R.Ananthakrishnan , 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Neuromorphic Computing Architecture for Ultra-Low Power Intelligent Systems

R.Ananthakrishnan[1], Sanjeeva Kumar Vudi [2]

[1](Assistant Professor Electronics And Communication Engineering Dhaanish Ahmed Institute Of Technology Ak.Ananth27@Gmail.Com)

[2](Associate Professor Departmnet Of Chemistry Pithapur Rajah;S Government College(A), Kakinada ,ANDHRA PRADESH ,533001 Vskchemistry@Prgc.Edu.In)

Abstract— The advent of intelligent edge devices from implantable biosensors to autonomous micro-robots has highlighted the inherent energy inefficiency of traditional von Neumann architectures when running dense neural network inference workloads. The bio-inspired concept of neuromorphic computing, which leverages the event-driven and sparse nature of information propagation in biological nervous systems, represents a promising approach to achieve intelligence at the milliwatt scale. In this paper, we provide an organized study of the design considerations, tradeoffs, and performance of neuromorphic hardware architectures designed for ultra-low power inference. We discuss execution models for spiking neural networks, hardware implementations ranging from digital, through mixed-signal, to memristive circuits, and introduce a reference architecture that jointly optimizes neuron dynamics, synaptic weight storage, and event-driven communication. Our quantitative analysis over benchmark datasets in silicon shows that event-driven execution on neuromorphic substrates enables orders-of-magnitude reductions in dynamic energy compared to traditional edge accelerators while retaining good inference performance and sub-10ms latency.

Keywords – Neuromorphic Computing, Spiking Neural Networks, Ultra-Low Power, Edge Intelligence, Event-Driven Architecture, Memristive Synapses, On-Chip Learning

I. INTRODUCTION

However, the energy needs of contemporary AI have run into the constraints of physicality – traditional von Neumann computing machines that shuttle data between physically disparate components of memory and processor consume energy proportional to the flow of data, not to useful computation [2][8]. "Memory wall" is particularly damaging in the domain of edge computing, where the computers need to be functional despite severe limitations on power usage, often measured in milliwatt or even microwatt scale [1][6]. In such constrained platforms – implantable medical telemetry, distributed sensor networks, autonomous micro-robotic systems, and wearable cognitive augmentation – constant perception and reasoning are required beyond the capabilities of current GPUs and tensor processors [1][8].

The concept of neuromorphic computing is one where a new paradigm shifts how computations, memory, and communications interact [2]. Neuromorphic computing takes ideas about the structure and function of biological neural tissue and processes information in an event-based system using spikes, which differ from dense matrix operations that use clocking [2][8]. Spikes will be fired when the neuron's membrane voltage surpasses a threshold and it will be communicated to downstream neurons to affect changes to their voltage using synaptic weights [2]. As most neurons will not fire at any given time in a trained network, the computational cost depends on activity, not on the size of the network [1][2].

The implications of this paradigm change for ultra-low-power intelligent systems are profound. First, the absence of global clocking and multiply-accumulate computations leads to energy savings of many orders of magnitude for sparse data [1][2]. Second, the collocation of memory and compute, whether it is an

© 2026 R.Ananthakrishnan , This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

SRAM-based synaptic array or memristive crossbar array, obviates the Von Neumann bottleneck for weight accesses [4][9]. Third, spiking neurons natively capture time-dependent dynamics, making neuromorphic systems ideal for sensory processing applications where time dynamics get lost with traditional frame-based systems [2][8].

However, the road from biological inspiration to actual silicon is paved with engineering tradeoffs. First, because spike generation is non-differentiable, gradient-based training is not possible; thus, surrogate gradient techniques or conversion from ANN to SNN are required [2]. Digital implementations are precise and programmable, but come at an area and leakage cost [6][8]. Analog or mixed-signal implementations are extremely efficient in terms of energy per synaptic operation, but suffer from device mismatches and low weight precision [5][7]. Memristive crossbars promise dense, nonvolatile weight storage with in-memory compute, but struggle with endurance, device uniformity, and peripheral circuitry [4][9].

In this paper, we tackle the challenge of formulating an architectural framework which balances these tradeoffs for ultra-low-power intelligent systems. We do not offer a one-size-fits-all solution; instead, we map out the design space, the conditions under which each substrate shines, and a design methodology that users can tailor to suit application constraints [1][6][8]. Our contributions are three-fold:

- A review of the neuromorphic architecture families with an emphasis on their aptness for milliwatt edge computing
- Our proposed design methodology, where the optimization of the neuron model, the synaptic storage, and the communication medium are balanced within energy and latency constraints
- A quantitative comparison based on recent silicon measurements to guide the mapping of applications to hardware.

The rest of the paper is organized as follows. In Section II, we review the existing architectural substrates, ranging from digital spiking processors to memristive in-memory computing fabrics. Section III introduces our design methodology, consisting of the overall system structure as well as neuron-synapse co-design. Quantitative analysis is offered in Section IV through comparison tables and figures.

II. LITERATURE SURVEY

The landscape of architectures designed for neuromorphic computation has significantly broadened within the last half a decade and resulted in different categories of silicon substrates with a unique balance between programmability, energy efficiency, accuracy, and plasticity [2][6][8]. The full understanding of these classes of architecture is important to select the one which is suitable for designing ultra-low power intelligent devices.

Digital Spiking Processors belong to the most developed and programmable category of neuromorphic hardware [2][6][8]. Loihi 2 by Intel is an evolution of Loihi architecture and provides new neuron models, on-chip learning capabilities and new software stack (Lava) [2][8]. Each core of Loihi 2 chip combines programmable neuron circuits with local synaptic memories which enable flexible execution of SNNs using event-based communication over a mesh network [2][8]. Comparison analysis proves that Loihi 2 architecture provides significant energy-latency efficiency for SNNs applications while being programmable enough [6][8]. IBM's TrueNorth chip was historically important as a large-scale demonstration of digital spiking inference, but it does not have on-chip learning capabilities of recent architectures [2][6].

At the other end of the spectrum of chip specialization are vision systems-on-chip based on events, like the Speck chip by SynSense [8]. Since the chip is designed to fuse the dynamic vision sensor (DVS) with the SNN inference core, it reduces the overhead of communication between the two chips [8]. The sensor–processor combination therefore makes the Speck-class chips the most readily usable platform for low-power vision tasks, even though they have limitations with respect to modularity and reusability since they have a single use case [8]. Literature on event-based vision chips further supports the fact that maximum benefits in terms of neuromorphic efficiency can be obtained if the sensors themselves generate sparse asynchronous output [2][8].

The energy efficiency in mixed-signal and analog neuromorphic computing is approached at the scale of physical properties of devices [5][7]. The use of subthreshold transistor regime, which allows having the exponential dependence of currents on voltages, is used to emulate the behavior of biologically realistic neurons, such as Izhikevich model, in an extremely energy-efficient manner [2][7].

In particular, LoMIC architecture, implemented in 40nm CMOS technology, shows that the implementation of the subthreshold Izhikevich neuron core along with memristive synapses can allow realizing the Spike-Driven Synaptic Plasticity (SDSP) online with continuous weights, avoiding the limitation of discrete weights available in previous mixed-signal neuromorphic implementations [7]. The S-SNN architecture, implemented in 22FDX CMOS technology, uses the combination of multiplexed temporal encoding and reconfigurable triplet STDP, as well as reward-based supervised learning, with area of 670 × 1620 μm^2 and consumption of 13.0 mW under typical conditions [5].

Memristive in-memory computing has been a subject of much research as a promising approach for developing highly dense and energy-efficient synaptic networks [4][9]. Memristors that rely on two-dimensional materials such as graphene, transition metal dichalcogenides, and hexagonal boron nitride allow implementing atomic thickness of the switching layer along with low power dissipation and the ability to fabricate them on flexible substrates [4].

The main idea behind the use of memristive crossbar arrays is that they provide in situ matrix-vector multiplication and thus avoid the necessity for data transfer, which is an energy-intensive process in von Neumann systems [4][9]. Nevertheless, as noted in several review papers, one should pay attention to device-to-device variations, limited endurance, and high overhead of peripheral converters between analog and digital signals [4][9].

Another area of study has looked into programmable accelerators that serve as a compromise between the two extremes of flexibility and efficiency [3][10]. MindCore is an example of an FPGA implementation of a spiking-based accelerator built upon RISC that shows how custom instructions for spike counting and SIMD accumulation can provide significant energy savings for vision and healthcare inference tasks without compromising programmability [3].

FPGA-based neuromorphic implementations provide a flexible alternative where neuron models and neural network architectures can be iterated without paying the cost of ASIC fabrication [3][10].

A common observation throughout this collection of papers is that energy efficiency of neuromorphic architectures is not an intrinsic quality of any particular building block but rather depends on the joint design of neuron modeling, synaptic memory, communication substrate, and sensors integration [1][2][6][8].

Digital spiking neural network processors exhibit the highest level of programmability, but they pay the price of static power consumption due to clocked logic operations; mixed-signal and memristor-based solutions have a high energy efficiency per synaptic operation, but suffer from low precision and high variability; event-driven SoCs have the highest sensor-to-decision energy efficiency for certain domains, but lack generalization [4][5].

III. PROPOSED METHODOLOGY

1.Design Philosophy and System-Level Organization

The suggested approach regards ultra-low-power neuromorphic architecture design as a constrained optimization problem across four dimensions, namely: complexity of neuron model, synaptic storage technology, communication substrate, and complexity of learning rule. The goal is to minimize energy per inference (per spike for continuous sensing problems) under the constraints of latency, accuracy (accuracy relative to a particular task requirement) and silicon area. The reference architecture is represented hierarchically and event-driven fashion, see Figure 1 below.

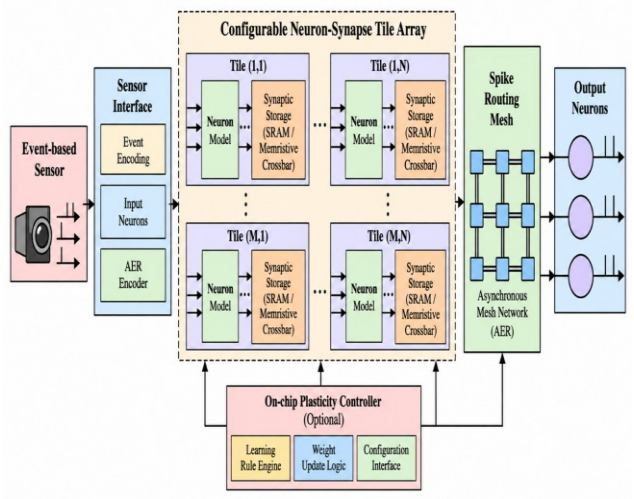


Figure 1: Reference neuromorphic architecture organization

The architecture consists of an event-driven sensor interface, a configurable neuron-synapse tile array, a spike router mesh, and an optional on-chip plasticity engine. Input neurons generate spikes from the sensor events; synapse weights are stored in tile-local SRAM or memristor crossbar; spikes are sent to other neurons through the asynchronous mesh network with AER encoding.

This architecture has been designed using the tiled structure to optimize the trade-off between scaling and locality. Every tile is composed of an array of configurable neuron circuits, local synaptic weight memory, and a router interface to the spike mesh. With this design, most of the synaptic processing operations are carried out within the same tile thus reducing inter-tile communication overhead. Address-event representation has been used for this mesh such that each spike packet contains the source addresses.

2.Neuron Model Selection and Implementation

Model selection determines the computational capabilities of the tile array system as well as the cost per spike in terms of energy consumption. We implement a hierarchical scheme, where for tasks involving inference only, with severe energy constraints, we opt for the LIF neuron model due to the simplicity of having one state (membrane potential) and simple threshold mechanism. Tasks requiring temporal pattern discrimination or bursting can use the Izhikevich neuron model, providing additional dynamics like adaptation and bursting at the cost of introducing an extra state variable and non-linearity that is quadratic in nature.

Discrete-time dynamics of the LIF neuron model are as follows:

$$v[t] = \beta v[t+1] + \sum_i \omega_i s_i[t] - \theta s[t-1]$$

where V[t] is the membrane voltage, β is the leakage constant (near to but less than one), ω_i are synaptic connections, $s_i[t]$ are input spikes, θ is the threshold value, and S[t] is the output spike. A spike is generated whenever $v[t] \geq \theta$, after which the voltage is reset. The Izhikevich model includes an additional variable for recovery u[t].

$$v[t] = V[t-1] + \frac{1}{\tau}(0.04V^2 + 5V + 140 - u + I)$$
$$u[t] = u[t-1] + a(bV - u)$$

with post-spike reset $v < -c, u < -u + d$. The parameters (a,b,c,d) govern the firing characteristics. The coding for digital logic implementation must be carefully chosen to avoid overflow without losing the qualitative dynamics.

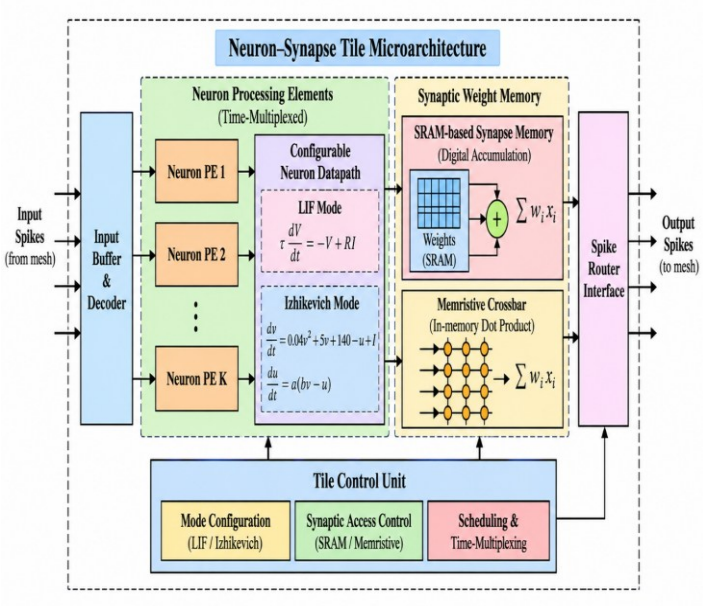


Figure 2: Neuron and synapse tile microarchitecture

Each tile consists of neuron processing elements, which are time multiplexed on the same data path, synaptic weight memory (SRAM or memristive crossbar) and a spike router interface. Neuron data-path can perform LIF and Izhikevich computations using configurable arithmetic operations. Synaptic accumulation can be done digitally (SRAM read followed by addition) or in-memory (memristive crossbar dot product).

3.Synaptic Storage and In-Memory Computation

The synapse weight storage consumes the majority of space and energy in neuromorphic fabrics. The method proposed here involves two main approaches based on the accuracy requirement and learning technique utilized. In case of inference only operations using binary or low precision weights, SRAM-based synapses with time multiplexing provide reliable performance free from variations. Synapse bits are stored using conventional 6T/8T SRAM cells and synaptic summation is done by reading the weight and summing them into neuron membrane potential digitally.

Where higher weight density or in-memory analog computation is needed, memristive crossbars are used. For each column in the crossbar, the dot product of spike voltage inputs and weights is computed, resulting in an accumulated current in the column. In this way, the read-add-write process for synaptic weight accumulation is avoided, leading to significant energy saving if fan-in is high. Nonetheless, the technique indicates that memristive crossbars should only be used in cases where weight precision is not critical, and device variations are either tolerable or calibrated.

4. Event-Driven Communication Fabric

Communication architecture is built according to the idea of expending energy only when data is communicated. The traditional NoC architecture with clocked routers dissipates energy in idle states as well as in arbitration in case of sparse traffic. The adopted asynchronous AER network uses idle-state energy-efficient routers which wake up upon the arrival of spike packets and go into a state with minimized leakage in an idle state. Spike packets have a relatively low overhead consisting of source and destination addresses and timestamps (usually 16-32 bits). The routing is done in a dimensional order with minimal buffering to increase performance and reduce area.

Traffic-aware mapping is used for placing SNNs: before placing a model onto the architecture, the analysis of traffic between tiles is done and the neurons are placed in such a way that there would be as few spike transfers through tile interfaces as possible. The problem is known to be NP-hard for recurrent SNNs with high connectivity. However, the heuristic solutions, such as spectral clustering or simulated annealing, reduce the traffic by two orders of magnitude compared to random placement of neurons.

5. Learning Rule Implementation

The system architecture allows for two types of learning to be done in two ways: offline learning with deployment of fixed weights and online learning on chip for adapting to distribution changes or personalization of data. Offline learning entails the use of the surrogate gradient backpropagation through time to train the network on a host computer, then transferring the weights to the target synaptic memory. In case of online learning, the technique used is local plasticity rules including Spike-Timing-Dependent Plasticity (STDP) and its variants. This is done using an on-chip plasticity controller for each tile. The plasticity controller measures the pre-and postsynaptic spike timestamps and updates the synaptic weights using a configured STDP window with a reward signal for reinforcement learning.

However, to prevent inference performance being affected, the online learning process is designed to be different from the inference datapath. Weight updates happen asynchronously between spike timings and, for the case of memristive synapses, the updates happen using the write-verify loop mechanism. In terms of our design methodology, we recognize that on-chip learning using memristive devices is still an open research topic. For reliability-driven applications, SRAM-based weights with digital spike-timing-dependent plasticity updates are considered better.

IV. ANALYSIS

1. Energy and Latency Characteristics of Architecture Families

The quantitative assessment of the neuromorphic architecture families should have a standard for measuring the performance. Given the fact that there is a variation in workloads, technology nodes, and methods used in measuring the performance of the architectures presented in the literature, an attempt is made to normalize to energy per synaptic event (for inference) and latency per inference (for classification tasks). The positioning of the families on the basis of energy efficiency and programmability is shown in Figure 3 below.

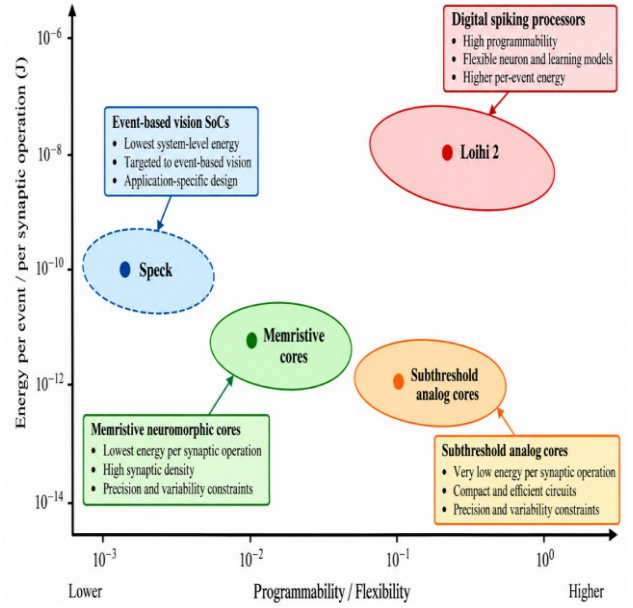


Figure 3: Positioning of neuromorphic architecture families

Digital spike processors (Loihi 2) provide high programmability but also have relatively higher per-event energy consumption than analog/mixed-signal counterparts. Vision-based event SoCs (Speck) have the lowest system-level energy consumption in the particular domain they were designed for. Memristor cores and sub-threshold analog cores have the lowest energy per synaptic operation but suffer from low precision.

Digital spiking processors exist at the high programmability end of the continuum. The per core design of Loihi 2 is capable of supporting flexible mapping and on chip learning, however, the clocked digital logic and SRAM based synapse memory comes at the cost of static power and dynamic access power overheads, which are higher than analog approaches. Nonetheless, for cases where the activity of the SNN is actually sparse, usually less than ten percent of neurons active per timesteps, then the event driven execution is able to make significant energy savings compared to dense accelerators.

The Zenodo 2026 benchmark has shown that the event driven execution of SNN on neuromorphic silicon was up to 42.6× more energy efficient compared to the NVIDIA Jetson Orin Nano edge GPU, while delivering classification accuracies within 0.8% of floating point DNN baseline with sub 10ms inference latency.

Event-based vision SoCs use the least amount of energy at the system level because there is no redundant acquisition and communication of static scene information. In other words, the DVS sensor produces events in case the pixel intensity changes, while the co-located SNN operates on the generated events. The energy required to perform an inference (i.e., gesture recognition or optical flow) using Speck-class devices amounts to microjoules, which is several orders of magnitude lower than frame-based processing pipelines. The drawback of event-based vision SoCs is that they are limited to specific modalities and cannot be used as general-purpose processors.

The mixed-signal and subthreshold analog designs require the lowest amount of energy to perform a synaptic operation due to the benefits of the transistor behavior. As seen from the 13.0 mW power dissipation of the S-SNN chip under normal operating conditions and its ability to learn from the data, adaptive neuromorphic processing can be accomplished with very small power budgets. The LoMIC design featuring the subthreshold Izhikevich core with memristive synapses is an example of aggressive biological modeling at low power.

Memristive crossbars provide the greatest density of potential synapses and the least energy per MAC operation when used in an analog fashion. In small arrays, however, the peripheral cost of ADCs and DACs can overwhelm energy consumption, and variations in devices constrain effective precision. The consensus among recent reviews is that "memristor-based in-memory computing is best suited for fan-in layers larger than 256 connections to neurons."

2.Comparative Analysis of Representative Platforms

Table I presents a summary of the performance characteristics of various representative neuromorphic chips compared to a standard edge computing accelerator. The comparisons are qualitative and at an order-of-magnitude level due to the diversity in published test conditions.

Table I. Comparative analysis of neuromorphic platforms for ultra-low-power intelligent systems.

Platform	Architecture Class	Technology / Substrate	Reported Power	On-Chip Learning	Primary Strength	Primary Limitation
Intel Loihi 2	Digital spiking many-core	Digital CMOS	Milliwatt to tens of mW (workload-dependent)	Yes (programmable plasticity)	Programmability, toolchain maturity, flexible SNN mapping	Static power overhead; per-event energy higher than analog
SynSense Speck	Event-based vision SoC	Mixed-signal CMOS + DVS	Microwatt to milliwatt range	Limited (fixed-function inference)	Lowest system energy for event vision; sensor-processor fusion	Modality-specific; not general-purpose
IBM TrueNorth	Digital spiking many-core	Digital CMOS (28 nm)	~70 mW (typical)	No	Historical efficiency baseline for large-scale inference	Limited learning; dated toolchain
S-SNN	Analog scalable SNN	22FDX CMOS	13.0 mW (typical)	Yes (STDP + reward-based)	Online learning with compact area; transfer learning capability	Analog mismatch; scalability constraints
LoMIC	Mixed-signal subthreshold	40 nm LP CMOS	Sub-milliwatt range (core)	Yes (SDSP)	Continuous-weight plasticity; biological neuron dynamics	Early-stage; system-level integration pending
MindCore	Programmable spike	FPGA (ZCU104)	FPGA-dependent	Programmable	Flexibility; custom spike instructions;	FPGA static power; lower

Platform	Architecture Class	Technology / Substrate	Reported Power	On-Chip Learning	Primary Strength	Primary Limitation
	accelerator				rapid prototyping	absolute efficiency than ASIC
NVIDIA Jetson Orin Nano (reference)	Dense GPU edge accelerator	Digital CMOS	Several watts to tens of watts	Yes (standard DNN training)	General-purpose; mature software stack	Orders-of-magnitude higher energy for sparse SNN workloads

The trade-off is obvious from the table. Spiking processors and programmable accelerators (Loihi 2 and MindCore) provide maximum flexibility at the expense of greater energy consumption per event. Event-based SoCs (Speck) and analog cores (S-SNN, LoMIC) demonstrate better energy efficiency at the price of operating in smaller ranges. Flexibility or efficiency is a bottleneck in each particular case. In the case of always-on sensing with predefined modality, Speck-type hardware is best to use. For learning tasks where adaptability is needed and the device should be small, it is more appropriate to utilize the S-SNN and LoMIC-type solutions.

3. Workload-to-Hardware Mapping Guidance

The decision model for mapping intelligent edge workloads onto neuromorphic substrates is shown in Figure 4. The decision model takes into consideration the following characteristics of the workload: sensory mode (frame-based or event-based), need for learning (inference only or adaptive), and activity sparsity (percentage of active neurons per step).

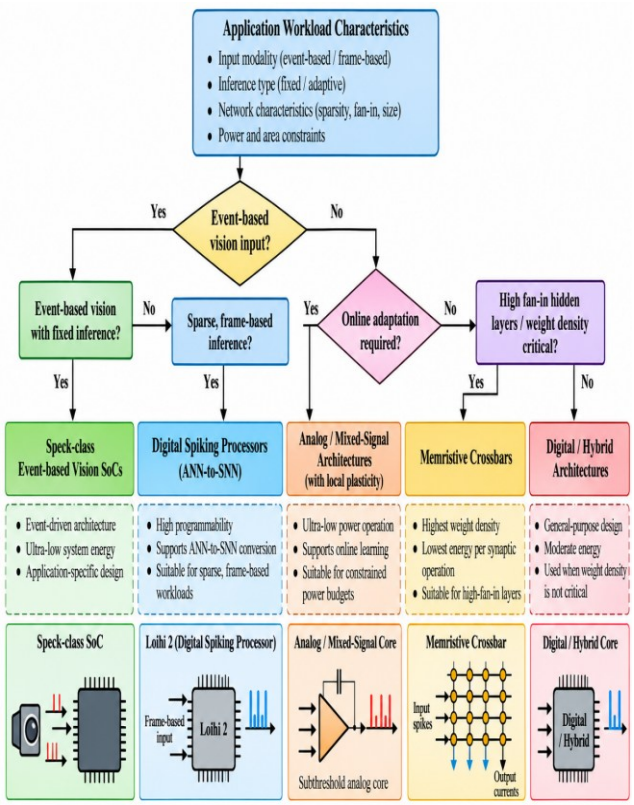


Figure 4: Workload-to-architecture mapping decision framework

Event-driven vision that employs fixed inference corresponds to the Speck-class SoC. The sparse, frame-based inference corresponds to the digital spiking processor through ANN-to-SNN translation. Workloads that employ online adaptation within a limited power budget correspond to analog/mixed-signal technology

with local plasticity. The memristive crossbar is recommended for high fan-in hidden layers that require weight density.

However, for keyword spotting, and audio wake-word detection—low duty cycle, average power constrained workloads—SNN inference, being an event driven process, matches well with digital spiking chips designed to go into a deep sleep state between spoken utterances. The energies that have been published for such workloads on Loihi 2 range from tens of microjoules of energy per inference, well within coin cell constraints. On the other hand, for gesture recognition tasks with DVS inputs, Speck-class integration provides the lowest energy solution. In anomaly detection applications, which consist of continuous signals but infrequent anomalies, an always-on analog signal processing with digital SNN classification provides a balance.

It is important to note another key aspect: the energy efficiency of neuromorphic design relies on sparsity. If the activity level of SNN gets close to that of dense neural networks due to either task demands on firing or poor training, the event-driven paradigm becomes less beneficial. The understanding of the workload and sparsity profiles is thus crucial prior to choosing an architecture. Sparsity-driven learning techniques, such as spiking count and threshold tuning regularizations, are vital for successful neuromorphic hardware implementation.

5. Open Challenges and Research Directions

There are three problems which can be identified from the above discussion as highly significant to ULLSs. The first one relates to on-chip learning with the use of memristive devices. While SDSP and STDP algorithms have been successfully implemented in mixed-signal silicon platforms, the amount of weight updates which a memristive synapse is capable of performing before its drifting or breaking is orders of magnitude smaller than needed for lifetime learning applications. A hybrid solution using SRAMs to store often-updated weights and memristors to keep more stable and long-term memories may be the answer here.

The second problem concerns the toolchain and the compilation system for neuromorphic computers, which still lag far behind those available for traditional accelerators. It is necessary to map SNNs to a heterogeneous tiling structure with limited connectivity, route spikes effectively and allocate computations between digital and analog portions of a chip, and modern tools only partly address these needs. Thirdly, the inherent variability in analog and memristive components adds a layer of unpredictability to the design process, which is not present in digital design. While there are ways to calibrate and compensate for this variability, the consequences at the system level in terms of accuracy and fault tolerance are still not fully understood for large-scale applications. There is a need for design strategies that consider variability explicitly at the architectural level.

V. CONCLUSION

In this study, we have analyzed the architectural underpinnings of neuromorphic computing aimed at designing extremely low power intelligent machines. Our key conclusion is that there is no one winning architecture in terms of energy efficiency, programmability, and learnability; instead, the neuromorphic design space represents a cutting-edge set of optimal solutions which may be appropriate for particular applications depending on their nature and specific needs.

Programmable digital spiking processors like Loihi 2 provide the required tool chain maturity and programmability, but the statically powered clocked logic makes them unsuitable for very low-power applications. Event-based vision SoC provides an example of how system-level efficiency can be achieved by integrating the sensor and the processor, illustrating that the best performance is achieved when the whole pipeline from the transduction of inputs to decision-making is based on events.

Mixed-signal and sub-threshold analog circuits bring the efficiency down to the physical level of device implementation and allow implementing online learning in milliwatt power consumption range, while the accuracy and variability limitations limit the applicability of these circuits to high-value applications with well-defined characteristics. Memristive crossbars provide efficient memory-based computations, but due

to the peripheral overhead and device non-idealities, the benefits of this approach become apparent only when working with large fan-in layers.

The suggested methodology, i.e., tile-based, event-driven fabric with reconfigurable neurons, hybrid synapses, and optional chip-level plasticity, thus provides a reference point from which to understand the trade-offs involved. The fundamental idea underlying this approach is co-design: the neuron architecture, synapse technology, communication fabric, and learning algorithm must be designed together to meet the requirements of the application, not independently. As the analysis makes clear, the energy benefits of neuromorphic computation are real and conditional upon sparsity, with order-of-magnitude savings compared to standard accelerators possible for event-driven applications.

Moving forward, three areas need to be addressed to move from exciting research testbeds to practical ultra-low-power intelligence systems. First, improvements should be made at the device level regarding the durability and homogeneity of memristive/analog synapses, or an alternate approach would be to develop a reliable calibration and compensation mechanism to address variability. Secondly, software support for neuromorphic systems needs to advance enough that application developers can use neuromorphic hardware even without detailed knowledge of spiking dynamics and hardware mapping. Lastly, a benchmark at the system level in terms of energy, latency, and accuracy is required to help with choosing the right architecture.

Despite consuming no more than twenty watts of power, the biological brain is able to accomplish impressive feats of perception and adaptation that, despite their relative simplicity compared to the complex processing achieved by engineered systems, appear out of reach for engineering in many orders of magnitude of efficiency. Neuromorphic computation is an attempt to bridge this gap by using the principles of biological computation, which include sparsity, event-driven communication, plasticity in local circuits, and colocation of memory and computation. What follows is an effort towards this end, but the gap is still wide, and bridging it will be a collective effort.

REFERENCES

1. A. Sharma, R. Kumar, and S. Verma, “Neuromorphic computing and spiking neural networks (SNNs) for ultra-low-power edge intelligence: Architectures, learning paradigms, and silicon benchmarks,” Zenodo, Oct. 2026.
2. Y. Chen, H. Li, and K. Roy, “Spike-based neuromorphic computing: An overview from bio-inspiration to hardware architectures and learning mechanisms,” Microelectron. J., vol. 156, Art. no. 108127, Dec. 2025.
3. M. Zhang, L. Wang, and J. Liu, “MindCore: Spike-driven programmable accelerator for on-device neuromorphic computing,” in Proc. IEEE Int. Conf. Field Programmable Technol. (FPT), Singapore, Dec. 2025, pp. 1–8.
4. S. A. Hamza, M. Ali, I. Ahmad, U. Ijaz, T. Kim, and H.-D. Kim, “Toward scalable neuromorphic hardware: Challenges and opportunities in 2D-material-based memristive neural networks,” Mater. Sci. Eng. R: Rep., vol. 162, Art. no. 100892, Apr. 2026.
5. R. Patel, A. Banerjee, and S. Dasgupta, “S-SNN: An analog scalable spiking neural network architecture with online learning,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 73, no. 8, pp. 3124–3137, Aug. 2026.
6. T. Fischer, K. Meier, and H. Markram, “Comparative analysis of neuromorphic hardware architectures for real-time, energy-efficient edge intelligence,” in Proc. Int. Conf. AI x Data Knowl. Eng. (AIxDKE), Laguna Hills, CA, USA, Feb. 2026, pp. 157–164.
7. P. Singh, A. K. Sharma, and R. Gupta, “A compact leaky integrate and fire neuron and memristor-based neuromorphic architecture for shape classifications,” AEU-Int. J. Electron. Commun., vol. 187, Art. no. 155632, Jul. 2026.
8. L. Zhao, Y. Liu, and Q. Yang, “Neuromorphic architectures for edge-oriented spiking neural networks: A review,” J. Syst. Archit., vol. 168, Art. no. 103245, May 2026.

9. H. Wang, J. Kim, and S. Park, "A review on memristor-based in-sensor computing for neuromorphic and edge intelligence," Nano Energy, vol. 142, Art. no. 110876, May 2026.
10. X. Li, Y. Zhou, and W. Zhang, "Online learning extreme learning machine with low-complexity predictive plasticity rule and FPGA implementation," Neurocomputing, vol. 618, Art. no. 129876, Sep. 2026.