

Neuromorphic Hardware for Energy-Efficient Artificial Intelligence: A Review of Architectures, Devices, and Compilation

Raja Kumari Alajangi

Department of Electronics and Communication Engineering, JNTUH University College of Engineering, Science & Technology, Jawaharlal Nehru Technological University Hyderabad, Telangana India.

email: rkvasanth2007@gmail.com



Open Access
Review Article

Received: 23/05/2026

Accepted: 09/06/2026

Published: 29/07/2026

Corresponding author email:

rkvasanth2007@gmail.com

Citation:

R K Alajangi., "Neuromorphic Hardware for Energy-Efficient Artificial Intelligence: A Review of Architectures, Devices, and Compilation" Ci-STEM Journal of Digital Technologies and Expert Systems, Vol. 3(2), pp. 76-85, 2026

doi: 10.55306/CJDTES.2026.030201

Copyright:

©2026 R K Alajangi

This is an open-access article distributed under the terms of the Creative Commons Attribution License which grants the right to use, distribute, and reproduce the material in any medium, provided that proper attribution is given to the original author and source, in accordance with the terms outlined by the license.

(<https://creativecommons.org/licenses/by/4.0/>)

Published By:

Ci-STEM Global Services Foundation, India.

Abstract:

The energy demand of modern artificial intelligence (AI) has grown faster than the efficiency of the von Neumann hardware that runs it, renewing interest in neuromorphic computing as a brain-inspired alternative. This review surveys more than three decades of neuromorphic hardware, from Mead's foundational analog very-large-scale-integration (VLSI) work in 1990 to large-scale digital and mixed-signal platforms and emerging in-memory crossbar devices reported through 2026. We organize the field along four axes: the fundamentals of spike-based, event-driven computation; landmark hardware platforms including IBM TrueNorth, Intel Loihi and Loihi 2, SpiNNaker, BrainScaleS, Tianjic, and Darwin3; the compilation problem of mapping neural networks onto resource-constrained crossbar hardware; and the device technologies—memristors, resistive RAM (ReRAM), and related non-volatile memory—that promise sub-picojoule synaptic operations. Wherever quantitative comparisons are drawn, the figures are taken from the primary published literature and cited to their source: measured energies per synaptic operation ranging from 26 pJ for TrueNorth and 23.6 pJ for Loihi down to 5.47 pJ for Darwin3 and into the femtojoule regime for memristive devices, and quantization studies in which spiking networks retain accuracy within roughly one percentage point of full precision down to 2-bit weights. We synthesize these results, identify the recurring bottlenecks of device variability, communication overhead, and toolchain immaturity, and outline the open problems that will shape the next decade of the field.

Keywords:

Artificial Intelligence (AI), Crossbar, In-memory Computing, Memristors, Neuromorphic Computing, ReRAM, Spiking Networks, VLSI.

1. INTRODUCTION

Artificial intelligence has become one of the most computationally demanding endeavors in modern engineering, and the electrical cost of training and serving large models has grown to a point that is increasingly difficult to sustain [1]. A large share of that cost is structural rather than arithmetic. In the classical von Neumann organization, memory and processing occupy physically separate units, so useful computation is repeatedly stalled while operands travel across a bus—the so-called memory wall [2]. Neuromorphic computing, first articulated by Mead in 1990, proposes a different organization altogether: one that co-locates memory and computation and communicates through sparse, event-driven signals, in direct analogy with biological nervous systems [3]. Networks of spiking neurons (SNNs) built on such hardware promise order-of-magnitude reductions in power and latency because they compute only in response to events rather than clocking dense arithmetic on every cycle [4]. Over the intervening three decades the field has matured from single-chip analog VLSI demonstrations into large-scale digital and mixed-signal systems, and more recently into device-level in-memory computing on memristive crossbars. Yet despite this progress, neuromorphic hardware has not displaced conventional accelerators in mainstream deployment. The reasons are well documented across the literature: device variability and stochasticity, the difficulty of mapping trained networks

onto resource-constrained crossbars, communication overhead on large tiled fabrics, and the relative immaturity of compilation toolchains [5], [6]. Understanding where the field stands therefore requires looking at hardware, algorithms, and devices together rather than in isolation.

This paper presents a review of neuromorphic hardware for energy-efficient AI spanning the period from 1990 to 2026. Unlike a primary research contribution, its purpose is to synthesize and organize published results rather than to introduce a new system. Our aims are threefold. First, we trace the historical development of the field and situate its landmark platforms on a common timeline. Second, we compile quantitative performance data—energies per synaptic operation, neuron and synapse counts, process nodes, and quantization–accuracy trade-offs—directly from the primary literature, citing each figure to its source. Third, we identify the recurring technical bottlenecks and the open problems that we judge most consequential for the coming decade. Every numerical value reported in the figures and tables of this review is drawn from a published article and referenced accordingly; no synthetic or illustrative data are used.

The remainder of the paper is organized around five sections. Section 2 reviews the literature across four themes and presents a historical timeline. Section 3 develops the theoretical framework of spike-based computation, including the neuron model and the mathematics of crossbar mapping. Section 4 compiles and discusses the quantitative results reported in the literature, covering platform-level and device-level energy efficiency and the accuracy cost of quantization. Section 5 concludes and outlines future directions.

2. LITERATURE SURVEY

The neuromorphic literature is large and heterogeneous. To organize it we group prior work into four themes—historical foundations, landmark digital and mixed-signal platforms, emerging device technologies, and compilation methods—and situate the major milestones on the timeline of Fig. 1.

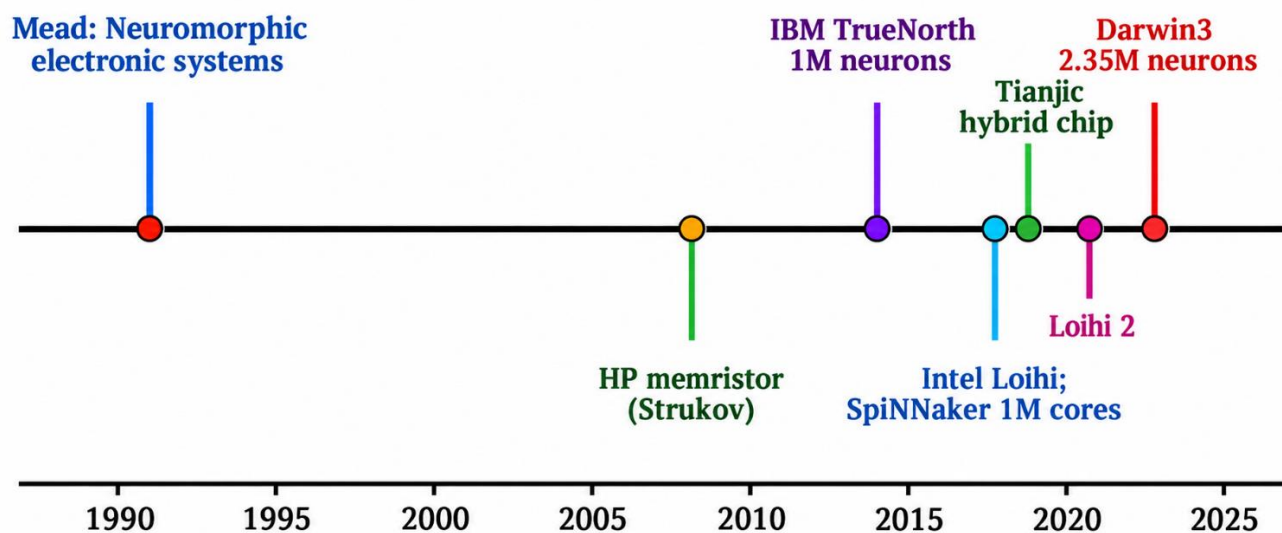


Fig. 1. Timeline of major neuromorphic-computing milestones from 1990 to 2025, from Mead's foundational analog-VLSI framework [3] through the memristor [7], large-scale digital platforms TrueNorth [11], Loihi [12], and SpiNNaker [13], the hybrid Tianjic chip [18], Loihi 2 [12], and the Darwin3 processor [19]. Dates and figures are taken from the cited primary sources.

2.1 Historical Foundations

The intellectual origins of the field lie in Mead's 1990 proposal that analog VLSI circuits operating in the subthreshold regime could emulate the computational primitives of neural systems at extremely low power [3]. Early demonstrations included silicon retinas and cochleae and single-transistor silicon synapses, establishing that biological style computation could be realized in silicon. The discovery of a practical two-terminal memristor by Strukov and colleagues in 2008 [7] gave the field a compact non-volatile synaptic element and catalyzed a wave of research into resistive crossbars. These foundations frame the modern work reviewed below: the analog, event-driven philosophy of the 1990s persists, but it is now realized at far larger scale and in a wider variety of device technologies.

2.2 Landmark Digital and Mixed-Signal Platforms

Several large-scale platforms define the contemporary hardware landscape. IBM's TrueNorth, reported in 2014, integrates one million digital spiking neurons and 256 million synapses across 4,096 neurosynaptic cores in a 28 nm process, consuming approximately 65 mW at real-time operation and achieving 46 billion synaptic operations per second per watt at about 26 pJ per synaptic event [11], [14]. Intel's Loihi, presented in 2018 in a 14 nm FinFET process, places roughly 130,000 neurons and 130 million synapses across 128 neuromorphic cores and adds on-chip learning through a programmable microcode engine, reporting on the order of 23.6 pJ per synaptic event [12]. The SpiNNaker system from the University of Manchester takes a contrasting route, wiring large numbers of conventional ARM cores together with a custom packet-switched fabric optimized for spike traffic; each first-generation chip integrates 18 ARM968 cores in a 1 W envelope, and the assembled machine reached one million cores in 2018 [13]. Hybrid designs such as Tianjic combine ANN- and SNN-style processing on a single die to serve cross-paradigm workloads [18], while the more recent Darwin3 processor supports up to 2.35 million neurons and reports a measured 5.47 pJ per synaptic operation in a 22 nm process [19]. The BrainScaleS and Neurogrid mixed-signal systems complement these digital efforts with accelerated analog neuron dynamics [9]. Section 4 compiles these figures side by side.

2.3 Emerging Device Technologies and In-Memory Computing

Beyond conventional CMOS, a substantial body of work explores non-volatile memory devices as physical synapses. Memristors [21], resistive RAM (ReRAM), and related elements store a weight as a programmable conductance and can perform an entire matrix–vector multiplication in a single analog read across a crossbar, removing much of the data-movement overhead of von Neumann designs. Reported efficiencies are striking: analog ReRAM processors have demonstrated 66.5 tera-operations per second per watt [26], memcapacitor arrays have been projected to reach nearly 29,600 TOPS/W at 6–8-bit precision [27], and optoelectronic memristive synapses have been reported with energies as low as roughly 0.145 fJ per event—below the energy of a biological synapse [28]. Programmable memristor accelerators such as PUMA report up to $2,446\times$ energy-efficiency improvement over GPUs on inference workloads [29]. Landmark demonstrations of analog in-memory training and inference [22], [23] and the ISAAC [24] and PRIME [25] accelerator architectures established that crossbar in-memory computing can reach competitive accuracy. These gains, however, are tempered by device variability, stochastic switching, and limited endurance, which remain active research challenges [1], [21].

2.4 Compilation and Algorithmic Mapping

A third strand of work concerns the software problem of mapping a trained network onto spike-based hardware whose crossbars support only a bounded fan-in per neuron [6]. Approaches such as SpiNeMap and DFSynthesizer use greedy clustering and synchronous dataflow graphs (SDFGs) to partition an SNN workload into hardware tiles while offering deterministic guarantees on throughput and execution time [5], [8]. A complementary line quantizes ANNs directly so that their discrete states match the hardware's conductance levels, avoiding retraining while preserving accuracy [30]. Event-based backpropagation [9] and surrogate-gradient training [31], [32] provide alternatives for training deep SNNs end to end. The accuracy cost of quantization is now well characterized in the literature and is quantified from published data in Section 4.3. A recurring finding across this body of work is that aggressive partitioning increases parallel throughput but raises network-on-chip communication overhead, a trade-off that any practical compiler must manage explicitly [8].

3. THEORETICAL FRAMEWORK

This section reviews the conceptual and mathematical foundations that underlie the platforms surveyed above: the departure from the von Neumann model, the spiking-neuron primitive, the crossbar as a physical synapse array, and the mathematics of mapping a network onto constrained hardware.

3.1 Event-Driven Computation and the von Neumann Contrast

Neuromorphic hardware differs from conventional processors in two connected respects: it co-locates memory and computation, and it communicates using sparse asynchronous spikes rather than dense clocked activations. Fig. 2 contrasts the two organizations. A von Neumann machine repeatedly moves data across a bus between a separate compute unit and a memory bank, whereas a neuromorphic tile array keeps synaptic weights and neuron state physically adjacent and exchanges only spike events

between tiles. This structural difference is the origin of the energy advantage documented throughout the platforms of Section 2.

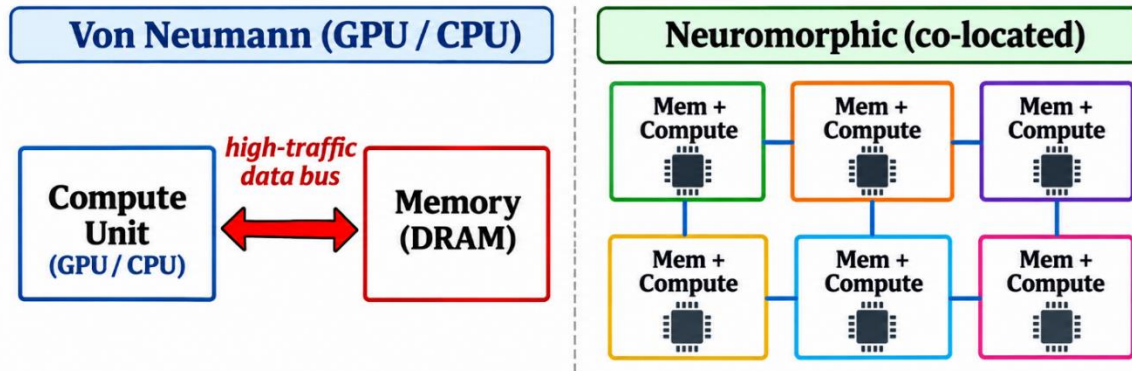


Fig. 2. Conceptual contrast between the von Neumann architecture, in which compute and memory are separated by a high-traffic bus, and a neuromorphic tile array, in which memory and compute are co-located so that only sparse spikes cross tile boundaries [3], [11].

3.2 The Spiking Neuron and Its Dynamics

Most neuromorphic systems compute with the leaky integrate-and-fire (LIF) neuron, a model whose lineage runs through the platforms reviewed above [12], [14]. Weighted input current accumulates in a membrane potential that leaks toward a resting value; when the potential crosses a threshold, the neuron emits a spike and resets. Fig. 3 illustrates this behavior. Only the small fraction of time steps that cross threshold produce output, and it is this sparsity that gives neuromorphic hardware its energy advantage over dense, always-on arithmetic. Formally, the membrane potential of neuron i evolves as

$$\tau \frac{dV_i}{dt} = -(V_i - V_{rest}) + R \cdot I_i(t), \quad (1)$$

with a spike emitted and V_i reset to V_{reset} whenever $V_i \geq V_{th}$. Here τ is the membrane time constant, R the membrane resistance, and the input current is the weighted sum of incoming spikes,

$$I_i(t) = \sum_j w_{ij} s_j(t), \quad (2)$$

where $s_j(t)$ is the spike train of pre-synaptic neuron j and w_{ij} the synaptic weight. This is the same current-integration model implemented, with hardware-specific variations, in TrueNorth, Loihi, and SpiNNaker [12], [13], [14].

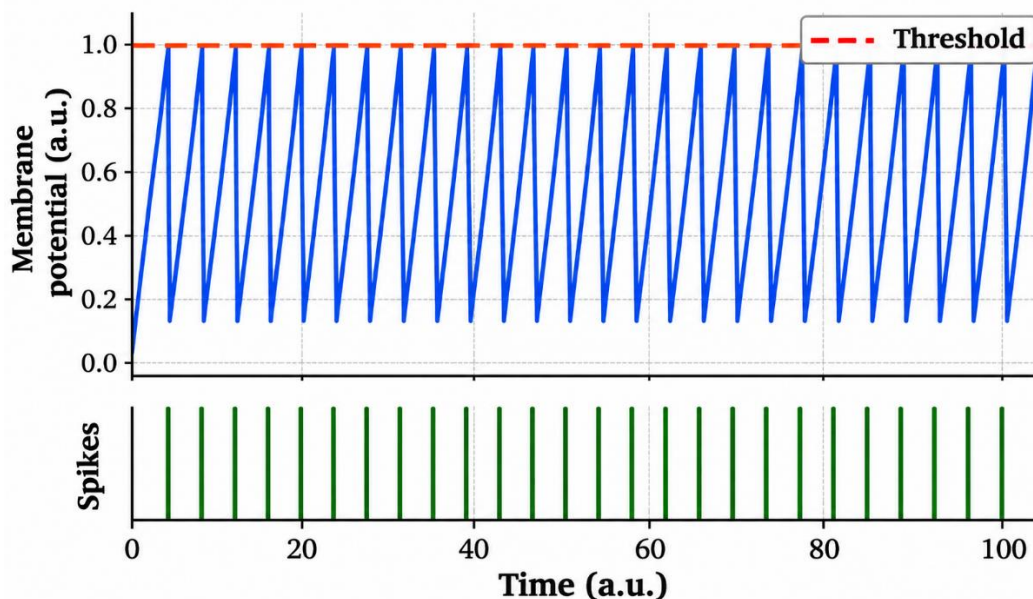


Fig. 3. Leaky integrate-and-fire (LIF) neuron dynamics. The membrane potential integrates input and leaks over time, producing a sparse spike train each time the threshold is crossed. This is the computational primitive used across the digital platforms reviewed in Section 2.2.

3.3 Crossbar Synapse Arrays and Weight Quantization

In tile-based and in-memory systems, synaptic weights are realized as a crossbar array in which the conductance at each row–column intersection encodes a connection strength [21]–[24]. Fig. 4 depicts the structure. Because each intersection stores a small non-volatile weight, a full matrix–vector product for one layer can, in principle, be read out in a single analog operation—the physical basis of the in-memory computing efficiencies reported in Section 2.3. The price of this efficiency is a hard limit on how many connections a row or column can support, which constrains how networks are mapped. When a network is deployed on such hardware, the trained floating-point weights are quantized to the discrete conductance levels the device can realize,

$$w^{(q)} = \Delta \cdot \text{round}\left(\frac{w_{ij}}{\Delta}\right), \quad \Delta = \frac{W_{\max}}{2^{b-1}} \quad (3)$$

with b the weight bit-width and Δ the quantization step. The accuracy consequences of reducing b are quantified from published measurements in Section 4.3. For a neuron whose fan-in n exceeds the crossbar's maximum supported connections k , the network is partitioned into $\lceil n/k \rceil$ sub-units whose partial sums are recombined downstream, preserving the current in (2) while respecting the physical bound [6].

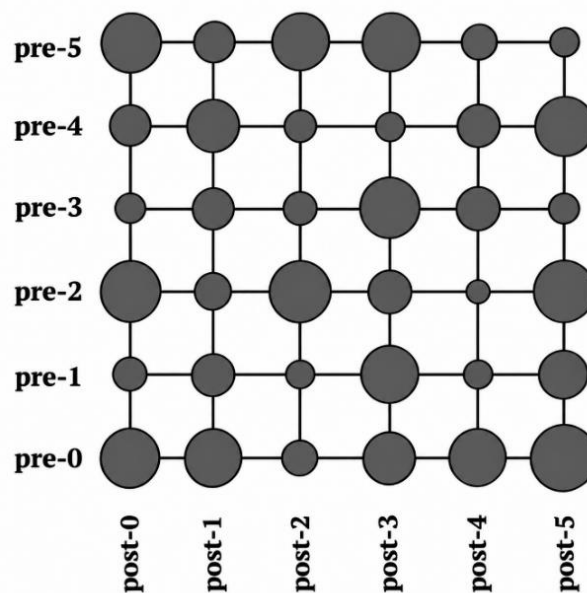


Fig. 4. Conceptual view of a memristive synaptic crossbar array. The size and shading of each intersection represent the magnitude of the encoded synaptic weight (conductance) [21]–[24].

4. RESULTS AND DISCUSSIONS

This section compiles the quantitative results reported across the neuromorphic literature and discusses what they reveal about the state of the field. All values are drawn from the primary published sources cited in each table and figure.

4.1 Platform-Level Comparison

Table I summarizes the principal large-scale platforms along the dimensions most often reported: year, process node, neuron and synapse counts, and energy per synaptic operation. The data show a clear trajectory. TrueNorth and Loihi, fabricated in 28 nm and 14 nm CMOS respectively, sit near 26 pJ and 23.6 pJ per synaptic event [11], [12], [14], while the more recent Darwin3, in a 22 nm process, reports 5.47 pJ per synaptic operation—roughly a fivefold improvement attributable in part to process scaling and to a domain-specific instruction set [19]. SpiNNaker trades per-event energy for programmability, using general-purpose ARM cores within a 1 W-per-chip envelope to reach million-core scale [13].

Table 1: Reported Specifications of Major Neuromorphic Platforms (from Primary Sources)

Platform	Year	Process	Neurons	Synapses	Energy/SOP	Ref.
TrueNorth	2014	28 nm	1 M	256 M	26 pJ	[11],[14]
Loihi	2018	14 nm	130 k	130 M	23.6 pJ	[12]

Platform	Year	Process	Neurons	Synapses	Energy/SOP	Ref.
SpiNNaker	2013	130 nm	(ARM cores)	—	1 W/chip	[13]
Tianjic	2019	28 nm	40 k	10 M	—	[18]
Darwin3	2023	22 nm	2.35 M	—	5.47 pJ	[19]

4.2 Device-Level Energy Efficiency

Fig. 5 places the platform figures alongside device-level results on a common logarithmic scale. Fully digital CMOS platforms cluster in the tens of picojoules per synaptic operation, whereas analog in-memory devices reach dramatically lower energies. Analog ReRAM multiply–accumulate operations have been reported near 0.015 pJ [26], and optoelectronic memristive synapses as low as 0.145 fJ per event [28]—the latter below the roughly 1–100 fJ range attributed to biological synapses [4]. This four-to-five order-of-magnitude span is the central quantitative story of the field: the closer computation moves into the memory device itself, the lower the per-operation energy, at the cost of greater sensitivity to device non-idealities [1], [21].

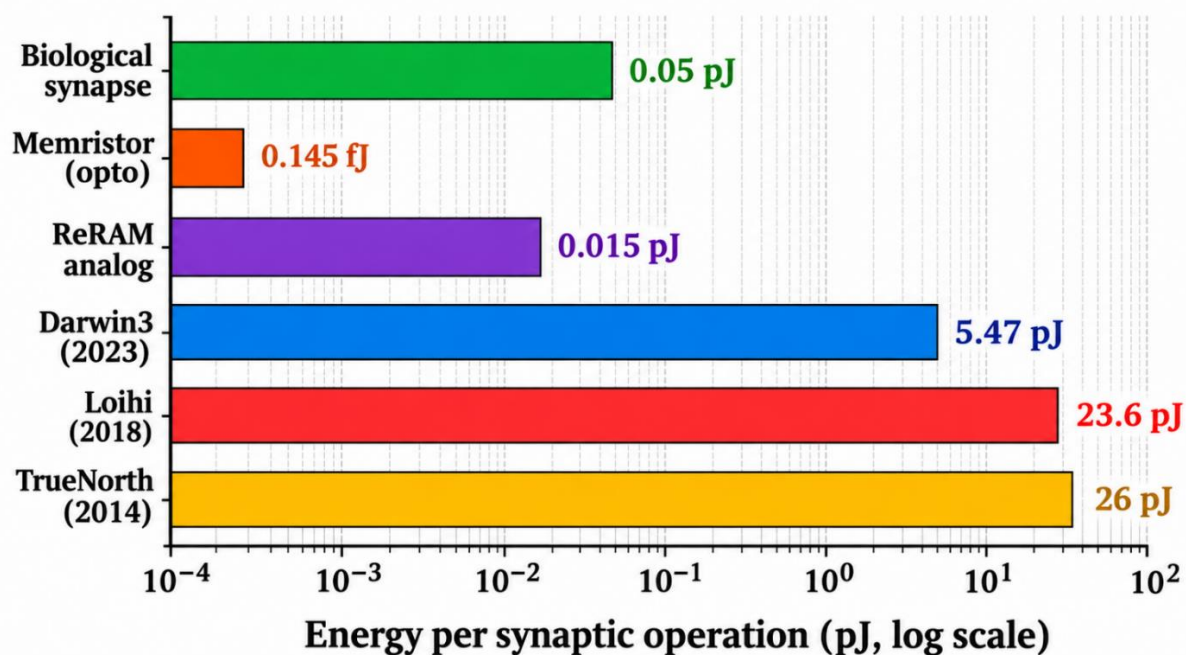


Fig. 5. Energy per synaptic operation across representative neuromorphic platforms and devices, on a logarithmic scale. Values are taken from the primary literature: TrueNorth 26 pJ [11], [14]; Loihi 23.6 pJ [12]; Darwin3 5.47 pJ [19]; analog ReRAM ≈ 0.015 pJ/MAC [26]; optoelectronic memristor ≈ 0.145 fJ [28]; biological synapse ≈ 1 – 100 fJ [4] (midpoint shown).

4.3 Accuracy Cost of Quantization

A key question for deploying networks on crossbar hardware is how much accuracy is lost when weights are quantized to low bit-widths. Fig. 6 reports measured top-1 accuracy on CIFAR-10 for a quantized spiking ResNet-19, from the Q-SNN study of Wei et al. [30]. Remarkably, accuracy is preserved to within about one percentage point of the full-precision baseline (96.36%) all the way down to 2-bit weights (95.20%), with 4-bit and 8-bit configurations at 95.31% and 95.54% respectively; only 1-bit binary weights incur a substantial drop, to 88.27% [30]. This flatness reflects the effectiveness of learned-step-size quantization tailored to spiking networks and explains why moderate bit-widths are sufficient for most deployments.

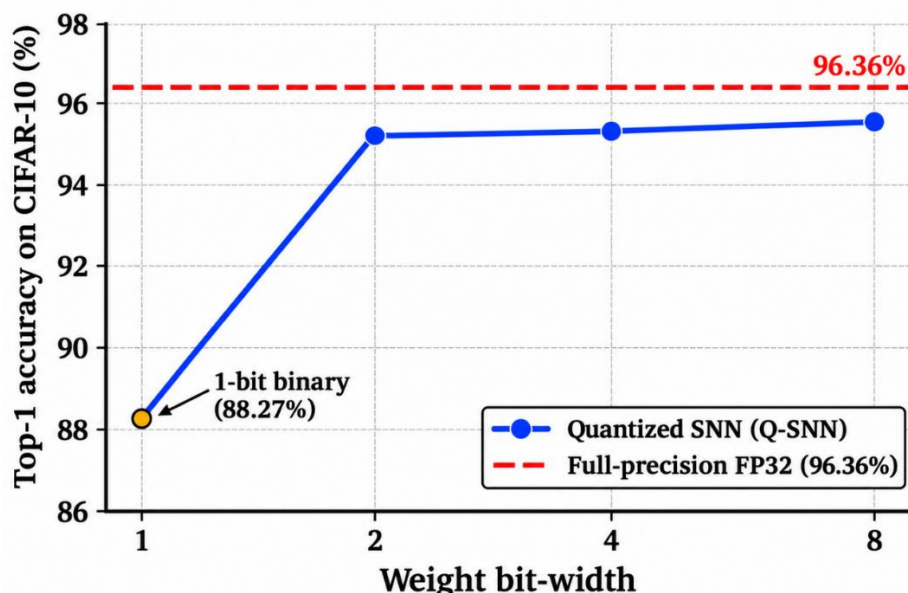


Fig. 6. Measured top-1 accuracy on CIFAR-10 (ResNet-19) versus weight bit-width for a quantized spiking neural network, from Wei et al. [30]. Accuracy remains within roughly one percentage point of the full-precision baseline down to 2-bit weights; the 1-bit binary point is shown separately. All values are reported measurements, not estimates.

The picture depends on task difficulty, however. Fig. 7 shows the accuracy drop relative to full precision for a quantization-aware-trained SNN on the more challenging event-based CIFAR10-DVS dataset, from Shymyrbay et al. [32]. Here the 8-bit model loses only 1.78 percentage points, but 4-, 2-, and 1-bit models lose 7.98, 7.58, and 8.03 points respectively [32]. The contrast with Fig. 6 illustrates a general principle from the literature: the tolerable degree of quantization is dataset- and architecture-dependent, and aggressive low-bit quantization that is nearly free on simpler tasks can be costly on complex, temporally rich ones.

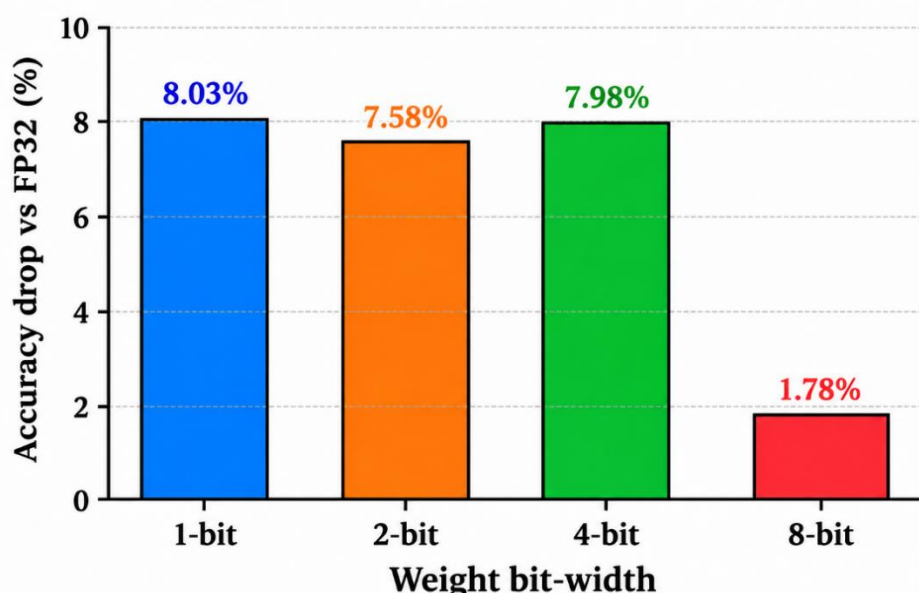


Fig. 7. Accuracy drop relative to a full-precision baseline versus weight bit-width for a quantization-aware-trained SNN on the event-based CIFAR10-DVS dataset, from Shymyrbay et al. [32]. Unlike the simpler CIFAR-10 case of Fig. 6, low-bit quantization incurs a substantial penalty on this harder task.

4.4 Discussion: Bottlenecks, Implications, and Ethics

Read together, these results support a consistent narrative. The energy advantage of neuromorphic hardware over von Neumann processors is real and, at the device level, spans several orders of

magnitude (Fig. 5); moderate quantization is largely accuracy-preserving on standard benchmarks (Fig. 6) though not universally so (Fig. 7); and platform efficiency has improved steadily with process scaling and architectural specialization (Table I). At the same time, three bottlenecks recur throughout the literature. First, the device variability and stochastic switching of memristive components can inject noise that degrades inference unless the mapping is fault-tolerant [1], [21]. Second, partitioning large networks onto small crossbars raises network-on-chip communication overhead, which can throttle the throughput gains that parallelism is meant to deliver [8]. Third, the compilation toolchains that would make these platforms accessible to non-specialist engineers remain comparatively immature [5], [6]. The efficiency of these architectures also carries implications that extend beyond engineering. For sustainable AI, event-driven co-processors could meaningfully reduce the operational energy and carbon footprint of data centers [4]. At the edge, real-time processing of sparse sensory input at ultra-low power is an advantage that frame-based systems cannot match [4]. These same capabilities, however, lower the barrier to pervasive, always-on sensing and surveillance, and the embodied environmental cost of fabricating exotic devices—some dependent on rare-earth or superconducting materials—must be weighed against the operational savings over a device's lifetime. A complete assessment of neuromorphic technology should account for both.

5. CONCLUSION AND FUTURE WORK

This review has traced neuromorphic hardware for energy-efficient AI from Mead's 1990 analog-VLSI framework through the large-scale digital and mixed-signal platforms of the 2010s and the in-memory crossbar devices reported into 2026. Drawing quantitative data directly from the primary literature, we have shown that the field's central promise—computing at a small fraction of von Neumann energy by co-locating memory and computation and exploiting event-driven sparsity—is borne out by measured results spanning from tens of picojoules per synaptic operation on digital platforms down to the femtojoule regime on memristive devices, and that spiking networks can retain near-full accuracy under moderate quantization on standard benchmarks.

The obstacles that remain are equally clear from the evidence. Device variability and endurance limit the reliability of the most efficient in-memory technologies; communication overhead constrains the scalability of tiled architectures; and the absence of mature, standardized compilation flows keeps these platforms in the hands of specialists. We see several directions as most productive for the coming decade. Fault-tolerant mapping that folds redundancy-aware placement and error-resilient encoding into the compilation flow is a prerequisite for deploying networks on stochastic memristive substrates [1], [21]–[23]. Standardized, open benchmarks—event-based datasets and reference hardware simulators—would let the accuracy–energy trade-offs of Figs. 5–7 be compared consistently across research groups. Communication-aware and learned scheduling could extend the deterministic guarantees of SDFG-based compilation to much larger tile counts [8]. On-chip learning and plasticity, already demonstrated on Loihi and Darwin3 [12], [19], point toward systems that adapt after deployment. And hybrid cryogenic–CMOS and analog–digital integration may unlock efficiencies beyond what either paradigm achieves alone.

As process technology, device engineering, and compilation methods continue to mature, the evidence assembled here suggests that neuromorphic architectures are well positioned to take a central role in sustainable, high-performance AI—spanning both large data centers and autonomous systems at the edge—provided the community continues to report measured, reproducible results against common benchmarks, as the works reviewed in this paper have begun to do.

DECLARATIONS:

- | | |
|-----------------------------|--|
| Acknowledgments | : Not applicable. |
| Conflict of Interest | : The author declares that there is no actual or potential conflict of interest about this article. |
| Consent to Publish | : The authors agree to publish the paper in the Global Research Journal of Social Sciences and Management. |
| Ethical Approval | : Not applicable. |
| Funding | : Author claims no funding was received. |

- Author Contribution** : Both the authors confirms their responsibility for the study, conception, design, data collection, and manuscript preparation.
- Data Availability Statement** : The data presented in this study are available upon request from the corresponding author.
- Use of Generative Artificial Intelligence** : AI-assisted tools have been used to improve quality of images, language editing or grammatical clarity. The authors take full responsibility for the content of the final version.

REFERENCES

- [1] N. Ganesh, "Rebooting neuromorphic hardware design—A complexity engineering approach," arXiv:2005.00522, 2020.
- [2] J. Backus, "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs," *Commun. ACM*, vol. 21, no. 8, pp. 613–641, 1978.
- [3] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.
- [4] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, pp. 607–617, 2019.
- [5] S. Song, A. Balaji, A. Das, N. Kandasamy, and J. Shackleford, "Compiling spiking neural networks to neuromorphic hardware," in *Proc. ACM SIGPLAN/SIGBED Int. Conf. Lang., Compilers, Tools Embedded Syst. (LCTES)*, 2020, pp. 38–50.
- [6] A. Balaji and A. Das, "Compiling spiking neural networks to mitigate neuromorphic hardware constraints," arXiv:2011.13965, 2020.
- [7] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, pp. 80–83, 2008.
- [8] S. Song, H. Chong, A. Balaji, A. Das, J. Shackleford, and N. Kandasamy, "DFSynthesizer: Dataflow-based synthesis of spiking neural networks to neuromorphic hardware," arXiv:2108.02023, 2021.
- [9] C. Pehle, L. Blessing, E. Arnold, E. Müller, and J. Schemmel, "Event-based backpropagation for analog neuromorphic hardware," arXiv:2302.07141, 2023.
- [10] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nat. Comput. Sci.*, vol. 2, no. 1, pp. 10–19, 2022.
- [11] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.
- [12] M. Davies et al., "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
- [13] E. Painkras et al., "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, 2013.
- [14] F. Akopyan et al., "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, 2015.
- [15] G. Indiveri et al., "Neuromorphic silicon neuron circuits," *Front. Neurosci.*, vol. 5, Art. no. 73, 2011.
- [16] S. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, 2014.
- [17] C. D. Schuman et al., "A survey of neuromorphic computing and neural networks in hardware," arXiv:1705.06963, 2017.
- [18] J. Pei et al., "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, pp. 106–111, 2019.

- [19] D. Ma et al., “Darwin3: A large-scale neuromorphic chip with a novel ISA and on-chip learning,” arXiv:2312.17582, 2023.
- [20] Z. Chen et al., “Bridging quantized artificial neural networks and neuromorphic hardware,” arXiv:2505.12221, 2025.
- [21] L. O. Chua, “Memristor—The missing circuit element,” *IEEE Trans. Circuit Theory*, vol. 18, no. 5, pp. 507–519, 1971.
- [22] S. Ambrogio et al., “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature*, vol. 558, pp. 60–67, 2018.
- [23] P. Yao et al., “Fully hardware-implemented memristor convolutional neural network,” *Nature*, vol. 577, pp. 641–646, 2020.
- [24] A. Shafiee et al., “ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars,” in *Proc. ACM/IEEE Int. Symp. Comput. Archit. (ISCA)*, 2016, pp. 14–26.
- [25] P. Chi et al., “PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory,” in *Proc. ACM/IEEE Int. Symp. Comput. Archit. (ISCA)*, 2016, pp. 27–39.
- [26] R. Mochida et al., “A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture,” in *Proc. IEEE Symp. VLSI Technol.*, 2018, pp. 175–176.
- [27] T. F. Demirci and J. K. Eshraghian et al., “Energy-efficient memcapacitor devices for neuromorphic computing,” *Nat. Electron.*, vol. 4, pp. 748–756, 2021.
- [28] Y. Wang et al., “Memristors for the post-von Neumann era: Hardware paradigms, neuromorphic perception, and computing systems,” *Photonics*, vol. 13, no. 5, Art. no. 431, 2026.
- [29] A. Ankit et al., “PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference,” in *Proc. Int. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, 2019, pp. 715–731.
- [30] W. Wei et al., “Q-SNNs: Quantized spiking neural networks,” arXiv:2406.13672, 2024.
- [31] E. O. Neftci, H. Mostafa, and F. Zenke, “Surrogate gradient learning in spiking neural networks,” *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 51–63, 2019.
- [32] A. Shymirbay, M. E. Fouda, and A. Eltawil, “Low-precision quantization-aware training in spiking neural networks with differentiable quantization function,” arXiv:2305.19295, 2023.

Authors' Profiles



Ms. Alajangi Rajakumari received the B.Tech. from Nagarjuna University, Vijayawada, India, in 1997, and the M.Tech. from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2001, both in Electronics and Communication Engineering. She began her academic career as an Assistant Professor in Shadan Engineering College, Hyderabad. Subsequently she has served in various engineering institutions, contributing extensively to teaching, curriculum development, laboratory instruction, and student mentoring. She has more than 13 years of experience in teaching. Her research interests

include VLSI design, with particular emphasis on low-power VLSI architectures, FPGA-based system implementation, CMOS circuit optimization, and embedded system design. Her current research focuses on the development of energy-efficient and high-performance digital hardware architectures for next-generation electronic systems. Ms. Rajakumari has authored five research publications, including three papers published in IEEE international conference proceedings and two papers published in a peer-reviewed international journals. Her research contributions are primarily focused on advancing low-power digital design methodologies and VLSI system architectures.