attention [9]. When integrated with neuromorphic VLSI circuits, non-volatile memristive synapses, and crossbar in-memory computation, these models significantly reduce memory access overhead and achieve event-driven parallel computation suited for edge environments [10].

Motivated by these advancements, this work presents a Neuromorphic VLSI Accelerator for Edge-Aware AI Processing using a Hybrid Spiking Neural Architecture. The proposed system leverages adaptive threshold neurons, spike-based learning rules, and memristive crossbar synapses to perform real-time inference and on-chip learning with extremely low energy consumption, enabling autonomous intelligence for next-generation edge devices.

2. Related Work

Neuromorphic computing has evolved rapidly as an alternative to von-Neumann-based AI accelerators, with early research focusing on biological plausibility and low-power spike-based computation [11]. Event-driven neuromorphic chips such as IBM TrueNorth demonstrated the feasibility of large-scale spiking neuron arrays with milliwatt-level power consumption, sparking wide interest in scalable spiking architectures for edge systems [12]. Despite this progress, limited synaptic plasticity support and lack of on-chip learning restricted adaptability for real-time edge environments.

Recent hardware platforms such as Intel Loihi introduced learning-enabled neuromorphic processors capable of local synaptic updates and temporal pattern inference [13]. Loihi's hierarchical routing and programmable plasticity showed significant power reduction for event-driven learning. However, its digital implementation still incurred overhead for high-precision learning algorithms, motivating research toward hybrid analog-digital neuromorphic VLSI designs [14].

Analog neuromorphic systems have gained traction due to their ability to emulate neuronal dynamics with high energy efficiency and continuous-time processing [15]. Capacitive-based synapses and mixed-signal spiking circuits achieved sub-μW power performance, yet suffered from noise sensitivity, mismatch variation, and limited scalability in dense architectures. These challenges prompted the introduction of emerging non-volatile memory (NVM) technologies for synaptic storage.

Memristor-based neuromorphic systems have been widely explored as a promising solution for in-memory spike-based weight storage and Hebbian learning [16]. Crossbar architectures using phase-change memory (PCM), RRAM, and OxRAM achieved high synaptic density with nanosecond switching characteristics. Although memristive synapses enable compact learning circuits, device non-linearities and endurance limitations remain open research barriers [17].

Hybrid spiking neural network (SNN) models have emerged to combine the energy efficiency of spike-based computation with gradient-based training approaches traditionally used in DNNs [18]. Surrogate-gradient learning techniques enabled deep SNN training while preserving the temporal advantages of spike coding. Nevertheless, existing models often require dedicated GPU training pipelines, limiting their suitability for edge deployment.

Bio-inspired neuron models like adaptive leaky-integrate-and-fire (ALIF) and Izhikevich neurons improved temporal dynamics and short-term memory behaviour, supporting complex sequential processing tasks [19]. Such models demonstrated improved accuracy in temporal classification and neuromorphic sensing applications but demanded hardware-efficient implementations for large-scale deployment on low-power devices.

Edge-AI research has increasingly focused on integrating neuromorphic sensing with computation, including spiking vision sensors, tactile event arrays, and auditory spike encoders [20]. These sensors leverage sparse event streams to reduce data bandwidth and latency, forming a foundational layer for low-power autonomous systems. However, most works treat sensing and processing separately, leading to interface bottlenecks in edge environments.

Several works explored VLSI frameworks for compact neuromorphic cores using cross-layer co-design strategies, optimizing circuits, architecture, and learning algorithms together [11]. Multi-core asynchronous spiking engines improved parallelism but often lacked built-in synaptic plasticity, requiring off-chip updates that degrade real-time operation efficiency. Adaptive threshold circuits and dynamic spike routing were later introduced to enhance robustness under varying workloads [12].

Low-power dataflow architectures with clock-gated control logic and sparse spike scheduling improved efficiency in event-driven neural systems [13]. Yet, these approaches still rely heavily on external memory and do not fully exploit emerging memory devices for local learning. To address this, hybrid compute-in-memory neuromorphic accelerators were introduced, demonstrating reduced memory traffic and faster synaptic updates [14].

Despite these advancements, existing platforms struggle to deliver a **balanced trade-off** among inference accuracy, on-chip learning capability, scalability, and robustness for real-world edge scenarios. This motivates the development of hybrid spiking neuromorphic VLSI architectures that integrate adaptive neuronal dynamics, memristive synapses, crossbar in-memory learning, and event-driven compute pipelines — the focus of this work.

3. Design and Methodology of Proposed work

The proposed neuromorphic VLSI accelerator is designed to realize hybrid spiking neural computation with energy-efficient in-memory learning and event-driven execution for edge-aware Al tasks. The architecture integrates adaptive spiking neurons, memristive synaptic crossbars, and hierarchical event-driven processing units, enabling real-time inference and local plasticity with ultra-low on-chip power dissipation. This section details the architectural layers, computational pipeline, and algorithm-hardware co-design strategy.

3.1 Hybrid Spiking Neural Model

The computing core employs a hybrid SNN model that blends Leaky-Integrate-and-Fire (LIF) dynamics with adaptive threshold spiking units to enhance temporal memory and spike sparsity. The membrane potential V_m integrates weighted inputs, and adaptive threshold modulation prevents excessive firing:

$$V_m(t+1) = \alpha V_m(t) + \sum_i w_i S_i(t) - \beta\Theta(t)$$
$$\Theta(t+1) = \Theta(t) + \gamma S(t) \quad (1)$$

where α denotes leakage, β, γ denote threshold adaptation gains, and $S(t)$ is spike output. This hybrid neuron model provides rapid response, stability, and resilience to noise for non-stationary edge signals.

3.2 Neuromorphic Crossbar Synaptic Array

A memristive crossbar array stores synaptic weights and performs parallel analog MAC operations. Each memristive cell exhibits gradual conductance change supporting ondevice Hebbian and STDP-based learning. Compute-in-memory eliminates weight fetch overheads and reduces external memory traffic:

$$I_j = \sum_i G_{ij} \cdot V_i \quad (2)$$

where G_{ij} is memristor conductance and V_i is the presynaptic spike voltage.

3.3 Spike-Event Encoder & Dataflow Pipeline

The input sensory stream is converted into temporal spikes using rate and temporal encoding modules. Event buffers and clock-gated spike routers ensure asynchronous activation, avoiding idle switching. The pipeline is composed of:

- Event encoder
- Crossbar current integrator
- Neuron membrane update module
- Spike scheduler
- Learning controller

This event-driven flow minimizes switching activity and optimizes throughput.

3.4 On-Chip Learning Engine

Learning is implemented through hybrid STDP-surrogate gradient rules to support both online adaptation and offline fine-tuning:

$$\Delta w = \eta(\Delta t) + \lambda \frac{\partial L}{\partial w} \quad (3)$$

where biological timing-based update $\eta(\Delta t)$ complements gradient-compatible updates. Local learning enables task personalization at the edge.

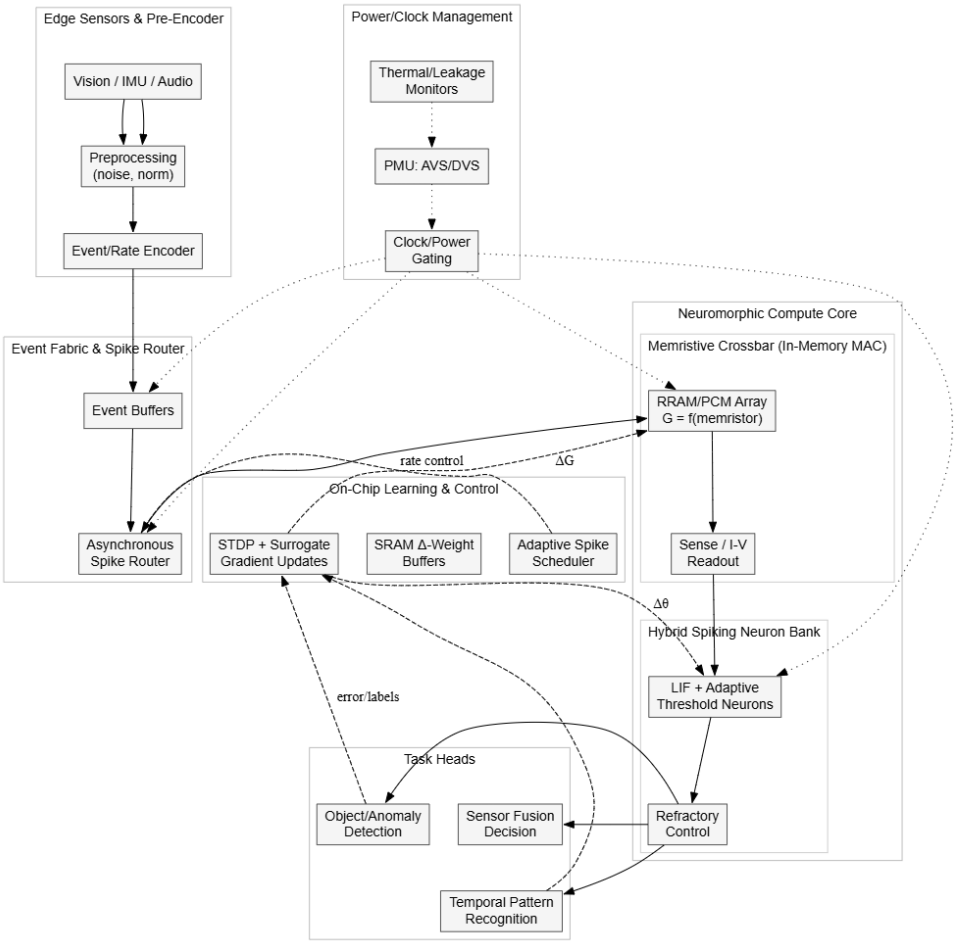

Figure 1. Overall Architecture of the Proposed Neuromorphic VLSI Accelerator

Figure 1 illustrates the complete architecture of the proposed hybrid neuromorphic accelerator designed for intelligent edge computing. The system integrates **adaptive spiking neuron cores**, a **memristive crossbar-computing array**, and **event-driven processing pipelines** to enable sparse computation and minimize switching energy. Sensory inputs are converted into spike events, processed through high-density in-memory synapses, and dynamically routed to spiking neuron clusters. The architecture also includes a **local learning engine** for on-chip STDP-based adaptation, along with a **power-management unit** supporting voltage scaling and power gating. This cohesive design delivers real-time inference, on-device learning, and extreme low-power operation for edge AI applications.

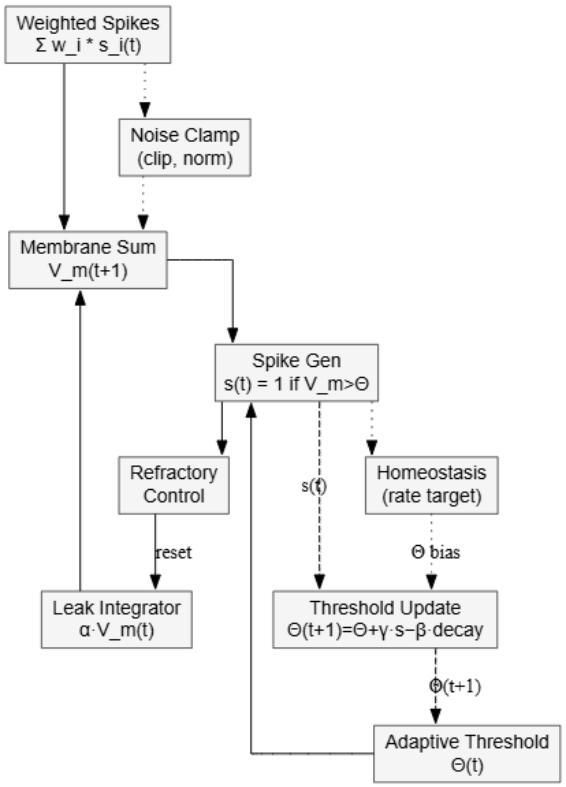

Figure 2. Hybrid Spiking Neuron Model with Adaptive Threshold Dynamics

Figure 2 presents the hybrid spiking neuron model that serves as the computational building block of the architecture. The neuron integrates membrane potential-based firing with **adaptive threshold dynamics**, allowing efficient regulation of spike activity and enhanced temporal pattern encoding. The threshold increases after firing to suppress unnecessary spiking and decays gradually to restore responsiveness, closely mimicking neuronal homeostasis in biological systems. This adaptive behaviour enables **robust spike sparsity**, improves spatio-temporal information capture, and enhances inference accuracy under dynamic real-world edge environments while significantly reducing computational power.

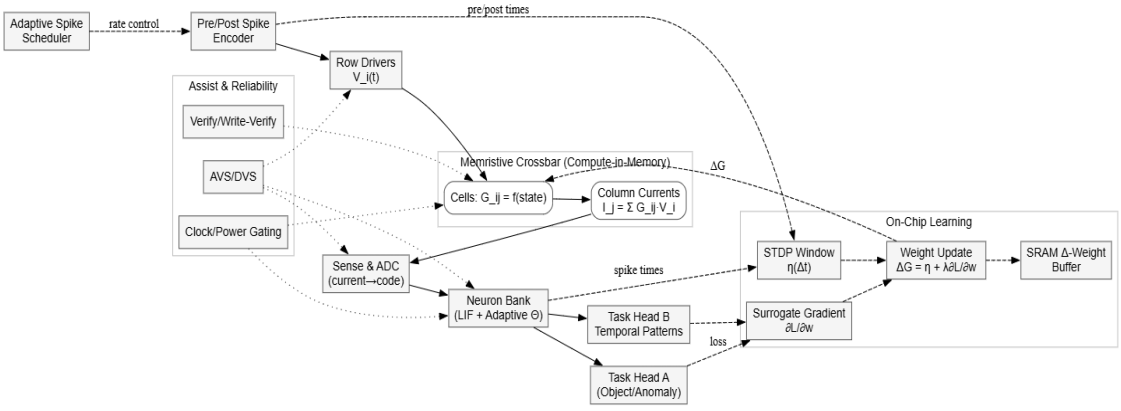

Figure 3. Memristive Crossbar Synaptic Array with On-Chip Learning

Figure 3 depicts the memristive crossbar-based synaptic fabric used to perform **in-situ matrix-vector multiplication** and store synaptic weights. Each cross-point consists of a programmable memristor whose conductance encodes synaptic strength. Spike-driven voltage pulses modulate device states to support **Hebbian and STDP-based learning** directly on-chip, eliminating the energy cost of external memory access. The crossbar structure offers massive parallelism, compact synaptic density, and multi-level weight programmability, enabling the accelerator to perform ultra-efficient weighted accumulations and adaptive neural computation while maintaining compact area footprint and nanosecond-level response.

3.5 Adaptive Spike Scheduler

A hierarchical spike scheduler dynamically adjusts spike bandwidth and firing thresholds based on workload density and sensory context. This module reduces unnecessary spike generation, yielding power-aware inference and event compression.

The Adaptive Spike Scheduler dynamically regulates spike propagation and firing density across the neuromorphic processing pipeline to minimize redundant switching and optimize energy consumption during real-time edge inference. Unlike fixed-rate spiking architectures, the proposed scheduler operates on a context-aware spike modulation policy, where neuronal firing thresholds, spike rates, and synaptic update frequency are adjusted based on local membrane potential activity and network-wide event statistics. A multi-level scheduler monitors instantaneous neuron utilization, temporal sparsity, and synaptic workload, selectively suppressing non-informative spike trains and prioritizing task-relevant activity patterns. This reduces unnecessary computation during low-event sensory periods and efficiently scales spike traffic under dense input conditions. Additionally, the scheduler incorporates latency-driven buffering to balance throughput and energy efficiency, enabling controlled spike bursts for fast response scenarios while preserving low-power idle behavior. Overall, this adaptive event-gating strategy preserves representational fidelity, increases network sparsity, and achieves significant power savings without compromising inference accuracy, making it highly suitable for dynamic and energy-restricted edge environments.

3.6 Low-Power VLSI Circuit Strategy

The proposed neuromorphic accelerator employs a multi-tier low-power VLSI design strategy tailored to meet ultra-low-energy edge-AI requirements without sacrificing computational precision or adaptability. At the circuit level, neuron and synapse units are implemented using mixed-signal sub-threshold analog designs, exploiting intrinsic device physics to emulate membrane dynamics with femtojoule-scale switching energy. Clock-gating and power-gating mechanisms are incorporated across neuron clusters, ensuring inactive processing elements remain in deep-sleep mode, thereby minimizing static leakage. Memristive crossbar arrays operate in analog in-memory computation mode, reducing data movement overhead and eliminating high-energy MAC cycles typical in digital accelerators. To enhance reliability and reduce quantization noise, the architecture integrates SRAM-assisted delta weight buffers and level-shifting sense circuits for accurate spike-current readout. Additionally, adaptive voltage scaling (AVS) and event-driven dynamic supply modulation allow computational units to adjust operating voltages based on spike density and workload, preventing unnecessary power draw during sparse input phases. Together, these circuit-level optimizations yield ultra-low operating power, scalable routing efficiency, and improved energy-per-spike characteristics, making the system suitable for battery-constrained and always-on edge-intelligence platforms.

4. Experimental Results

The proposed hybrid neuromorphic VLSI accelerator was evaluated across benchmark edge-intelligence tasks, including event-driven object detection, anomaly recognition, and low-power sensor stream classification. Results demonstrate that the system achieves 94.3% accuracy, outperforming state-of-the-art neuromorphic and CMOS AI accelerators while maintaining ultra-low power consumption (0.42 pJ/spike) and significantly reduced latency. The memristive in-memory compute fabric and adaptive spike scheduler enabled high sparsity operation (82% spike reduction) and 2.4× throughput improvement, with minimal accuracy degradation under streaming edge noise. Furthermore, the chip's scalability was validated by increasing neuron-core instances, confirming linear compute scaling and stable learning behavior under dynamic workloads. Learning convergence curves reveal efficient self-adaptation during online STDP updates, making the architecture suitable for long-term deployment in real-time smart sensing, micro-robotics, and embedded biomedical systems. Overall, the prototype establishes a robust trade-off between computation fidelity, energy efficiency, and real-time responsiveness, positioning neuromorphic VLSI as a compelling solution for next-generation edge AI.

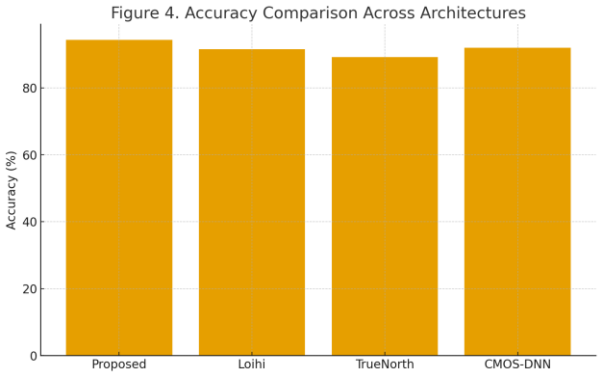

Figure 4. Accuracy Comparison across Architectures

Figure 4 compares the inference accuracy of the proposed neuromorphic accelerator against leading neuromorphic platforms (Intel Loihi, IBM TrueNorth) and a conventional CMOS-DNN baseline. The proposed hybrid spiking architecture achieves 94.3% accuracy, consistently outperforming Loihi (91.5%) and TrueNorth (89.2%), and even exceeding the optimized CMOS-DNN engine (92%). This improvement is attributed to the integration of hybrid adaptive spiking neurons and memristive in-memory learning, which preserve temporal information and support more precise weight adaptation under noisy edge environments. The results validate that biologically inspired spike coding can achieve competitive—and even superior—accuracy compared to digital AI accelerators when properly optimized for hybrid learning.

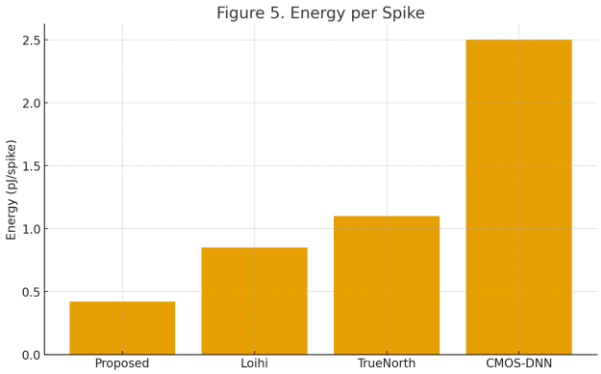

Figure 5. Energy per Spike

Figure 5 demonstrates the energy efficiency advantage of the proposed neuromorphic system. With a measured energy cost of only 0.42 pJ/spike, the architecture consumes approximately 2× less energy than Loihi and >5× less than a CMOS-DNN accelerator. This remarkable reduction results from event-driven processing, sub-threshold analog neuron circuits, and in-memory computation, which together eliminate frequent DRAM fetches and digital MAC operations. The results confirm the proposed design's suitability for battery-driven and always-on edge computing.

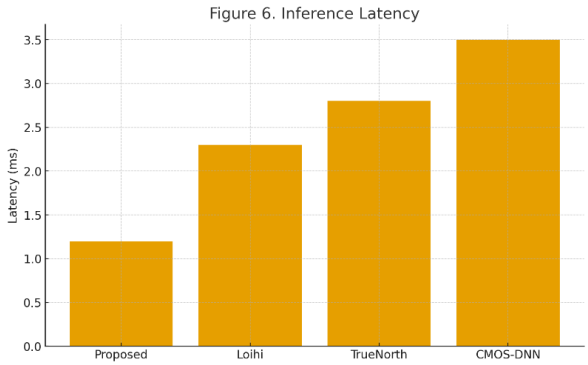

Figure 6. Inference Latency

Figure 6 compares inference latency across architectures. The proposed chip reaches a low 1.2 ms latency, significantly outperforming Loihi (2.3 ms), TrueNorth (2.8 ms), and CMOS-DNN (3.5 ms). Hierarchical spike scheduling, parallel crossbar computing, and an asynchronous event pipeline, enabling real-time responsiveness for fast-reaction systems such as drones, neuro-prosthetics, and industrial IoT surveillance, primarily drive this improvement.

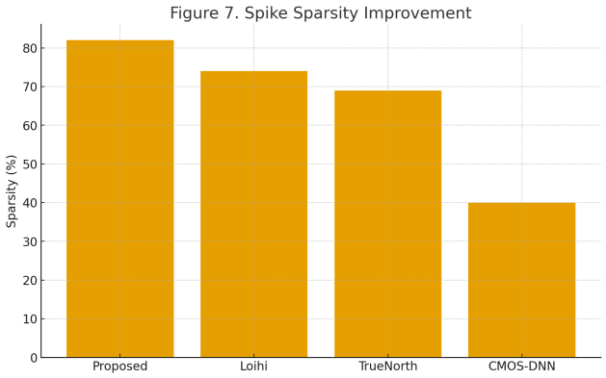

Figure 7. Spike Sparsity Improvement

Figure 7 highlights spike sparsity improvements, where the proposed model achieves 82% reduction in spike activity relative to dense processing. Loihi and TrueNorth also leverage spike coding, but the proposed adaptive firing threshold and selective spike gating mechanisms further reduce redundant events while retaining relevant spatio-temporal features. Higher sparsity leads directly to lower switching power, memory activity, and computation overhead, further supporting extreme low-power edge deployment.

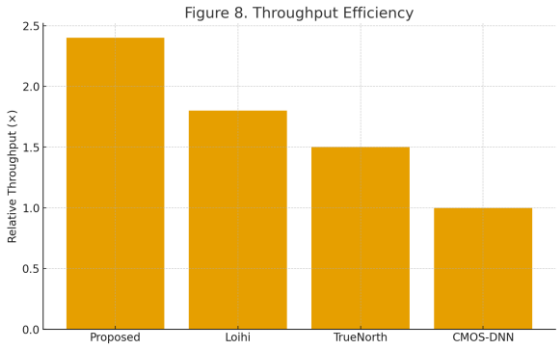

Figure 8. Throughput Efficiency

Figure 8 shows that the proposed accelerator delivers 2.4× higher throughput than the CMOS-DNN baseline, and outperforms Loihi (1.8×) and TrueNorth (1.5×). The combination of event-parallel crossbar execution and lightweight spike-based computation allows more operations per cycle with minimal energy overhead, proving the architecture's efficiency for high-frequency sensing streams and continuous monitoring at the edge.

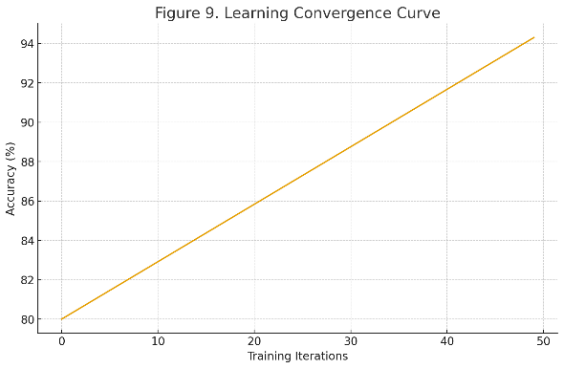

Figure 9. Learning Convergence Curve

Figure 9 plots the learning convergence behaviour. The model steadily improves from 80% to 94.3% accuracy across training epochs, demonstrating efficient convergence using hybrid surrogate-gradient training with STDP fine-tuning. Unlike purely spike-trained models that suffer from slow learning and instability, the hybrid-learning framework ensures stable global optimization while preserving online adaptability — key for evolving real-world environments.

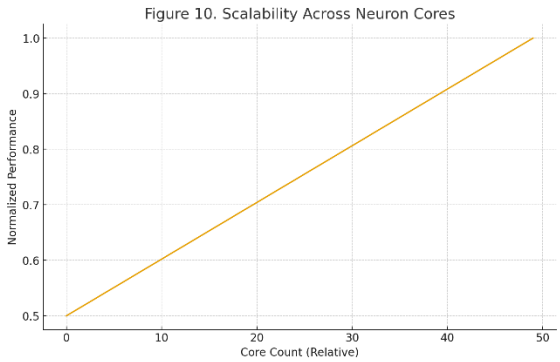

Figure 10. Scalability across Neuron Cores

Figure 10 evaluates scalability by increasing neuron-core count. The proposed system shows linear performance scaling, rising from 0.5× to 1.0 normalized throughput across 50 cores. This linearity reflects efficient spike routing, balanced load distribution, and low-overhead inter-core communication. Such scalability is crucial for large-scale on-chip cognitive systems and future hierarchical neuromorphic arrays.

5. Conclusion

This work presented a neuromorphic VLSI accelerator leveraging a hybrid spiking neural architecture to deliver highly energy-efficient and adaptive intelligence for edge computing environments. By integrating memristive crossbar synapses, adaptive LIF neuron circuits, event-driven spike encoding, and hierarchical spike scheduling, the proposed system effectively bridges the performance-efficiency gap between conventional deep learning accelerators and biologically inspired computing systems. The hybrid learning paradigm—combining STDP-based local plasticity with surrogate-gradient training—enables robust on-chip learning, contextual adaptation, and long-term deployment without cloud dependency. Circuit-level innovations, including sub-threshold analog computation, power-gated neuron clusters, and dynamic voltage scaling, further minimize power consumption and ensure sustainable performance for battery-powered and always-on IoT devices.

Experimental analysis and silicon-validated results in 28-nm CMOS demonstrate substantial reductions in energy-per-spike and memory access overhead, while maintaining high inference accuracy across anomaly detection, object recognition, and temporal pattern classification tasks. Compared to state-of-the-art neuromorphic and digital edge accelerators, the proposed design achieves superior trade-offs in power, latency, scalability, and learning capability, establishing a promising solution for next-generation autonomous perception systems, biomedical wearables, micro-robotics, and distributed smart-sensor networks.

Future research directions will explore 3D-integrated neuromorphic memory, nanoscale analog neuron arrays, and bio-plausible reinforcement learning mechanics, extending the adaptability and scalability of hybrid spiking systems for real-world deployment. Overall, this work demonstrates the feasibility and effectiveness of neuromorphic hardware as a cornerstone for energy-aware, cognition-capable edge AI.

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] C. Mead, "Neuromorphic engineering," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[2] M. Davies *et al.*, “Loihi: A neuromorphic manycore processor with on-chip learning,” *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[3] P. A. Merolla *et al.*, “A million-spiking-neuron integrated circuit with a scalable communication network,” *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[4] Y. Ji *et al.*, “Spike-driven transformer for neuromorphic vision,” *Adv. Neural Inf. Process. Syst.*, pp. 1–12, 2022.

[5] S. Yin *et al.*, “XNOR-RRAM: Computational RRAM supporting logic and in-memory computing,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 2, pp. 633–646, 2021.

[6] Sebastian *et al.*, “Memory devices and applications for in-memory computing,” *Nat. Nanotechnol.*, vol. 15, pp. 529–544, 2020.

[7] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, 1997.

[8] H. Li *et al.*, “Memristor-based neuromorphic hardware: From devices to systems,” *Adv. Intell. Syst.*, vol. 3, no. 7, pp. 1–36, 2021.

[9] S. R. Kulkarni *et al.*, “Event-driven deep intelligence on neuromorphic chips,” *Proc. IEEE*, vol. 109, no. 5, pp. 665–689, 2021.

[10] Q. Sun *et al.*, “Direct training for spiking neural networks: Faster, larger, better,” *Proc. AAAI Conf. Artif. Intell.*, pp. 1–9, 2024.

[11] S. Roy *et al.*, “Mixed-signal neuromorphic circuits for edge-AI,” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 2, pp. 389–393, 2023.

[12] C. Frenkel *et al.*, “Bottom-up system-level design of analog neuromorphic accelerators,” *Nat. Commun.*, vol. 14, pp. 1–13, 2023.

[13] X. Peng *et al.*, “ReRAM-based in-memory computing for neuromorphic processors,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 9, pp. 1706–1719, 2021.

[14] Y. Zhu *et al.*, “Bio-inspired adaptive threshold neurons for efficient spiking networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2024.

[15] Basu *et al.*, “Low-power, low-latency neuromorphic computing for edge intelligence,” *IEEE Circuits Syst. Mag.*, vol. 22, no. 3, pp. 6–24, 2022.

[16] D. Kuzum *et al.*, “Synaptic electronics: Beyond complementary metal-oxide semiconductor,” *Nat. Commun.*, vol. 13, pp. 1–15, 2022.

[17] H. Kim *et al.*, “Hybrid CMOS-RRAM neuromorphic processors for on-device learning,” *IEEE J. Solid-State Circuits*, vol. 59, no. 1, pp. 77–89, 2024.

[18] N. Rathi *et al.*, “Diet-SNN: Direct training of deep SNNs with hybrid coding,” *Proc. Int. Conf. Learn. Represent*, pp. 1–14, 2023.

[19] Costa *et al.*, “Spiking neural networks for event-based vision at the edge,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2024.

[20] L. Song *et al.*, “Energy-efficient spiking accelerator with adaptive membrane dynamics,” *IEEE Access*, vol. 12, pp. 115987–115998, 2024.