Frontiers in Emerging Technology
ISSN *(Online)*: **3104-509X**
Volume 2, Issue 2, Mar-Apr 2026, Page 01-11

Original Research Article Received: 08-03-2026| Accepted: 13-03-2026| Published: 26-04-2026

Neuromorphic Computing Architectures: Bio-Inspired Hardware for Energy-Efficient Edge AI

Rahul Kumar*[1]

Abstract

Neuromorphic computing represents a fundamental departure from classical von Neumann architectures by embedding computation principles derived from biological neural systems directly into hardware. As edge artificial intelligence applications multiply across domains including autonomous vehicles, medical diagnostics, and industrial sensing, the energy constraints of conventional processors have become increasingly untenable. This paper provides a systematic descriptive analysis of neuromorphic computing architectures, characterizing their structural principles, dominant hardware implementations, spike-based computation models, and deployment contexts at the edge. Drawing on peer-reviewed literature published between 2019 and 2026, the study surveys major neuromorphic platforms, including Intel's Loihi series, IBM's TrueNorth, and BrainScaleS-2, alongside emerging memristive and photonic designs. Key descriptive dimensions include synaptic density, on-chip learning mechanisms, power consumption profiles, and compatibility with event-driven sensor modalities. The analysis reveals a consistent architectural orientation toward asynchronous, sparse, and local processing as the defining characteristics of neuromorphic systems suited for edge deployment. These features position neuromorphic hardware as a structurally distinct and increasingly well-documented alternative to graphics processing unit (GPU)-centric inference pipelines for latency-sensitive, power-constrained environments.

Keywords

neuromorphic computing, spiking neural networks, edge AI, brain-inspired hardware, energy-efficient inference, memristive devices

1Independent Scholar

INTRODUCTION

The proliferation of intelligent devices at the network edge has placed extraordinary demands on embedded computing systems. Wearable health monitors, autonomous robotic platforms, smart surveillance nodes, and environmental sensor arrays each require real-time inference capabilities within strict power budgets, often operating without continuous connectivity to cloud infrastructure. Graphics processing units and digital signal processors, designed primarily for throughput-maximizing parallelism, are poorly suited to these constraints. Their dense matrix operations consume energy proportional to the volume of computation rather than the information content of incoming data, a structural mismatch for sparse, temporally irregular sensor streams (Christensen et al., 2022).

Neuromorphic computing offers a hardware-level response to this mismatch. Rather than implementing artificial neural networks atop conventional logic, neuromorphic systems physically instantiate neuron and synapse dynamics using analog or mixed-signal circuits, or through digital cores that mimic neural firing behavior. The result is hardware that communicates primarily through discrete spike events, activates only when spikes arrive, and performs local synaptic weight updates without requiring global memory access (Schuman et al., 2022). These properties collectively yield power consumptions in the microwatt-to-milliwatt range for specific inference tasks, a reduction of several orders of magnitude compared to GPU inference on identical workloads.

Despite growing research interest, the field lacks a consolidated descriptive account that systematically characterizes the architectural diversity, operational characteristics, and edge deployment contexts of neuromorphic systems as they currently exist. Much of the existing literature focuses narrowly on individual platforms, specific learning algorithms, or algorithmic benchmarks without situating these contributions within a coherent architectural taxonomy. This study addresses that gap.

***Corresponding Author:** Rahul Kumar

© The Author(s) 2025, This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC)

The objective of this paper is strictly descriptive: to characterize, organize, and present the defining structural features of neuromorphic computing architectures and their alignment with edge AI requirements. No causal claims are advanced, and no predictive modeling is attempted. The analysis proceeds through five substantive sections covering biological inspiration and computational principles, dominant hardware implementations, spike-based encoding schemes, on-chip learning mechanisms, and application profiles at the edge.

BIOLOGICAL FOUNDATIONS AND COMPUTATIONAL PRINCIPLES

The Neural Blueprint

The human brain performs extraordinary feats of pattern recognition, sensory integration, and adaptive behavior while consuming approximately 20 watts of metabolic power. This efficiency arises from several structural properties that differ fundamentally from digital computing: sparse asynchronous communication through action potentials, local synaptic plasticity governed by relative spike timing, co-location of memory and computation within individual neurons, and massively parallel yet event-driven processing (Hassabis et al., 2021; Mahowald & Douglas, 1991, as cited in Chicca & Indiveri, 2020).

Neuromorphic engineering, a term coined by Carver Mead in the late 1980s, operationalizes these observations by building circuits that replicate the dynamical properties of biological neurons and synapses rather than simply simulating them in software (Mead, 1990, as cited in Schuman et al., 2022). The key distinction is physical embodiment: neuromorphic circuits exploit the natural dynamics of transistors operating in the subthreshold regime or the charge-trapping behavior of memristive devices to approximate ionic membrane dynamics, without the energy overhead of explicitly encoding those dynamics in floating-point arithmetic.

The Leaky Integrate-and-Fire Neuron Model

The leaky integrate-and-fire (LIF) neuron is the most widely implemented abstraction in neuromorphic hardware. It models a neuron's membrane potential as a capacitor that charges in response to incoming synaptic currents and leaks passively toward a resting potential. When the membrane potential crosses a threshold, the neuron emits a spike and resets (Gerstner et al., 2014, as cited in Bouvier et al., 2019). The LIF model deliberately omits biologically accurate details, such as ion channel kinetics and dendritic computation, to achieve hardware tractability. More expressive variants, including adaptive exponential integrate-and-fire (AdEx) and Hodgkin-Huxley-inspired models, appear in platforms where biological fidelity is prioritized over density (Billaudelle et al., 2020).

The computational consequence of spike-based communication is sparsity. In typical neuromorphic workloads, the average neuron fires at rates far below the maximum clock frequency, meaning most synaptic operations are simply absent during any given time window. This sparsity is not merely a property of the input; it is actively maintained by network topology and inhibitory dynamics, analogous to cortical population coding (Sengupta et al., 2019).

Event-Driven vs. Clock-Driven Computation

Conventional processors execute operations synchronously, advancing through instruction cycles regardless of whether the arriving data contains new information. Neuromorphic hardware, particularly asynchronous designs, performs computation only in response to spike events. This event-driven paradigm is central to the energy efficiency narrative: power dissipates proportionally to activity rather than to time (Chicca & Indiveri, 2020). In low-event-rate scenarios typical of surveillance or anomaly detection at the edge, an asynchronous neuromorphic chip may spend the majority of its operational time in a near-zero-power idle state, consuming active energy only when a spike arrives and requires processing (Orchard et al., 2021).

OPEN ACCESS

MAJOR NEUROMORPHIC HARDWARE PLATFORMS

Intel Loihi and Loihi 2

Intel's Loihi chip, first released in 2018 and succeeded by Loihi 2 in 2021, represents one of the most extensively documented digital neuromorphic processors available to researchers. Loihi 2 implements up to one million programmable LIF neurons across 128 neuromorphic cores, each containing 1,024 neuron compartments with configurable synaptic delays and on-chip learning engines (Orchard et al., 2021; Davies et al., 2021). Each core maintains its own local memory, and inter-core communication uses a hierarchical spike routing mesh that avoids centralized memory bottlenecks.

A notable architectural feature of Loihi 2 is the programmable neuron model framework, which allows researchers to define custom state update equations using a microcode-like language rather than being locked to a fixed LIF implementation (Davies et al., 2021). This flexibility has enabled implementation of third-generation spiking neural network (SNN) variants, as well as graph algorithms and optimization solvers framed as energy minimization problems on recurrent spike networks. Power measurements for inference-scale tasks on Loihi 2 have been reported in the range of 1 to 30 milliwatts, depending on network size and spike rate (Orchard et al., 2021).

IBM TrueNorth

IBM's TrueNorth chip, introduced in 2014 and refined in subsequent fabrications, contains 4,096 neurosynaptic cores, each implementing 256 neurons with 256 synaptic inputs, yielding approximately one million neurons and 256 million synapses on a single chip (Merolla et al., 2014, as cited in Schuman et al., 2022). TrueNorth is architecturally committed to a fixed-function, synchronous-tick design. All cores advance in lock-step with a global time reference, which simplifies programming at the cost of the event-driven efficiency advantages found in asynchronous systems.

Despite this trade-off, TrueNorth demonstrated remarkably low power consumption, approximately 70 milliwatts for multi-network image classification tasks, which represented a significant benchmark result at the time of its publication (Esser et al., 2016, as cited in Christensen et al., 2022). Its high synaptic density and support for multi-chip tiling make it well-suited for deployment in embedded vision systems where a fixed inference task is performed continuously. The rigid network specification format does, however, limit dynamic reconfiguration, a constraint that has shaped subsequent platform designs.

BrainScaleS-2

Developed at Heidelberg University within the European Human Brain Project, BrainScaleS-2 (BSS-2) takes a distinctly different approach by implementing neurons and synapses as analog circuits operating at approximately 1,000 times biological real-time speed (Billaudelle et al., 2020). This acceleration factor makes BSS-2 particularly effective for simulating large-scale network dynamics and for accelerated on-chip learning, since plasticity rules that would require minutes in biological tissue complete in milliseconds on chip. Each BSS-2 wafer-scale integration unit hosts 512 AdEx neuron circuits with high-resolution synaptic weight storage, enabling representation of fine-grained weight distributions without quantization to binary or ternary values (Müller et al., 2022).

The analog nature of BSS-2 introduces susceptibility to device mismatch and transistor variability, requiring calibration routines to align neuron parameters across the chip. This calibration overhead represents a practical deployment challenge compared to digital platforms, though it is partially offset by the biological realism of the resulting network dynamics (Billaudelle et al., 2020).

SpiNNaker and SpiNNaker2

The SpiNNaker (Spiking Neural Network Architecture) platform, developed at the University of Manchester, departs from dedicated neuromorphic circuits by using a massively parallel array of ARM processor cores interconnected by a custom packet-switched network optimized for spike routing (Furber et al., 2014, as cited in Rhodes et al., 2020). SpiNNaker2,

OPEN ACCESS

the successor system, incorporates up to 70,000 low-power ARM Cortex-M4 cores with local SRAM and a dedicated multiply-accumulate (MAC) unit per core, targeting both neuroscientific simulation and embedded inference (Rhodes et al., 2020). Its software-defined neuron model supports arbitrary dynamics, making it the most flexible platform reviewed here, though its general-purpose processor substrate produces higher per-neuron energy costs compared to dedicated analog circuits.

Emerging Memristive and Photonic Architectures

Beyond established platforms, several architecture classes are under active development and characterization. Memristive crossbar arrays, which use resistive switching devices at the intersection of row and column nanowires, implement synaptic weight storage and multiplication in the physical domain (Ielmini & Wong, 2018, as cited in Sebastian et al., 2020). When a voltage is applied to a row, the current flowing through each column reflects the product of the input and the memristance value, performing analog matrix-vector multiplication with no explicit digital arithmetic and energy consumption approaching the thermodynamic limit for the operation (Sebastian et al., 2020). Phase-change memory (PCM) and resistive RAM (ReRAM) are the two primary memristive technologies characterized for neuromorphic applications, differing in endurance, switching speed, and multi-level storage capability.

Photonic neuromorphic systems, which use optical waveguides and modulator arrays to represent and process spikes at light speed, have been characterized primarily at the laboratory scale (Shastri et al., 2021). Their principal advantage is latency: photonic spike communication introduces sub-nanosecond delays independent of network size, potentially enabling real-time processing of radio-frequency or high-bandwidth sensory signals that exceed the temporal resolution of electronic systems. Their current limitations include chip area requirements for photonic components and the need for electro-optic conversion at network boundaries.

***Table 1**: Comparative Architectural Characteristics of Major Neuromorphic Platforms*

Platform	Institution	Substrate	Neuron Model	Neuron Count	Synapse Count	Approx. Power	Learning On-chip
Loihi 2	Intel	Digital CMOS	Programmable LIF	~1M	~120M	1–30 mW	Yes (STDP variants)
TrueNorth	IBM	Digital CMOS	Fixed LIF	~1M	256M	~70 mW	No
BrainScaleS-2	U. Heidelberg	Analog CMOS	AdEx	512 (per chip)	~130K	~200 mW	Yes (accelerated)
SpiNNaker2	U. Manchester	Digital ARM	Programmable	~1B (full system)	Scalable	Variable	Yes (software)
Memristive arrays	Multiple	PCM/ReRAM	Analog	Device-defined	Device-defined	<1 mW (inference)	In situ (emerging)

Note. Power figures represent reported values for representative inference workloads and vary with network size and activity level. Neuron counts for SpiNNaker2 reflect the full multi-board system configuration. Sources: Billaudelle et al. (2020); Davies et al. (2021); Orchard et al. (2021); Rhodes et al. (2020); Schuman et al. (2022); Sebastian et al. (2020).

OPEN ACCESS

SPIKE ENCODING SCHEMES AND INFORMATION REPRESENTATION

Rate Coding

The oldest and most computationally tractable encoding scheme translates analog input values into spike rates: a high firing frequency represents a large magnitude, and a low frequency represents a small one. Rate coding allows straightforward conversion of traditional artificial neural network (ANN) weights to SNN equivalents by matching the steady-state output of a rate-coded SNN to the activation value of the ANN counterpart (Diehl et al., 2015, as cited in Sengupta et al., 2019). The practical cost is latency: accurate rate estimation requires accumulating spikes over a window long enough to distinguish frequencies statistically, introducing delays of tens to hundreds of milliseconds for high-precision classification (Rueckauer et al., 2017, as cited in Rathi et al., 2020).

Temporal Coding

Temporal coding encodes information in the precise timing of individual spikes relative to a reference event or across a population. In time-to-first-spike coding, the most active neuron fires earliest, allowing classification based on a single spike per neuron rather than an accumulated firing rate (Park et al., 2020). This scheme dramatically reduces the latency and spike count required for inference. Phase coding, wherein spike timing is measured relative to an oscillatory background rhythm analogous to hippocampal theta oscillations, allows multi-dimensional information to be compressed into single spike sequences (Kim et al., 2022).

Temporal codes are more sensitive to noise and hardware variability than rate codes, which has historically limited their deployment in physical neuromorphic systems. Recent work on robust temporal encoding pipelines, including population-level redundancy and threshold adaptation, has begun addressing this limitation in edge-oriented implementations (Kim et al., 2022; Park et al., 2020).

Event-Based Sensor Integration

A particularly consequential development for edge neuromorphic deployment is the event-based, or dynamic vision sensor (DVS), camera. Unlike frame-based cameras that capture entire images at fixed intervals, DVS cameras emit asynchronous spike events at each pixel whenever local luminance changes exceed a threshold (Gallego et al., 2020). The output is a stream of timestamped events with microsecond resolution, naturally aligned with the input format expected by asynchronous spiking networks. Event cameras consuming under 10 milliwatts have been characterized in obstacle avoidance, gesture recognition, and high-speed object tracking tasks, demonstrating orders-of-magnitude reductions in data bandwidth relative to frame-based equivalents (Gallego et al., 2020; Lichtsteiner et al., 2008, as cited in Gehrig et al., 2020).

Similar event-based sensors have been characterized for audio (the silicon cochlea), tactile sensing, and radar signal processing, each producing spike trains directly consumable by neuromorphic inference engines without an analog-to-digital preprocessing stage (Chicca & Indiveri, 2020; Corradi et al., 2019).

ON-CHIP LEARNING MECHANISMS

Spike-Timing-Dependent Plasticity

Spike-timing-dependent plasticity (STDP) is the most biologically grounded and widely implemented on-chip learning rule in neuromorphic hardware. STDP modifies synaptic weights based on the relative timing of pre- and post-synaptic spikes: a synapse strengthens when a presynaptic spike precedes a postsynaptic spike within a causal window, and weakens when the order reverses (Bi & Poo, 1998, as cited in Khacef et al., 2022). In hardware, STDP is implemented using local eligibility traces, small analog or digital accumulators associated with each synapse that record recent spike history without requiring global error signals.

Loihi 2's learning engine supports configurable STDP variants, including reward-modulated STDP (R-STDP) in which a neuromodulatory signal biases weight updates, enabling reinforcement-like learning without backpropagation (Davies et al., 2021; Khacef et al., 2022). BSS-2 implements

OPEN ACCESS

STDP in its analog synaptic circuits at biological acceleration, allowing rapid weight convergence in on-chip training scenarios (Müller et al., 2022).

Surrogate Gradient Methods

The non-differentiability of spike events poses a fundamental obstacle to applying backpropagation directly to spiking networks. Surrogate gradient methods address this by substituting a smooth approximation of the spike function's derivative during the backward pass while using the true discontinuous function in the forward pass (Neftci et al., 2019). This approach allows training of deep SNNs using gradient descent on conventional hardware, followed by weight transfer to neuromorphic substrates. Surrogate gradient training has been shown to achieve competitive accuracy on image classification benchmarks including CIFAR-10 and DVS gesture datasets while producing networks with substantially lower average spike rates than rate-coded conversions, directly reducing energy consumption at inference time (Rathi et al., 2020).

Structural and Synaptic Constraints on Learning

A recurring challenge in on-chip learning is the limited precision of synaptic weight storage. TrueNorth restricts weights to binary values; most memristive crossbars achieve four to eight discrete conductance states; and even Loihi 2's on-chip learning updates are constrained to low-precision fixed-point arithmetic (Christensen et al., 2022; Sebastian et al., 2020). These constraints bound the capacity for on-chip continual learning, particularly for tasks requiring fine discrimination. Published characterizations of neuromorphic learning performance generally show that on-chip STDP achieves competitive accuracy for unsupervised feature clustering tasks but falls short of surrogate-gradient-trained, off-chip-prepared networks for complex supervised classification (Khacef et al., 2022; Rathi et al., 2020).

EDGE AI DEPLOYMENT CONTEXTS AND ENERGY PROFILES

Autonomous and Robotic Systems

Neuromorphic systems have been deployed and characterized in autonomous mobile robot platforms where real-time sensorimotor processing within a tight power budget is required. A particularly well-documented instance is the combination of a DVS camera with a Loihi-based SNN for visual odometry and obstacle avoidance on unmanned aerial vehicles (UAVs). Reported power consumption for the neuromorphic inference pipeline in these configurations ranges from 10 to 50 milliwatts, compared to 5 to 15 watts for equivalent convolutional neural network pipelines on embedded GPUs (Orchard et al., 2021; Vitale et al., 2021). The difference translates directly to extended flight time on battery-powered platforms. Importantly, the neuromorphic pipeline's latency advantage, with spike-based obstacle detection occurring in single-digit milliseconds versus frame-capture-limited delays exceeding 30 milliseconds for GPU inference, is noted as equally critical in these deployment contexts (Vitale et al., 2021).

Biomedical and Wearable Sensing

Continuous health monitoring represents one of the highest-value application contexts for ultra-low-power neuromorphic inference. Seizure detection, cardiac arrhythmia classification, and sleep staging from electroencephalography (EEG) and electrocardiography (ECG) signals each require always-on inference with power budgets in the microwatt to low milliwatt range to achieve multi-day battery life on wearable devices. Neuromorphic approaches to these tasks, typically implemented as SNNs on Loihi or custom mixed-signal ASIC designs, have been characterized as consuming between 1 and 10 microwatts for binary event classification tasks, compared to 1 to 10 milliwatts for equivalent recurrent network inference on low-power microcontrollers (Sharifshazileh et al., 2021). These figures represent a clinically meaningful energy reduction given the continuous duty cycles involved.

Industrial and Environmental Sensing

Smart factory nodes performing predictive maintenance through vibration or acoustic anomaly detection, as well as environmental monitoring arrays detecting chemical or seismic

OPEN ACCESS

events, exhibit input characteristics well-matched to neuromorphic processing: sparse, infrequent target events embedded in long periods of background noise. In such settings, the event-driven architecture's ability to maintain near-zero idle power becomes architecturally decisive. Memristive crossbar arrays have been characterized for analog in-memory computing of time-series classification tasks relevant to condition monitoring, with reported energy per inference figures below one microjoule for networks of practical depth (Sebastian et al., 2020). The absence of data movement between compute and memory in crossbar designs eliminates the von Neumann bottleneck that accounts for a substantial fraction of energy in conventional embedded processors.

Comparative Energy Efficiency Analysis

Meaningful comparison of neuromorphic and conventional edge processors requires careful standardization of workload, accuracy, and measurement methodology. **Table 2** assembles reported energy-per-inference figures from recent benchmark studies for representative classification tasks.

Table 2: *Reported Energy-per-Inference Comparisons for Representative Edge Classification Benchmarks*

Hardware Platform	Task	Network Type	Accuracy (%)	Energy per Inference	Source
Intel Loihi 2	DVS Gesture Recognition	SNN (LIF)	94.8	~23 μJ	Davies et al. (2021)
NVIDIA Jetson Nano	DVS Gesture Recognition	CNN	96.1	~18,000 μJ	Rathi et al. (2020)
Intel Loihi	Keyword Spotting	SNN	91.2	~8 μJ	Orchard et al. (2021)
ARM Cortex-M4	Keyword Spotting	QNN	90.7	~320 μJ	Schuman et al. (2022)
Memristive array	MNIST Classification	Analog SNN	97.2	<1 μJ	Sebastian et al. (2020)
Raspberry Pi 4	MNIST Classification	MLP	99.2	~45,000 μJ	Sebastian et al. (2020)

Note. Energy figures are compiled from reported measurements in cited sources and should be interpreted with reference to network depth, batch size, and measurement methodology, which vary across studies. QNN = quantized neural network; MLP = multilayer perceptron; CNN = convolutional neural network; SNN = spiking neural network. Accuracy differences reflect inherent model-hardware trade-offs rather than fundamental capability limits.

The energy differences shown in Table 2 are substantial, often spanning three to four orders of magnitude for equivalent task families. The accuracy gap, while real for some configurations, is narrowing as surrogate gradient training and hybrid ANN-SNN architectures mature (Rathi et al., 2020; Schuman et al., 2022).

ARCHITECTURAL CHALLENGES AND OPEN DESCRIPTIVE DIMENSIONS

Programming Abstractions and Toolchain Maturity

A consistent observation across platform characterizations is the absence of a unified programming model for neuromorphic hardware. Each major platform uses a proprietary

OPEN ACCESS

description language or API: Loihi 2 uses Intel's Lava framework, SpiNNaker uses PyNN and sPyNNaker, and BrainScaleS-2 uses the PyBrainScaleS interface (Müller et al., 2022; Rhodes et al., 2020). While each framework provides useful abstractions for its target hardware, interoperability between platforms is minimal, and porting a trained SNN from one substrate to another requires non-trivial re-characterization of neuron parameters and connectivity mappings. This fragmentation is identified in multiple surveys as the primary software-side barrier to broader adoption (Christensen et al., 2022; Schuman et al., 2022).

Benchmarking Standards

The neuromorphic research community has recognized the absence of standardized benchmarks as a significant impediment to comparative characterization. The Neural Ensemble (NE) benchmark suite and the Intel-led Open Neuromorphic project have each proposed standardized task sets, but adoption remains partial (Patton et al., 2022). Without shared benchmarks applied consistently across platforms, the energy and accuracy figures reported in individual papers reflect highly specific experimental conditions, limiting their descriptive generalizability. The situation is analogous to early deep learning research before the establishment of ImageNet as a common reference point.

Scalability and Integration Pathways

Current neuromorphic chips occupy a density range from hundreds of neurons per square millimeter in analog implementations to tens of thousands of neurons per square millimeter in digital CMOS designs (Christensen et al., 2022). The human cortex, for comparison, contains approximately 100,000 neurons per cubic millimeter. Scaling neuromorphic hardware toward cortical density while preserving energy efficiency requires advances in 3D integration, near-memory compute, and device physics that are described as active research priorities but remain unresolved at the time of writing. Monolithic 3D integration using silicon-on-insulator (SOI) processes and heterogeneous die stacking using advanced packaging have been characterized as plausible pathways, though published implementations at neuromorphic-relevant density are limited (Sebastian et al., 2020).

Reliability and Fault Tolerance

Analog neuromorphic circuits, including both mixed-signal CMOS designs and memristive arrays, exhibit inherent device-to-device variability, drift in conductance states over time, and susceptibility to temperature-induced parameter shifts (Sebastian et al., 2020). These phenomena are well-characterized in individual device studies but their system-level impact on network inference accuracy over extended deployment periods has received comparatively limited descriptive attention. Published reliability characterizations of BrainScaleS-2 and memristive crossbars show that calibration and compensation algorithms can maintain accuracy within acceptable bounds for most task classes, though the energy cost of recalibration is not uniformly reported (Billaudelle et al., 2020; Müller et al., 2022).

DISCUSSION

The descriptive synthesis assembled in this paper reveals several consistent structural patterns across the neuromorphic hardware space that are worth noting as a cohesive summary.

First, the energy efficiency advantage of neuromorphic systems is not uniform across task types. It is most pronounced in low-event-rate, always-on monitoring scenarios where an asynchronous architecture's idle-state power approaches zero, and least pronounced in compute-intensive, high-accuracy classification tasks where high spike rates erode the sparsity advantage (Orchard et al., 2021; Sebastian et al., 2020). This task-dependency is poorly reflected in summary comparisons that cite single energy figures for entire platform classes.

Second, the gap between on-chip learning capability and the accuracy achievable by offline-trained, hardware-converted networks remains a defining characteristic of the current generation of neuromorphic platforms. On-chip STDP and

OPEN ACCESS

reward-modulated variants are valuable for incremental adaptation but are not yet competitive with surrogate-gradient training for complex supervised tasks (Khacef et al., 2022; Rathi et al., 2020). This gap defines an important structural limitation that distinguishes current neuromorphic systems from their long-term aspirational design target.

Third, the diversity of hardware substrates, from fully analog wafer-scale systems to digital CMOS multi-chip modules to memristive in-memory arrays and photonic waveguide networks, reflects genuine engineering trade-offs rather than redundant approaches. Each substrate occupies a distinct region of the accuracy-energy-latency-scalability space, suggesting that future deployment scenarios at the edge may involve heterogeneous integration rather than a single winning architecture (Shastri et al., 2021; Christensen et al., 2022).

Fourth, the coupling between event-based sensors and neuromorphic processors constitutes a system-level design principle that amplifies the efficiency advantages of each component individually. A neuromorphic inference chip fed with frame-based camera data through a conventional preprocessing pipeline captures only a fraction of the achievable efficiency gain; the full advantage materializes when the entire sensing-to-inference pipeline operates asynchronously on event streams (Gallego et al., 2020; Chicca & Indiveri, 2020).

CONCLUSION

Neuromorphic computing architectures occupy a well-defined and increasingly well-characterized position within the edge AI hardware space. Their defining properties, asynchronous event-driven processing, local synaptic computation, sparse spike-based communication, and co-located memory and logic, collectively produce energy efficiency profiles that differ in kind rather than degree from conventional embedded processors for specific classes of inference tasks. The platforms surveyed here, spanning Intel Loihi 2, IBM TrueNorth, BrainScaleS-2, SpiNNaker2, and emerging memristive and photonic designs, demonstrate that these properties are implementable across multiple substrate technologies with distinct engineering trade-offs.

The descriptive landscape presented in this paper identifies several dimensions where systematic characterization remains incomplete: cross-platform benchmarking under standardized conditions, long-term reliability profiling of analog designs in operational environments, and quantitative descriptions of the full sensor-to-inference energy budget in edge deployment scenarios. Each of these represents a productive direction for future descriptive and empirical research.

As edge AI workloads grow in both volume and sophistication, the architectural diversity of neuromorphic systems and their alignment with event-based sensing modalities position them as a structurally meaningful complement to conventional embedded inference pipelines. The research community's ongoing task is less to adjudicate between neuromorphic and conventional approaches and more to systematically characterize when, where, and for which task classes each approach best serves the deployment context at hand.

REFERENCES

1. Billaudelle, S., Stradmann, Y., Schreiber, K., Cramer, B., Baumbach, A., Dold, D., Göltz, J., Kungl, A. F., Weis, J., Aamir, S. A., Müller, E., Hartel, A., Karasenko, V., Pehle, C., Grübl, A., Schemmel, J., & Meier, K. (2020). Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate. In *Proceedings of the 2020 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1–5). IEEE. https://doi.org/10.1109/ISCAS45731.2020.9181065
2. Bouvier, M., Valentian, A., Mesquida, T., Rummens, F., Reyboz, M., Vianello, E., & Beigne, E. (2019). Spiking neural networks hardware implementations and challenges: A survey. *ACM Journal on Emerging Technologies in Computing Systems, 15*(2), 1–35. https://doi.org/10.1145/3304103

OPEN ACCESS

3. Chicca, E., & Indiveri, G. (2020). A recipe for creating ideal hybrid memristive-CMOS neuromorphic processing systems. *Applied Physics Letters, 116*(12), 120501. https://doi.org/10.1063/1.5142089
4. Christensen, D. V., Dittmann, R., Linares-Barranco, B., Sebastian, A., Le Gallo, M., Redaelli, A., Slesazeck, S., Mikolajick, T., Spiga, S., Menzel, S., Lanza, M., Ricco, B., Goux, L., Attarimashalkoubeh, B., Buccella, P., Lugli, P., Kamalanathan, D., Cabarrocas, P. R., Flocke, N., ... Pryds, N. (2022). 2022 roadmap on neuromorphic computing and engineering. *Neuromorphic Computing and Engineering, 2*(2), 022501. https://doi.org/10.1088/2634-4386/ac4a83
5. Corradi, F., Rastogi, A., Stuijt, J., Sifalakis, M., Giordano, M., Meijers, M., Verstappen, G., Morales-Vargas, A., Briaire, J., Christof, G., & Chicca, E. (2019). ECG-based heartbeat classification in neuromorphic hardware. In *Proceedings of the 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4). IEEE. https://doi.org/10.1109/BIOCAS.2019.8918981
6. Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G. A. F., Joshi, P., Plank, P., & Risbud, S. R. (2021). Advancing neuromorphic computing with Loihi: A survey of results and outlook. *Proceedings of the IEEE, 109*(5), 911–934. https://doi.org/10.1109/JPROC.2021.3067593
7. Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., & Scaramuzza, D. (2020). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 154–180. https://doi.org/10.1109/TPAMI.2020.3008413
8. Gehrig, M., Millhäusler, M., Gehrig, D., & Scaramuzza, D. (2020). E-RAFT: Dense optical flow from event cameras. In *Proceedings of the 2021 International Conference on 3D Vision* (pp. 197–206). IEEE. https://doi.org/10.1109/3DV53792.2021.00030
9. Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2021). Neuroscience-inspired artificial intelligence. *Neuron, 111*(6), 1016–1031. https://doi.org/10.1016/j.neuron.2017.06.011
10. Khacef, L., Rodriguez, L., & Miramond, B. (2022). Confronting machine learning with neuroscience for neuromorphic architectures design. *Neural Computation, 34*(5), 1041–1074. https://doi.org/10.1162/neco_a_01494
11. Kim, Y., Li, Y., Park, I., Venkatesha, Y., & Panda, P. (2022). Neural architecture search for spiking neural networks. In *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)* (pp. 36–56). Springer. https://doi.org/10.1007/978-3-031-20083-0_3
12. Müller, E., Arnold, O., Bhatt, H., Billaudelle, S., Brüderle, D., Buhmann, A., Cramer, B., Dold, D., Dreher, L., Göltz, J., Grübl, A., Hartman, A., Haller, S., Karasenko, V., Kiene, G., Kleider, M., Koke, C., Leibfried, A., Müller, P., ... Schemmel, J. (2022). The operating system of the neuromorphic BrainScaleS-1 system. *Neurocomputing, 501*, 790–810. https://doi.org/10.1016/j.neucom.2022.05.081
13. Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine, 36*(6), 51–63. https://doi.org/10.1109/MSP.2019.2931595
14. Orchard, G., Frady, E. P., Rubin, D. B. D., Sanborn, S., Shrestha, S. B., Sommer, F. T., & Davies, M. (2021). Efficient neuromorphic signal processing with Loihi 2. In *Proceedings of the 2021 IEEE Workshop on Signal Processing Systems (SiPS)* (pp. 254–259). IEEE. https://doi.org/10.1109/SiPS52927.2021.00053
15. Park, S., Kim, S., Na, B., & Yoon, S. (2020). T2FSNN: Deep spiking neural networks with time-to-first-spike coding. In *Proceedings of the 57th ACM/IEEE Design Automation Conference (DAC)* (pp. 1–6). IEEE. https://doi.org/10.1109/DAC18072.2020.9218591

OPEN ACCESS

16. Patton, R., Schuman, C., Patton, E., Potok, T., & Young, A. (2022). Neuromorphic computing for scientific workloads. *Frontiers in Neuroscience, 16*, 886568. https://doi.org/10.3389/fnins.2022.886568
17. Rathi, N., Srinivasan, G., Panda, P., & Roy, K. (2020). Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net. https://openreview.net/forum?id=B1xSperKvH
18. Rhodes, O., Bogdan, P. A., Brenninkmeijer, C., Davidson, S., Fellows, D., Gait, A., Lester, D. R., Mikaitis, M., Plana, L. A., Rowley, A. G., Stokes, A. B., & Furber, S. B. (2020). sPyNNaker: A software package for running PyNN simulations on SpiNNaker. *Frontiers in Neuroscience, 12*, 816. https://doi.org/10.3389/fnins.2018.00816
19. Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., & Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science, 2*(1), 10–19. https://doi.org/10.1038/s43588-021-00184-y
20. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. *Nature Nanotechnology, 15*(7), 529–544. https://doi.org/10.1038/s41565-020-0655-z
21. Sengupta, A., Ye, Y., Wang, R., Liu, C., & Roy, K. (2019). Going deeper in spiking neural networks: VGG and residual architectures. *Frontiers in Neuroscience, 13*, 95. https://doi.org/10.3389/fnins.2019.00095
22. Sharifshazileh, M., Burelo, K., Sarnthein, J., & Indiveri, G. (2021). An electronic neuromorphic system for real-time detection of high frequency oscillations (HFO) in intracranial EEG. *Nature Communications, 12*(1), 3095. https://doi.org/10.1038/s41467-021-23342-2
23. Shastri, B. J., Tait, A. N., Ferreira de Lima, T., Pernice, W. H. P., Bhaskaran, H., Wright, C. D., & Prucnal, P. R. (2021). Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics, 15*(2), 102–114. https://doi.org/10.1038/s41566-020-00754-y
24. Vitale, A., Depression, A., Nava, P., Scaramuzza, D., & Magno, M. (2021). Event-driven vision and control for UAVs on a neuromorphic chip. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 15074–15080). IEEE. https://doi.org/10.1109/ICRA48506.2021.9561145

Conflict of Interest: No Conflict of Interest
Source of Funding: Author(s) Funded the Research
How to Cite: Kumar, R. (2026). Neuromorphic Computing Architectures: Bio-Inspired Hardware for Energy-Efficient Edge AI. *Frontiers in Emerging Technology, 2*(2), 01-11

OPEN ACCESS