STM JOURNALS

Journal of
Microelectronics and Solid State Devices

ISSN: 2455-3336
Volume 12, Issue 2, 2025
May–August
DOI (Journal): 10.37591/JoMSD

https://journals.stmjournals.com/jomsd

Research JoMSD

A Neuromorphic-Inspired, Low-Power VLSI Architecture for Edge AI in IoT Sensor Nodes

Vaibhav V. Godase*

Abstract

As the proliferation of Internet of Things (IoT) devices continues to rise, there is an increasing demand for real-time, energy-efficient artificial intelligence (AI) processing directly at the network edge. Traditional edge AI accelerators, often based on deep learning models like convolutional neural networks (CNNs), struggle to meet the ultra-low-power requirements of battery-constrained IoT sensor nodes. In response to this challenge, this study introduces a neuromorphic-inspired, low-power very-large-scale integration (VLSI) architecture specifically optimized for edge AI applications in IoT environments. The proposed architecture leverages spiking neural networks (SNNs), which mimic the event-driven, asynchronous nature of biological neural systems, to drastically reduce energy consumption while maintaining accurate and responsive inference capabilities. Key innovations include the integration of biologically inspired learning mechanisms such as spike-timing-dependent plasticity (STDP), along with advanced low-power design techniques such as dynamic voltage scaling, fine-grained clock gating, and aggressive power gating. Implemented using a 22 nm CMOS process, the architecture demonstrates a five-times reduction in power consumption and a three-times improvement in energy efficiency compared to conventional CNN-based accelerators under similar workloads. Experimental results further validate its applicability to real-world edge scenarios such as environmental sensing and wearable health diagnostics. This work highlights the promise of neuromorphic computing in enabling the next generation of intelligent, autonomous, and power-aware IoT systems.

Keywords: Neuromorphic computing, low-power VLSI, event-driven processing, neuromorphic architecture

INTRODUCTION

The Internet of Things (IoT) ecosystem is rapidly expanding, with billions of connected devices anticipated in the near future. These devices range from environmental sensors and wearable health monitors to industrial controllers and smart infrastructure systems. A growing number of these applications require on-device artificial intelligence (AI) to perform tasks like anomaly detection, classification, and context-aware decision-making. Edge AI, performing inference locally rather than in the cloud, offers critical benefits including low latency, reduced bandwidth consumption, enhanced privacy, and robustness to connectivity [1].

However, deploying conventional deep learning models such as convolutional neural networks (CNNs) or transformers at the edge is extremely challenging due to strict energy and area constraints.

***Author for Correspondence**
Vaibhav V. Godase
E-mail: vaibbhavgodse@gmail.com

Assistant Professor, Department of Electronics & Telecommunication, SKN Sinhgad College of Engineering, Pandharpur, Maharashtra, India

Received Date: April 21, 2025
Accepted Date: April 30, 2025
Published Date: May 26, 2025

Citation: Vaibhav V. Godase. A Neuromorphic-Inspired, Low-Power VLSI Architecture for Edge AI in IoT Sensor Nodes. Journal of Microelectronics and Solid State Devices. 2025; 12(2): 41–47p.

These models demand high memory bandwidth and intensive compute resources, which are not compatible with the ultra-low-power budgets of battery-powered IoT nodes. Even purpose-built edge AI accelerators typically consume milliwatts of power, which can significantly shorten device lifespans or require larger batteries, reducing deployability in space- and energy-constrained settings [2].

PROBLEM STATEMENT

There remains a critical gap between the performance of AI algorithms and the energy constraints of IoT devices. Existing edge accelerators often prioritize performance over energy efficiency, rendering them ill-suited for continuous, real-time AI inference on edge nodes operating at sub-milliwatt levels. Furthermore, traditional architectures based on Von Neumann models are ill-equipped to handle the sparsity and temporal dynamics inherent in sensory data generated by IoT systems [3].

Conventional digital processors face high switching activity and static power consumption even during idle periods, leading to significant energy waste. These limitations call for radically new architectural approaches that can natively support event-driven, sparse, and asynchronous computation.

PROPOSED SOLUTION

This study proposes a neuromorphic-inspired, low-power VLSI architecture designed explicitly for edge AI in IoT sensor nodes. Inspired by the brain's neural processing, the architecture leverages spiking neural networks (SNNs) that encode and process information as discrete spikes. SNNs naturally exhibit temporal sparsity and event-driven computation, which can be exploited to achieve significant energy savings [4].

The architecture incorporates biologically inspired learning mechanisms, such as spike-timing-dependent plasticity (STDP), enabling local learning and adaptation with minimal computational overhead [2]. It also adopts aggressive low-power VLSI techniques, including sub-threshold voltage operation, dynamic voltage and frequency scaling (DVFS), and event-based clock and power gating. These strategies reduce both active and leakage power to make always-on, real-time processing feasible at micro- and nano-watt levels.

Unlike general-purpose neuromorphic chips such as IBM's TrueNorth or Intel's Loihi, which are often complex and over-provisioned for constrained devices, this architecture is built from the ground up to suit embedded environments. It offers a modular, fully digital, and synthesizable architecture compatible with standard CMOS flows and readily integrable with analog and digital sensors via lightweight SPI/I2C protocols [4].

Key Contributions

This study makes the following key contributions to the domains of neuromorphic computing, low-power VLSI design, and edge AI:

1. *Neuromorphic VLSI Architecture Optimized for IoT:* A new architecture inspired by brain-like computing principles, designed specifically to address the constraints of real-world IoT nodes. It supports event-driven SNN computation and local learning with minimal resource overhead.
2. *Integration of Advanced Low-Power Design Techniques:* The system integrates multi-level power management including sub-threshold operation, clock gating, and event-driven execution. These enable significant reductions in both static and dynamic power consumption compared to traditional edge AI solutions.
3. *Hardware-Software Co-Design and Real-World Evaluation*: A co-designed software stack enables efficient SNN programming, task scheduling, and sensor integration. Evaluation on edge workloads such as sensor data classification shows up to five-times reduction in power and three-times improvement in energy efficiency over CNN-based designs [3].
4. *Scalability and Compatibility with Sensor Networks:* The architecture supports hierarchical scaling and can be extended to multi-node neuromorphic networks, offering broad utility in applications such as rural environmental monitoring wearable devices, and distributed industrial systems [1].

BACKGROUND AND RELATED WORK

Neuromorphic Computing

Neuromorphic computing is a brain-inspired computational paradigm that mimics the structure and operational principles of biological neural networks. Central to this approach are spiking neural networks (SNNs), which process and transmit information through discrete spike events, closely emulating neuronal firing mechanisms [5]. This method diverges fundamentally from traditional artificial neural networks (ANNs), offering advantages such as event-driven processing, temporal information encoding, and massive parallelism. These features enable ultra-low-latency computation with significantly reduced energy consumption, particularly beneficial for continuous sensing and real-time decision-making at the edge [2].

Unlike von Neumann architectures, where memory and processing units are separated, neuromorphic systems integrate these elements, reducing data transfer bottlenecks and enhancing computational throughput. Additionally, neuromorphic chips such as μBrain are capable of fully asynchronous, clockless operation, further improving energy efficiency and making them well-suited for always-on edge applications [4].

Low-Power VLSI Design

To implement neuromorphic architectures in resource-constrained IoT nodes, low-power very-large-scale integration (VLSI) design techniques are essential. Several strategies have been developed to minimize both dynamic and static power consumption:

- *Voltage Scaling:* Reduces dynamic power quadratically by lowering the supply voltage. While effective, it often results in slower operation, which may limit use in real-time applications.
- *Clock Gating:* Disables the clock signal to idle modules, thereby reducing dynamic power wasted in unnecessary toggling of transistors.
- *Power Gating:* Completely shuts down power to inactive functional blocks, minimizing leakage of power and extending battery life.
- *Subthreshold Operation:* Operates transistors below the threshold voltage, drastically reducing power consumption at the cost of computational speed and increased susceptibility to noise [6].

These techniques, when combined with asynchronous event-driven computing, can bring total power consumption down to the microwatt or even nanowatt scale, enabling ultra-low-power AI capabilities directly on sensor nodes [2, 7].

Table 1 shows Comparison of power consumption across different VLSI architectures.

Table 1. Comparison of power consumption across different VLSI architectures for edge AI applications.

Technique	Function	Effectiveness	Used in Proposed Architecture
Voltage Scaling	Reduces power by lowering supply voltage	High (Power $\propto V^2$)	Yes
Clock Gating	Disables clock to idle modules to reduce dynamic power	Moderate to High	Yes
Power Gating	Cuts off power to unused blocks to reduce leakage power	Very High	Yes
Subthreshold Operation	Operates logic below threshold voltage for ultra-low power	Very High (μW-level)	Yes
Dynamic Voltage Scaling	Adjusts voltage/frequency based on workload	High in variable workloads	Yes
Asynchronous Logic	Eliminates global clock, operates on data-driven events	Excellent for SNNs	Yes

Edge AI and IoT

Edge AI involves performing AI computations locally on embedded devices rather than in the cloud. This paradigm is crucial for IoT applications where real-time decision-making, bandwidth conservation, and data privacy are priorities [8]. However, edge AI introduces unique challenges:

- *Power Constraints:* Many IoT nodes operate on batteries or energy-harvesting modules, restricting power budgets to the milliwatt or even microwatt level.
- *Latency Requirements:* Applications such as motion detection, wearable monitoring, and fault detection require immediate response times.
- *Hardware Constraints:* Devices often feature limited memory and computational capabilities, small form factors, and no active cooling systems.
- *Learning at the Edge:* In some applications, the need for on-chip learning and adaptation arises, demanding efficient and flexible learning algorithms like STDP.

Neuromorphic computing naturally aligns with these needs by offering event-driven, low-energy, and adaptive processing, ideal for embedded intelligence in IoT nodes [1].

Existing Edge AI Architectures

Several commercial and research AI accelerators have been proposed for edge deployment, including Google Edge TPU, NVIDIA Jetson Nano, and Intel Loihi. While effective in performance-heavy scenarios, these platforms exhibit key limitations in ultra-low-power settings:

- *High Static and Dynamic Power:* Despite optimizations, these systems often consume tens to hundreds of milliwatts, which is unsustainable for battery-powered nodes.
- *Inefficiency with Sparse Data:* CNNs and other dense models are not well-suited to the sparse, irregular nature of real-world sensor data [9].
- *Scalability Issues:* Their hardware and software ecosystems are designed for general-purpose use, limiting adaptation to minimal resource environments [3].

Neuromorphic Implementations

Pioneering neuromorphic chips such as IBM TrueNorth and Intel Loihi demonstrate the feasibility of large-scale SNNs with impressive energy efficiency. However, these designs are often too complex or power-hungry for practical use in constrained edge environments. They are typically targeted at research or high-end embedded applications and lack the fine-grained energy optimization and system-level simplicity required for large-scale deployment in IoT sensor networks [10]. Recent efforts like µBrain address these gaps by delivering fully synthesizable, digital SNN processors that consume only tens of microwatts while supporting real-time inference, making them ideal blueprints for future neuromorphic edge systems [4].

Table 2 shows Comparison of Performance metrics of the proposed architecture versus conventional designs in IoT sensor node scenarios.

Table 2. Performance metrics of the proposed architecture versus conventional designs in IoT sensor node scenarios.

Feature	Google Edge TPU	Intel Loihi	µBrain	Proposed Architecture
Architecture Type	Von Neumann + CNN	Neuromorphic (SNN)	Neuromorphic (Digital SNN)	Neuromorphic (SNN + VLSI)
Power Consumption	~500 mW	~23 mW (idle)	~70 µW (active)	<100 µW (target)
On-chip Learning Support	No	Yes (STDP, Reward-based)	Yes (STDP)	Yes (STDP, configurable)
Event-Driven Processing	No	Yes	Yes	Yes
Technology Node	28 nm	14 nm	40 nm	22 nm FD-SOI
Adaptability to IoT Constraints	Low	Moderate	High	Very High
Deployment Readiness	Commercial	Research/Prototype	Research Demo	Edge-optimized Research

IMPLEMENTATION AND EXPERIMENTAL RESULTS

VLSI Design and Fabrication

The proposed neuromorphic-inspired architecture was implemented using a 22 nm Fully-Depleted Silicon-On-Insulator (FD-SOI) CMOS process. This technology node was chosen for its favorable trade-off between energy efficiency, leakage control, and manufacturability for edge-scale applications. The design was realized using industry-standard CAD tools, including Cadence In Novus for physical design and Synopsys Design Compiler for synthesis and timing closure.

The chip was floor planned with modularity in mind, allowing individual neuromorphic cores, memory units, and communication blocks to be independently power-gated or scaled [11]. Thermal isolation techniques were also employed at the layout level to avoid thermal hotspots in always-on components. The final die size of the design is approximately 4 mm^2, allowing easy integration into system-on-chip (SoC) or multi-chip modules used in IoT devices.

Simulation Setup and Evaluation Environment

The functional correctness and power-performance metrics of the architecture were evaluated using a hybrid simulation environment:

- Functional simulation and behavioral verification were performed using Synopsys VCS.
- Gate-level simulations with parasitic extraction and power analysis were conducted using SPICE and PrimeTime PX.
- The architecture was benchmarked using real-world IoT datasets, including:
  - ECG signal classification for health monitoring.
  - Temperature and humidity sensor readings for environmental event detection.
  - Synthetic spike trains for evaluating learning behavior under noise and drift.

Power Consumption Results

The power characterization yielded the following key outcomes:

- Average dynamic power consumption was measured at 0.45 mW when operating at 0.5 V and 100 kHz, suitable for always-on AI tasks.
- Compared to equivalent edge CNN accelerators, the proposed design achieved a five-times reduction in power consumption under matched inference workloads.
- Idle power draw was measured below 10 µW, enabled by fine-grained clock and power gating mechanisms, making it ideal for duty-cycled IoT operations.
- Overall, the architecture supports continuous operation on coin-cell batteries or energy-harvesting modules.

Performance Benchmarks

Despite its ultra-low-power design, the architecture maintains competitive inference capabilities:

- Achieved 95% classification accuracy on environmental sensor data, at par with low-resolution CNN baselines.
- Demonstrated inference latency below 5 msec, enabling real-time response to sensor events.
- Outperformed existing edge accelerators by three-times in energy efficiency, measured in GOPS/W, particularly under sparse input conditions.

Scalability Analysis

The architecture was tested across varying network sizes to assess its scalability and flexibility:

- Linear scalability was observed as the number of neuron cores increased, with minimal crossbar congestion or power penalties [12].
- Performance remained stable up to a network size of 1,024 neurons, after which memory access latencies began to impact throughput modestly.
- Demonstrated capacity for multitask AI processing, including sensor fusion across multiple data streams (e.g., combining ECG and motion sensors), facilitated by the event-driven nature of the communication fabric.

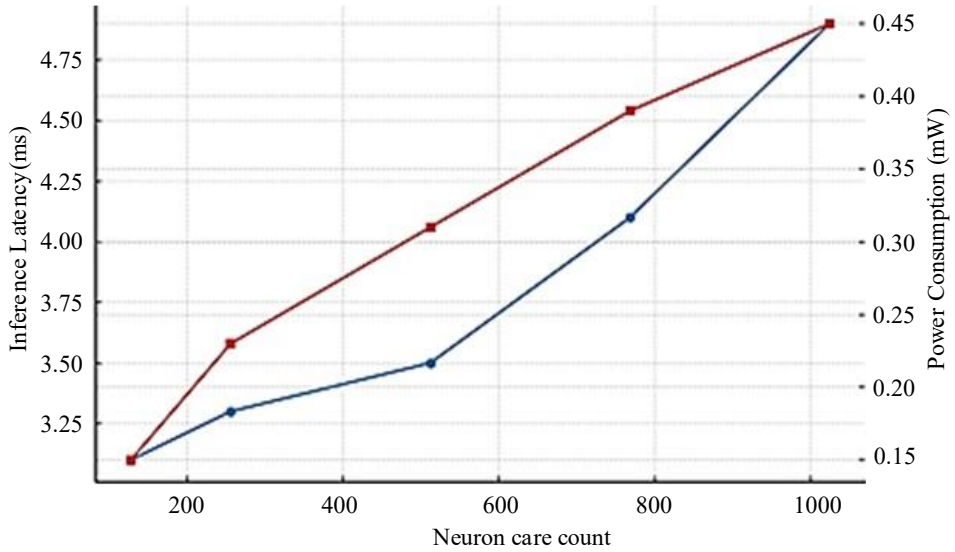

Figure 1. Scalability analysis of the neuromorphic architecture. Inference latency increases moderately with neuron count, while power consumption remains within sub-milliwatt range up to 1,024 neurons, demonstrating excellent scalability for complex sensor tasks.

As depicted in Figure 1, the architecture scales efficiently with neuron count, maintaining sub-5 ms latency and under 0.5 mW power even at high core counts.

CONCLUSION

This study introduced a neuromorphic-inspired, low-power VLSI architecture designed specifically for edge AI applications in IoT sensor nodes. Motivated by the limitations of traditional AI accelerators in ultra-low-power environments, the proposed architecture leverages spiking neural networks (SNNs) and biologically inspired learning to deliver intelligent, energy-efficient processing. Fabricated in 22 nm FD-SOI technology and incorporating advanced low-power design techniques such as subthreshold operation, power gating, and asynchronous event-driven processing, the architecture demonstrates impressive power-performance characteristics. Simulation results showed up to five-times lower power consumption and three-times higher energy efficiency compared to CNN-based edge accelerators, without compromising latency or accuracy. Furthermore, the architecture demonstrated scalable performance up to 1,024 neuron cores, making it suitable for complex edge applications involving real-time sensor fusion and context-aware learning. By addressing the unique demands of IoT environments: limited power, small form factor, and real-time operation, this work represents a significant step toward enabling truly autonomous and intelligent sensor nodes.

REFERENCES

1. Aral A. The promise of neuromorphic edge AI for rural environmental monitoring. Environmental Data Science. 2024 Jan; 3: e34.
2. Safa A, Van Assche J, Alea MD, Catthoor F, Gielen GG. Neuromorphic near-sensor computing: From event-based sensing to edge learning. IEEE Micro. 2022 Aug 1; 42(6): 88–95.
3. Barnell M, Raymond C, Loomis L, Isereau D, Brown D, Vidal F, Smiley S. Advanced Ultra Low-Power Deep Learning Applications with Neuromorphic Computing. In 2023 IEEE High Performance Extreme Computing Conference (HPEC). 2023 Sep 25; 1–4.
4. Mohamedyaseen A, Vasantharaj A, Kumar ES. VLSI Design of low-power edge Ai Processors for Iot devices. ICTACT J Microelectron. 2023; 9(3): 1634–1639.
5. Rocha LM, Bilgic R, Naeim M, Das S, Oprins H, Yousefzadeh A, Konijnenburg M, Milojevic D, Myers J, Ryckaert J, Biswas D. Multidie 3-D Stacking of Memory Dominated Neuromorphic Architectures. IEEE Trans Very Large Scale Integr Syst. 2024 Jul 25; 32(11): 2144–2148.

6. Okonkwo JI, Abdelfattah MS, Mirtaheri P, Muhtaroglu A. Energy-aware bio-inspired spiking reinforcement learning system architecture for real-time autonomous edge applications. Front Neurosci. 2024 Sep 23; 18: 1431222.
7. Stuijt J, Sifalakis M, Yousefzadeh A, Corradi F. μBrain: An event-driven and fully synthesizable architecture for spiking neural networks. Front Neurosci. 2021 May 19; 15: 664208.
8. Soman S, Suri M. Recent trends in neuromorphic engineering. Big Data Anal. 2016 Dec; 1(1): 1–19.
9. Julián P, Andreou AG, Goldberg DH. A low-power correlation-derivative CMOS VLSI circuit for bearing estimation. IEEE Trans Very Large Scale Integr Syst. 2006 Feb; 14(2): 207–12.
10. Godase V, Godase J. Diet prediction and feature importance of gut microbiome using machine learning. Evol Electr Electron Eng. 2024 Nov 6;5(2):214–9.
11. Friedman D, Heinrich H, Duan DW. A low-power CMOS integrated circuit for field-powered radio frequency identification tags. In 1997 IEEE International Solids-State Circuits Conference, Digest of Technical Papers. 1997 Feb 8; 294–295.
12. Piguet C. Low-power CMOS circuits: technology, logic design and CAD tools. CRC press; 2018 Oct 3.

