

# Engineering Neuromorphic Computing Systems for Brain-Inspired Processing

MOSES ADEOLU AGOI<sup>1,a</sup>, OLUWANIFEMI OPEYEMI AGOI<sup>2,b</sup>

<sup>1</sup>Lagos State University of Education  
Lagos  
NIGERIA

<sup>2</sup>Obafemi Awolowo University  
Osun  
NIGERIA

<sup>a</sup><https://orcid.org/0000-0002-8910-2876>

<sup>b</sup><https://orcid.org/0009-0007-9461-8870>

**Abstract:** - Neuromorphic computing is a computer-system design approach inspired by the structure and information-processing principles of biological neural systems. Unlike conventional von Neumann architectures, neuromorphic systems integrate computation and memory more closely and exploit event-driven communication, sparse activity, parallelism, and, in some implementations, on-chip learning. This paper presents a structured, source-audited review of neuromorphic computing, focusing on architecture, spiking neural networks, learning mechanisms, hardware implementation, scalability, energy efficiency, edge intelligence, and engineering challenges. Rather than combining heterogeneous measurements through unsupported statistical meta-analysis, the review distinguishes documented hardware specifications, workload-specific experimental measurements, and reported performance claims. Representative platforms, including IBM TrueNorth, Intel Loihi, Loihi 2, and Hala Point, are examined in terms of neuron and synaptic capacity, architectural organization, learning capabilities, and deployment context. Evidence indicates that neuromorphic systems can provide energy and latency advantages for selected sparse, event-driven workloads, although performance depends strongly on workload characteristics, algorithms, implementation, and measurement methods. TrueNorth demonstrated large-scale digital neuromorphic integration, while Loihi introduced programmable on-chip learning and Loihi 2 expanded neuron-model flexibility and capacity. Hala Point demonstrates further multi-chip scaling potential, but its specifications should not be interpreted as biological equivalence. Recent initiatives such as NeuroBench improve evaluation consistency by distinguishing algorithm-level and deployed-system measurements. Key challenges include training large spiking networks, hardware variability, software interoperability, benchmarking, memory and communication constraints, and integration with established AI workflows. The review concludes that neuromorphic computing is a complementary paradigm with strong potential for sparse temporal workloads, low-latency applications, and power-constrained environments.

**Key-Words:** - Brain-Inspired Computing Neuromorphic Architecture Spiking Neural Networks Synaptic Plasticity.

Received: April 8, 2026. Revised: June 2, 2026. Accepted: July 9, 2026. Published: October 9, 2026.

## 1 Introduction

The high rate of artificial intelligence (AI) application growth, consisting of autonomous vehicles, precision healthcare, complex robotics, cybersecurity analytics, smart agriculture, and Internet of Things (IoT) ecosystems have significantly increased the need to have low-latency, energy-efficient, and adaptable

processing provided by computational infrastructures. The modern AI tasks and especially neural networks and real-time analytics demand large memory bandwidth and parallel processing. Nevertheless, the traditional von Neumann architectures are getting limited more by the so called von Neumann bottleneck which is a structural limitation that exists due to the physical distance between the central processing unit

(CPU) and memory storage. This isolation requires the constant exchange of data between the shared buses and this results in the overhead of latency, higher energy usage, and lower throughput. These inefficiencies increase as the model sizes increase to the billions of parameters and edge devices become more and more prevalent across the planet. Currently, the data centres have already consumed 1-2 percent of the electricity consumed worldwide with the workloads in AI training and inference taking an increasing proportion. Thus, pressure is growing to look beyond the current computing paradigm and consider new computing paradigms that challenge the principles of architectural design to attain sustainability and scalability. Neuromorphic computing as a paradigm change comes in to solve these architectural weaknesses through the inspiration of biological neural systems. In contrast to the classical digital processors that are based on a chain of executed instructions, and the hierarchical memory structure, neuromorphic systems imitate the structure and behavior of the human brain, in which the computation and memory are co-located in the network of neurons and their connections [1]. In the biological neural networks, the processing of information happens through interactions that are distributed and parallel amongst about 86 billion neurons [2] which are linked up through trillions of synapses. These networks are asynchronous, event-driven and incredibly energy efficient (around 20 watts required to execute complicated perceptual and cognitive computations). Neuromorphic engineering aims to recreate these properties in silicon substrates and, hence, minimize energy overhead, increase adaptability, and scalability. Brain-inspired processing models take advantage of massively parallel providers in which the artificial neurons interact by discrete electrical signals, or spikes. The spiking neural networks (SNNs) are a temporal encoding of information (spike sequences) unlike the traditional artificial neural networks (ANNs) which make use of continuous-valued activations and synchronous updates, which is more similar to the dynamics of biological signalling [3]. Computation in the SNNs is event-driven: neurons are only active in response to meaningful signals and as a result, processing is sparse and asynchronous. This sparsity makes a great deal of unnecessary computations redundant thereby reducing the amount of energy used and enhancing the efficiency of the computational process. The temporal coding mechanisms also enable SNNs to learn more naturally the spatiotemporal dependencies and are therefore especially appropriate in activities that involve sensory streams, like vision, audition and touch. Event-driven computation is

particularly beneficial in an environment with edge AI, where power capabilities and thermal limits are highly constrained. The real-time inference functions are necessary in devices that are installed in remote or embedded environments, including wearable medical devices, autonomous robots, and smart sensors, and do not need constant connectivity to the cloud. The neuromorphic systems allow localized learning and inference to control data transmission and latency to achieve on-device intelligence. In addition, neuromorphic processors can also be updated on a dynamically changing connection weights through adaptive synaptic plasticity mechanisms, including Spike-Timing-Dependent Plasticity (STDP), which can enable continuous learning without having to retrain the connection weights. The recent progress in the semiconductor fabrication technologies, nanoscale devices, and new materials including memristors have made the creation of neuromorphic hardware platforms much faster. The memristive devices, which have the ability to store the resistance states as per the historical flow of current, are highly similar to the biological synapses. Memristors provide in-memory computing model through the combination of memory and processing into a single physical component that minimizes the data transfer and the energy expenditure. Moreover, analogue and mixed-signal circuit design allows the low count of transistors of neuron and synapse models to be readily implemented than full-digital architecture. All these innovations result in a better power-performance trade-off and an increase in the densities of synaptic. Neuromorphic processors such as large-scale processors indicate the realistic implementation of brain-inspired computing. The example of IBM is the TrueNorth chip which combines one million programmable spiking neurons and 256 million synapses and also consumes about 70 milliwatts of power when operating [4]. It has a non-von Neumann architecture that allows distributed memory in the cores of neurons allowing scalable communication via an event-driven routing network. On the same note, Intel Loihi processor has a built-in learning functionality, which can be programmed to offer synaptic plasticity and real-time adaptation [5]. Loihi has orders-of-magnitude energy efficiency improvements over the traditional CPU and graphics platforms over sparse workloads, notably in optimization, signal processing, and sensor fusion. The Loihi architecture is also a second-generation that has increased the density of neurons, inter-chip scalability, and flexible algorithmicity, which makes the neuromorphic hardware viable to complex AI applications. In addition to hardware improvements, the neuromorphic ecosystem has also been expanded to feature domain-specific software systems and

Toolchains to close the gap between modeling inspired by neuroscience and implementation in engineering. Such development platforms as Nengo, Brian and Lava allow researchers to design and execute SNN-based applications on neuromorphic chips and simulation platforms. There is also the emergence of hybrid architecture with neuromorphic accelerators and other types of conventional processors, which allows a smooth transition to existing AI processes. These co-design plans allow the incremental adoption process without breaking the compatibility with the existing deep learning pipelines. Neuromorphic engineering offers a viable future of scalable cognitive computing as the field of artificial intelligence goes beyond the distributed and edge-centric paradigms. Neuromorphic accelerator-based edge devices have the ability to make localized inference and adaption and avoid reliance on centralized cloud systems. This decentralization increases the privacy level, reduces the bandwidth of communication as well as increases the resilience of the system. In addition, brain-inspired networks have the intrinsic fault-tolerance and redundancy that provide resilience to failure of hardware and environmental noise that is important to safety-sensitive systems like autonomous navigation and medical diagnostics. However, some serious problems still exist. The computational complexity of training deep spiking neural networks at scale still makes them difficult to train by methods other than surrogate gradient or converting traditional ANNs. The standardization of programming models and benchmarking methodology as well as fabrication processes are in the process of development. Nevertheless, the challenges notwithstanding, the potential of neuromorphic computing is a prospective field of AI that can be explored by the combination of insights into neuroscience, inventions in the materials science field, and scalable hardware design. To conclude, the weaknesses of conventional von Neumann architectures to the growing AI requirements have spurred the study of neuromorphic computing as an inspirational alternative to the brain. The neuromorphic systems can achieve significant improvements in energy efficiency, latency reduction and scalability through event-driven processing, co-localized memory-computation structures, and adaptive learning mechanisms. With the current technological innovations still evolving, neuromorphic engineering will change the principles of intelligent computing, which will underpin the sustainable and resilient AI infrastructure in a variety of areas of application.

## 2 Literature Review

Neuromorphic computing originates with the work of Carver Mead who in 1990 officially defined the idea of implementing neurobiological architectures with analogue very-large-scale integration (VLSI) circuits [1]. Instead of symbolic computation and clock-based digital logic, Mead suggested that the engineers should design silicon based systems to simulate the physical laws of the biological neural networks. His initial work showed that the subthreshold analogue circuits would replicate the behaviour of neuronal membranes and synaptic interactions with a very low level of energy consumption. This initial understanding made neuromorphic engineering an interdisciplinary science which has merged neuroscience, electrical engineering, materials science, and computer architecture. Fundamentally, neuromorphic computing aims at imitating the co-location of memory and computation found in biological synapses thus eliminating the von Neumann bottleneck of conventional architectures. In the 1990s and early 2000s, studies in neuromorphic systems shifted away from completely analogue models of neuromorphic systems to hybrid analogue-digital or completely digital models. Analog circuits were more energy efficient, more biological but were sensitive to noise and could not easily be scaled because of fabrication variation. Digital methods, in their turn, were stronger, programmable and compatible with the CMOS fabrication technology. A more viable tradeoff was the mixed-signal implementation, which took the low-power nature of analogue computation but ensured the reliability and scalability of digital control circuits [6]. This development in architecture made the neuromorphic much more applicable to experimental systems of large scale beyond laboratory prototyping. One of the greatest advances in the large-scale neuromorphic hardware was the introduction of the TrueNorth processor by IBM. Merolla et al. [4] indicated that TrueNorth was made of one million programmable spiking neurons and 256 million configurable synapses on a single chip produced by use of 28 nm CMOS technology. TrueNorth can run at just around 70 milliwatts and is therefore extremely energy efficient compared to other traditional processors. The operations of each synaptic use approximately 26 picojoules, which is orders-of-magnitudes lower than the deep learning accelerators of GPUs. The architecture of the chip uses 4,096 neurosynaptic cores that are linked through a packet switched network that is asynchronous. Every core consists of memory and computing locally, using an implementation of a non-von Neumann system where the weights of the synapses are stored in the same

location as the neuron models. This architecture facilitates event based communications i.e. power is used when spikes are sent. Experimental evidence indicated that TrueNorth was capable of doing real-time image recognition experiments with low thermal emissions, which made it ideal in embedded and edge cases. On this development, Intel has announced the Loihi neuromorphic research processor, which pushes the programmability, and on-chip learning capabilities even higher. Loihi is reported by Davies et al. [5] as a manycore processor with a design of about 130,000 neurons and 130 million synapses in the first generation design fabricated on the 14 nm process technology. In contrast to TrueNorth, Loihi uses programmable synaptic plasticity which allows on-chip learning using spike-timing-dependent plasticity (STDP) and other adaptive rules. Experimental appraisals demonstrated that Loihi had the capability of reaching a 1,000 times less energy use than traditional CPU and GPU executions of sparse optimization and constraint satisfaction issues. The processor has hierarchical scaling by mesh interconnects, where many chips can be assembled into huge neuromorphic clusters. The second generation Loihi 2 design goes on to enhance the flexibility of neuron models, inter-chip communication bandwidth, and energy efficiency, which will enhance the scalability of neuromorphic systems to distributed AI systems. Although neuromorphic computing is focused on the hardware innovations, algorithmic models, specifically spiking neural networks (SNNs) are the theoretical basis of the field. Maass [3] proved that the networks of spiking neurons form the third-generation neural network models which can be much more computationally expressive than the traditional perceptron-based models. The spikes allow the encoding of information in the exact timing they occur, unlike rate-coded artificial neural networks (ANNs), and this means that the SNNs can have temporal dynamics closer to those of biological neural processing. This time coding also has the inherent benefits to work with sequential and sensory streams of data. Moreover, SNNs are asynchronous, that is, neurons do not update their states, but do it whenever some events happen. The resulting event based paradigm saves a lot of redundant computation and also consumes very little energy, particularly when dealing with sparse data. Plasticity rules that are biologically inspired frequently control learning in SNNs. Spike-timing-dependent plasticity (STDP) was experimentally confirmed by Bi and Poo [7] who showed that the strength of a synaptic connection between two neurons is determined by the temporal relationship between the pre- and post-synaptic spikes. In case a

presynaptic spike comes before a postsynaptic spike within a critical window, the synaptic strength is reinforced, otherwise, the synaptic strength is diminished. STDP allows unsupervised learning and adaptive behaviours, which offers a way of learning continuously without the need to have gradient backpropagation that is centrally located. In neuromorphic hardware, local implementation of STDP in synaptic networks makes the implementation more scalable and energy efficient as it does not require global weight updates. Irrespective of such developments, it is difficult to train large-scale SNNs. The standard backpropagation algorithms are based on the use of differentiable activation functions, but spiking neurons have discontinuous spikes. To solve this problem, scholars have come up with surrogate gradient techniques and ANN-to-SNN conversion techniques which are used to estimate gradients when training. These methods are usually inaccurate trade-offs to deep convolutional neural networks (CNNs), though effective. Therefore, the current studies aim at coming up with training frameworks that ensure that they are biologically plausible and can deliver competitive performance standards. Recent systematic reviews focus on the increased injection of the neuromorphic systems into edge AI and IoT ecosystems. Indiveri and Liu [6] claimed that neuromorphic processors can be applied especially well to ultra-low-power embedded intelligence, in which real-time processing and energy efficiency are the key considerations. During the implementation of IoT, it is prone to latency and privacy issues since raw sensor data is being transmitted to the cloud servers. Localized inference and anomaly detection can be done on neuromorphic edge devices which use less bandwidth and provide better data security. As an illustration, the event-based vision sensors with neuromorphic processors can be used to track objects at high speed using considerably lower power consumption than the frame-based vision systems. Other opportunities include memristive and resistive memory devices to be used in artificial synapses. Memristors also show resistance change depending on past current flow which is more similar to the weight change of synapses. Embedding the arrays of memristive with the neuromorphic circuits allows in-memory computing, which saves on the cost of energy needed to move data. Research has been found to indicate that crossbar-based memristive arrays are able to execute matrix-vector multiplications, which is the fundamental algorithm of neural networks, parallel and with low latency. The variability of devices and the lack of endurance is however an impediment to commercialization. Neuromorphic systems have high potential of distributed deployment

in terms of scalability. Multi-chip neuromorphic clusters enable the scale of millions of neurons to be simulated together to provide large-scale cognitive modelling and real-time adaptive control systems. Experimental results on distributed clusters of Loihi show a scaling efficiency of almost linear on selected workloads, and this implies that neuromorphic data centres of large scale may be possible in the future. However, the connection with general AI systems like TensorFlow and PyTorch is still low, which points to the necessity of standardized software systems. The strongest argument of neuromorphic computing is that of energy efficiency. It is estimated that human brain is able to carry out about 1015 operations per second consuming about 20 watts of power. In comparison, mega-watt-hours of electricity can be spent to train a state-of-the-art deep neural network. Neuromorphic processors reduce this performance disparity by taking advantage of event-based sparsity and localized computing. Neuromorphic inference tasks have been empirically compared to make less than 1% of the energy demands of the sparse workloads of the corresponding tasks done on a GPU [5]. This efficiency is especially important since AI is becoming battery-powered and sustainable computing infrastructures. Although these benefits are present, there are still great limitations. To begin with, there is no scalable methodologies of SNNs training that can compete with deep learning. Second, analogue and memristive devices are fabricated, which will be variably and this can impact reliability. Third, the benchmarking of standards of neuromorphic performance is not consistent making it difficult to compare across platforms. Lastly, the use of large scale programming needs to be facilitated through the use of readily available programming models and developer tools that hide the complexity of the hardware. To conclude, neuromorphic computing has developed on the further ideas of Mead as analogue VLSI into a high-end platform of digital, analogue and mixed-signal models that can model millions of spiking neurons. TrueNorth and Loihi hardware platforms are energy-efficient and scalable unlike any other hardware, and SNNs and STDP algorithmic innovations make adaptive and event-driven processing possible. It is integrated with edge AI and IoT systems, which highlights the usefulness of neuromorphic engineering to sustainable and distributed intelligence. The difficulties in training, standardization and reliability of the devices have not been solved, but research in the interdisciplinary fields is going on to perfect the engineering principles required to make the brain- inspired computational systems more reliable.

6.  
7.

## 3 Methodology

### 3.1 Review Design

This study was reconstructed as a structured, source-audited literature review and comparative engineering analysis. This is important because the literature contains measurements obtained from different hardware architectures, workloads, software stacks, process technologies, and evaluation procedures. Combining such measurements into a single mean or percentage without harmonization would create a false impression of statistical comparability.

The review therefore used three evidence categories:

1. Architectural specifications, including neuron and synapse capacity, core organization, process technology, and learning mechanisms.
2. Experimental performance measurements, including energy, latency, throughput, and task-level efficiency where the measurement conditions were explicitly reported.
3. Benchmarking and review literature, used to interpret differences among architectures and identify methodological limitations.

### 3.2 Source Identification

The literature base was centered on peer-reviewed research papers, authoritative technical publications, and primary hardware documentation. Foundational sources included Mead [1], Maass [3], Bi and Poo [7], Merolla et al. [4], Indiveri and Liu [6], Davies et al. [5], and Davies et al. [8]. More recent evidence was incorporated from benchmark and emerging-device research, including the NeuroBench framework and recent work on neuromorphic materials and devices.

The search and verification process prioritized sources for which bibliographic information, technical specifications, DOI information, or authoritative institutional documentation could be independently checked.

### 3.3 Inclusion Criteria

Sources were included when they addressed one or more of the following:

- neuromorphic processor architecture;
- spiking neural networks;
- synaptic plasticity or on-chip learning;
- neuromorphic hardware implementation;
- energy or latency evaluation;
- scalability;
- edge or event-driven intelligence;

8. emerging neuromorphic devices;
9. benchmarking of neuromorphic systems.

Sources were excluded from quantitative comparison when their numerical measurements could not be traced to a clearly identified workload, platform, metric, or experimental procedure.

### 3.4 Data Extraction

For each representative platform, the review extracted:

1. platform name;
2. year;
3. architecture type;
4. neuron capacity;
5. synapse capacity;
6. learning capability;
7. reported process technology where available;
8. power or energy metric;
9. workload or application;
10. measurement type;
11. source.

The extraction procedure deliberately avoided converting unlike metrics into a common average. For example, energy per synaptic operation was not treated as equivalent to energy per inference, and maximum system power was not interpreted as energy per task.

### 3.5 Comparative Analysis

The principal analytical method was descriptive comparison. Numerical values were reported only when they could be connected to a documented source. Comparisons were grouped into architecture, energy, scalability, edge intelligence, learning, and engineering challenges.

The review did not calculate a pooled effect size because the available evidence does not satisfy the assumptions required for a meaningful meta-analysis. In particular, workloads, hardware configurations, measurement boundaries, input representations, model sizes, and software stacks differ substantially.

This methodological decision is consistent with the motivation behind NeuroBench, which identifies the lack of standardized neuromorphic benchmarks as a major challenge and separates hardware-independent algorithm evaluation from deployed-system evaluation.

## 4 Results

### 4.1 Energy Efficiency

Based on the literature reviewed, it can be concluded that for certain types of sparse workloads and event-driven workloads, there are significant energy-efficiency benefits to be gained with the use of neuromorphic systems. But the degree of the benefit is not necessarily common to all neuromorphic computing. TrueNorth is an excellent example of such a location. Merolla et al. [4] announced a million-neuron digital neurosynaptic processor that used around 70 mW power under the conditions of the experiment. Its architecture distributes memory and computation throughout the neurosynaptic cores, and communicates using event-driven packets. There's another example called Loihi. Davies et al. [5] showed an on-chip learning-enabled many-core architecture and showed significant energy savings for specific sparse optimization/learning workloads. The "up to 1,000-fold" comparison is meant to convey a comparison of workloads, not a ratio of Loihi to all CPUs or GPUs. The principle engineering reason for these benefits is conditional sparsity. In an event-driven system, the system could choose not to execute operations on inactive inputs if there is little activity on them. Conventional accelerators, on the other hand, can continue to work well in workloads for which their hardware and software ecosystem is well optimized for dense matrices.

The importance of network activity and neuromorphic circuit implementation on energy consumption in SNNs has also been highlighted by Sorbaro et al. [9]. The question then isn't whether it is always better to use neuromorphic computing, but when. So it's not so much whether it is always better to use neuromorphic computing, but when. To make the comparison, it must distinguish between power and energy. Power is measured in watts, energy is dependent upon the length and type of computation. A processor that has a low nominal power rating may not be using less power doing a specific operation, if that operation requires a long computation time.

### 4.2 Scalability

There are multiple measures of scalability: Number of Neurons per chip, number of synapses per chip, number of chips, communication bandwidth, memory, and system throughput. A single chip of TrueNorth was used to show integration of one million neurons [4]. Loihi 2 expanded the number of neurons to around 1 million per chip and has been programmed to work with up to 120 million synapses, thereby boosting Loihi's flexibility. [10] Hala Point is an additional leap forward in system-level scaling. The 1.15 billion neurons and 128 billion synapses are spread across a total of 1,152 Loihi 2 processors, Intel

says. It has 140,544 neuromorphic processing cores and consumes a max of 2,600W. [11] These numbers reflect engineering achievements, but don't necessarily mean that we can talk of computational equivalence to the human brain by mere counting of neurons. The number of synapses does not directly reflect cognitive ability and a biological neuron and an artificial neuron are not alike. Thus, only scale comparisons can be made with biological brains and not the equivalent of performance. There is a scaling problem there as well, related to communication. More NEPs make the bandwidth, routing, synchronization, memory access, thermal management and fault tolerance of the interconnectivity more critically. So, going forward, it will not be only about transistor density; system stack improvements will be important in achieving scaling.

### 4.3 Real-Time Edge Intelligence

Because the information from sensors in edge applications is often temporal, sparse, and always changing, neuromorphic computing will be especially applicable. Event-based vision is an example. Event-based cameras differ from traditional cameras in that they capture a frame when the individual pixels around the event change. If there is only a small number of changes in a scene, then such representations can help to minimize redundant data. These features are well suited to such architectures, as SNNs can process temporal events without having to convert all inputs into a dense frame representation. These can be used in gesture recognition, robot perception, autonomous navigation, auditory processing, and always on sensing. But, it's not wise to make claims for universal latency superiority. Latency is determined by the entire processing chain, from sensing and data pre-processing, through data transfer, neural computation, post-processing and onward to the communication with the host system.

It was therefore a problem to compare 164.5 ms and 250 ms, as a representative example in one experiment would be used as evidence for a general neuromorphic advantage. In those two instances, the difference between the two measurements is about 34.2%, which is calculated by subtracting the two.

$$\frac{250 - 164.5}{250} \times 100 \approx 34.2\%$$

The correct reading is that under the reported experimental conditions the reported neuromorphic implementation had about 34.2% lower latency. It must not be offered as a universal latency reduction for neuromorphic systems. The broader literature lends credence to the potential benefits of real-time processing with neuromorphic architectures but system-level benchmarking is still necessary. To

address this, NeuroBench explicitly demands evaluation and reporting at the task level and reports the methodology used for system measurements.

### 4.4 Learning, Programmability and Engineering Challenges

The evidence that was reviewed shows that there are some challenges which are hindering larger scale uptake.

Large-scale SNN training is still challenging. Conventional gradient based optimization is complicated because of the discontinuous events produced by spiking neurons. Surrogate gradients and ANN-to-SNN conversion have been shown to increase the practicality of training, but pose trade-offs in terms of computational cost, temporal accuracy, conversion quality, and the complexity of the model.

Second, ecosystems for programming are not as developed as those of the mainstream deep-learning frameworks. Most of the time, there are specific considerations: mapping, models of neurons, event representations, and hardware considerations that need to be addressed for neuromorphic hardware. Software frameworks like Lava, Nengo and other SNN environments have made software more approachable, yet cross-platform interoperability is not yet an engineering solution but a goal to be achieved.

Thirdly, there are variability and reliability issues in emerging-device implementations. While memristive and related devices offer potentially desirable physical mechanisms for synaptic storage and computation, device-to-device variation, endurance, retention, noise, and peripheral-circuit overhead need to be dealt with. The latest study is still exploring ways to control or exploit stochasticity of devices in neuromorphic systems.

Fourth, benchmarking still is a big issue. The development of NeuroBench was exactly motivated by the fact that the tasks and measurement procedures used in neuromorphic studies have been heterogeneous so far. It provides a framework for distinguishing algorithm level evaluation from deployed system evaluation and the need for contextual information on architecture, algorithms, workflow and measurement procedures.

Last but not least, there is an integration with the existing AI system. In practice, neuromorphic processors could work alongside CPUs, GPUs, traditional accelerators, sensors and communications to achieve results. Therefore, hybrid architectures could be a realistic approach to incremental adoption, instead of replacing conventional computing with a

completely new approach to computing: neuromorphic computing.

## 5 Discussion

From the reconstructed evidence, it appears that neuromorphic computing is not a one-size-fits-all solution to replace traditional AI accelerators. The key benefits of this are attributed to its architecture that incorporates distributed computation, close memory to processor, asynchronicity in communication and sparse event-driven activity [1], [6]. These qualities are especially important for some workloads that have temporal and spatial sparse data, such as sensory processing, robotics, optimisation and edge intelligence. Spiking neural networks (SNNs) also facilitate this way of working by encoding information in the spike trains and temporal evolution of the spikes instead of the traditional continuous valued activations [3]. The evolution of TrueNorth and Loihi has been a demonstration of key phases in the evolution of neuromorphic hardware. TrueNorth's distributed architecture for event-driven communication consisted of a million programmable spiking neurons and 256 million synapses, a large-scale digital integration. The event-driven communication, distributed architecture used by TrueNorth, included a million programmable spiking neurons and 256 million synapses [4]. Loihi took this approach further with programmable on-chip learning and synaptic plasticity, and increased flexibility for adaptive neuromorphic applications [5]. Future developments of Loihi and related systems have focused on expanding the application opportunities, scaling up, enabling neuron models and communication between different chips. [8]. The advances are indicative of the progress in neuromorphic engineering, which is moving beyond the specialized circuit level implementations into increasingly programmable and scalable computing systems.

Care must be taken in interpreting the evidence on energy efficiency. Event-driven architectures can benefit from significant energy savings for certain sparse workloads as they can skip unnecessary computation when activity is limited [5], [9]. The results presented here should not be extrapolated to a general efficiency comparison between neuromorphic and GPUs, however. Energy performance depends on the task, the model, the representation of the input, the level of the activity, the software implemented and the measurement boundary and the hardware configuration. The SNN energy consumption is also closely dependent on the characteristics of the SNN implementation and its activity, as previously pointed

out by Sorbaro et al. [9]. Thus, it is important to make a distinction between metrics like power, energy per operation, energy per inference, and throughput per watt, and avoid using them as synonyms. Similarly, for latency, the entire computational pipeline should be considered. Neuromorphic architectures could be beneficial under conditions of low temporal density and when processing the temporal events can be done directly on the event-driven representations. When applications need to support large amounts of pre-processing, dense inputs, translation between frame based and event based representations, and communication to conventional processors, however, the potential benefits of the group may be diminished. This further highlights the need to evaluate the workload itself instead of blanket statements on the advantages of neuromorphic capabilities.

Scalability is yet another critical factor to take into account. The first is the integration at the chip level (TrueNorth), and the second is the programmable learning capabilities followed by the addition of capacity and flexibility over the generations (Loihi) [4], [5], [8]. The development of large-scale systems like that of Hala Point is another step in the direction of linking multiple neuromorphic processors together into a system. However, actual numbers of neurons in the artificial network cannot be compared with numbers of neurons in the biological network. Biological neural systems are very heterogeneous with respect to their neurons, synaptic structures, connectivity patterns, biochemical mechanisms, and their learning processes. Thus, at the system level, the number of neurons and synapses in a system may be an appropriate engineering metric of scale, but is not a measure of system functionality. Understanding the biological bases of neuromorphic engineering, therefore, must be interpreted as an architectural influence instead of a biological mimicry of cognitive processes. The challenges of learning continue to be at the heart of engineering. An important biologically inspired mechanism for modifying the strength of a synapse is the timing relationship between the pre- and postsynaptic spikes in the activity of the two cells—STDP [7]. These local learning mechanisms are appealing since they allow the adaptation to be achieved without central weight updates. However, their training of large-scale deep SNNs is still difficult due to the discontinuous nature of the spike generation and direct gradient-based optimization is not suitable. This rapid progress in programmable neuromorphic processors and learning mechanisms therefore requires algorithms and software frameworks to be improved as well as training algorithms [8]. Another roadmap for future neuromorphic development is emerging materials. Memristive and low-dimensional

devices can enable physical integration of storage and computation, which can help minimize data transfers and enable artificial synaptic functions (Ahmed et al., 2023). But other engineering factors such as device variability, reliability, endurance, retention, and noise are also significant factors. Recent studies also indicate that stochasticity of memristive synapses can be used as a computational resource instead of just considered as a problem [12]. This suggests that future neuromorphic systems could benefit from devices which explicitly model physical device properties in computational models.

Another significant problem stated in the reconstructed review is the issue of benchmarking. Direct numerical comparisons of neuromorphic studies are often difficult because of the evaluation of different tasks, models, hardware configurations, software pipelines and measurement boundaries. The NeuroBench framework tackles this by separating between algorithm-level and deployed-system evaluation, and prioritizing the architecture, algorithm, workflow and measurement methodology [13]. This will give a more solid base for comparative studies than summing together non homogeneous numbers of energy, latency or throughput. Combining all these facts together, it can be concluded that there is a need for multi-level co-design of the devices, circuits, architectures, algorithms, software, sensors and benchmarking in order to achieve progress in the field of neuromorphic engineering. The evolution of the field from the neuromorphic principles of Mead [1] to the SNN and plasticity studies of Maass [3], Bi & Poo [7] and on to the large-scale platforms of TrueNorth and Loihi [4], [5] illustrates the interdisciplinary aspect of the research. More recent studies on energy optimization, novel materials, memristive stochasticity, and standardized benchmarking to this end reinforce the need for system level integration for further advances [9]; (Ahmed et al., 2023); [12], [13]. Neuromorphic computing therefore is best suited as an additional paradigm for computing that is best utilized in scenarios where sparse temporal information, adaptation, low latency and energy constraints are critical.

## 6 Conclusion

Neuromorphic computing is a research area of great significance for brain-inspired and event-driven information processing. Distributed computation, close integration of memory and processing, asynchronous communication, sparse activity, spiking neural networks and, in some systems, local learning mechanisms are some of its main engineering

features. The results in this paper suggest that for certain workloads, especially those having sparse temporal input and sparse internal computation, a neuromorphic system has significant potential for reducing energy and latency. However, these benefits are dependent on the workload and shouldn't be taken as a blanket statement of superior performance to CPUs or GPUs. This evolution from large-scale neurosynaptic integration to programmable learning and multi-chip neuromorphic systems is illustrated through the development of TrueNorth, Loihi, Loihi 2 and Hala Point. The reported 1.15 billion neurons and 128 billion synapses at Hala Point, and the reported maximum system level power of 2,600 W illustrate the growing power of engineered neuromorphic systems, and the need to consider scale alongside the system-level requirements for energy and communication. The review also highlights the need for clearly defined workload, measurement boundary, hardware configuration, evaluation procedure, software, and model in quantitative claims of neuromorphic research. The metrics of energy per synaptic operation, energy per inference, total system energy, throughput per watt, and latency do not relate directly to each other, and thus should not be confused.

Moving forward, advancements in SNN training, software portability, hardware reliability, interconnects, new memory technologies, benchmark standardization and integration with legacy AI systems will be critical. The NeuroBench initiative offers valuable underpinning to future evaluations for greater comparability and transparency. Therefore, neuromorphic computing should not be seen as a solution to replace traditional computing or simply as a biological copy of the brain. Its engineering significance is in taking advantage of architectural features which are specifically suitable for sparse, temporal, adaptive, and resource-constrained computation.

## Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

## References:

- [1] Mead, C. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE*, 78(10), 1629–1636. <https://doi.org/10.1109/5.58356>
- [2] Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, Article 31. <https://doi.org/10.3389/neuro.09.031.2009>
- [3] Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9), 1659–1671. [https://doi.org/10.1016/S0893-6080\(97\)00011-7](https://doi.org/10.1016/S0893-6080(97)00011-7)
- [4] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., Jackson, B. L., Modha, D. S., & others. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197), 668–673. <https://doi.org/10.1126/science.1254642>
- [5] Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G. D., Joshi, P., Imam, N., Jain, S., Liao, Y., Lin, C.-K., Lines, A., Liu, R., Mathiakutty, D., McCoy, S., Paul, A., Tse, J., Venkataramanan, G., Weng, Y.-H., Wild, A., Yang, Y., & Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1), 82–99. <https://doi.org/10.1109/MM.2018.11213035>
- [6] Indiveri, G., & Liu, S. C. (2015). Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, 103(8), 1379–1397. <https://doi.org/10.1109/JPROC.2015.2444094>
- [7] Bi, G. Q., & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24), 10464–10472. <https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998>
- [8] Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Fonseca Guerra, G. A., 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>
- [9] Sorbaro, M., Liu, Q., Bortone, M., & Sheik, S. (2020). Optimizing the energy consumption of spiking neural networks for neuromorphic hardware. *Frontiers in Neuroscience*, 14, Article 662. <https://doi.org/10.3389/fnins.2020.00662>
- [10] Intel Labs. (2021). Taking neuromorphic computing to the next level with Loihi 2: Technology brief. Intel Corporation. <https://download.intel.com/newsroom/2021/new-technologies/neuromorphic-computing-loihi-2-brief.pdf>
- [11] Intel. (2024, April 17). Intel builds world's largest neuromorphic system to enable more sustainable AI. Intel Newsroom. <https://www.intel.com/content/www/us/en/newsroom/news/intel-builds-worlds-largest-neuromorphic-system.html>
- [12] Das, H., Patel, K. P., Febbo, R. D., Schuman, C. D., & Rose, G. S. (2025). Leveraging stochasticity in memristive synapses for efficient and reliable neuromorphic systems. *npj Unconventional Computing*, 2, Article 3. <https://doi.org/10.1038/s44335-024-00017-x>
- [13] Yik, J., Van den Berghe, K., den Blanken, D., Bouhadjar, Y., Fabre, M., Hueber, P., Ke, W., Khoei, M. A., Kleyko, D., Pacik-Nelson, N., Pierro, A., Stratmann, P., Sun, P. S. V., Tang, G., Wang, S., Zhou, B., Ahmed, S. H., Vathakkattil Joseph, G., Leto, B., ... Reddi, V. J. (2025). The neurobench framework for benchmarking neuromorphic computing algorithms and systems. *Nature Communications*, 16(1), Article 1545. <https://doi.org/10.1038/s41467-025-56739-4>

### **Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)**

The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

### **Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself**

No funding was received for conducting this study.

### **Conflict of Interest**

The author has no conflicts of interest to declare that are relevant to the content of this article.

### **Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)**

This article is published under the terms of the Creative Commons Attribution License 4.0 [https://creativecommons.org/licenses/by/4.0/deed.en\\_US](https://creativecommons.org/licenses/by/4.0/deed.en_US)