From CPUs and GPUs to Neuromorphic Chips: A Paradigm Shift in Computing Architecture

Dr.Manasi M. Patil

Department of Artificial Intelligence and Machine Learning, Walchand College of Engineering , Sangli, India

Abstract

The exponential growth of data-intensive applications, especially in artificial intelligence (AI) and machine learning (ML), has pushed conventional computing architectures—CPUs and GPUs—to their limits in terms of power efficiency, parallelism, and scalability. Neuromorphic chips, inspired by the human brain's structure and function, promise to overcome these limitations by enabling event-driven computation, massive parallelism, and ultra-low energy consumption. Continuous comparative framework that highlights performance power, and programmability metrics relative to traditional processors. Each neuromorphic system embodies divergent trade-offs. For instance, TrueNorth's 4,096-core fabric excels in extreme low power for inference but requires an explicit mapping of neural topologies. Intel's Loihi balances programmability and efficiency through on-chip learning but is constrained by fixed memory hierarchies. BrainChip's Akida, leveraging a unified memory and custom instruction set, targets real-time sensor fusion in low-footprint environments. By quantifying these metrics, we identify specific problem classes that merit a transition to neuromorphic computing, including spatiotemporal signal processing, anomaly detection in sensor networks, and continual learning in resource-constrained devices.

Beyond hardware, also surveyed the software models that translate conventional neural networks into spiking representations. Case studies illustrate the accelerated inference times and reduced energy footprints achieved on neuromorphic processors for real-world vision, speech, and robotic control tasks. Finally, we discuss the anticipated scalability of neuromorphic systems, arguing that future heterogeneous deployments will integrate these chips within mixed architectures to offload specific workloads, thereby extending the performance and efficiency spectrum of computing infrastructure beyond the limits of traditional CPUs and GPUs.

Keywords: AI, Computing, CPU, GPU, Neuromorphic

1. Introduction

Over decades, the landscape of computing architectures has determined the pace of innovation, influencing everything from climate modelling to the latest AI breakthroughs. Central Processing Units have long provided the backbone of computing, well-suited to the orderly, stepwise logic of most programs. Yet as data volumes surged, particularly in fields like machine learning and simulations of complex phenomena, the industry turned to Graphics Processing Units, whose thousands of cores could tackle multiple data elements simultaneously. Together, the CPU and GPU, still arranged around the classic von Neumann principle of memory and processing separation, have met the challenges of ever-more demanding workloads without breaking stride.

However, the fast-growing amount of data and the increasing complexity of the AI models have shown the limits of traditional hardware. Even though CPUs and GPUs are powerful, they

have some natural problems like poor energy use, limited ability to do many tasks at once, and memory bandwidth issues from the von Neumann architecture, often called the "von Neumann bottleneck. As making smaller transistors becomes harder and Moore's Law slows down, these issues have gotten worse, especially in areas that need real-time, low-power, and event-based processing, like edge AI, autonomous robots, and IoT devices.

Neuromorphic computing has become a big change to solve these problems by looking at how the human brain works. Unlike regular processors, which separate memory and computation, neuromorphic chips combine these parts, copying how biological neurons work in parallel and distributed ways. This lets neuromorphic systems be very energy-efficient and highly responsive. Spiking Neural Networks (SNNs), which use sudden spikes to send information instead of continuous signals, are central to neuromorphic designs, allowing for asynchronous and event-based computing.Interest in neuromorphic computing has led to new hardware like IBM's TrueNorth, Intel's Loihi, BrainChip's Akida, and SpiNNaker.These chips are built to handle AI tasks more efficiently than traditional CPUs and GPUs. They not only use much less power but also open up new opportunities for brain-inspired computing, like neuromorphic vision sensors and advanced robotics.This paper gives a full look at the shift from CPUs and GPUs to neuromorphic chips.It starts with a review of the history of computing, showing how it moved from doing one task at a time to doing many tasks at once. Then, it explores the shortcomings of current CPUs and GPUs and why they can't keep up with modern AI needs. The main part of the paper explains neuromorphic architecture, its design ideas, major examples, and performance benefits. Finally, it covers possible uses, challenges, and future steps in neuromorphic computing, showing why this change is important for the next big step in computing, IMARC Group(2025), Transparency Market Research(2025), SNS Insider(2023)[1-3].

2. Literature Review

Neuromorphic computing, which is inspired by the structure and function of the human brain, has made big progress lately, especially with the use of photonic technologies and new designs. The following review covers important findings from recent research. Biasi et al.(2023) [4], look into the potential of photonic neural networks (PNNs) made using integrated silicon microresonators. Their work points out the benefits of using photonic systems, such as faster processing, the ability to handle multiple tasks at once, and lower power use compared to traditional electronic systems. The authors show experimental results of photonic neuron-like behavior and talk about how silicon photonics can be scaled up for neuromorphic computing. They also suggest that using photonics could help solve the problem of slow electronic connections, making it a great option for fast AI hardware. Li et al.(2023)[5], give a detailed overview of how photonics can be used in neuromorphic computing. They focus on the basics, the types of devices used, and future possibilities. The paper describes various photonic devices like microring resonators, Mach-Zehnder interferometers, and phase-change materials that can imitate the behavior of synapses and neurons. They also discuss how combining photonic and electronic systems can help with issues like energy use and bandwidth. This review shows that photonics could be a big part of creating next-gen neuromorphic processors that work better than the traditional von Neumann architecture. Greatorex et al.(2024)[6], introduce TEXEL, a neuromorphic processor that can learn directly on the chip, which makes it suitable for use with next-gen technologies beyond standard CMOS. Unlike traditional AI accelerators, TEXEL allows learning inside the hardware, which cuts down on delays and energy use. The authors talk about the principles of combining device design with circuit design and highlight how TEXEL works well with new devices like memristors and spintronic elements. This study helps in building highly adaptable and energy-efficient neuromorphic systems that are similar to how the brain works. Le D et al. (2025)[7], review the memory wall problem in neuromorphic computing, which is a major issue that limits the performance of hardware accelerators. The

paper explains the main causes of this problem, like the extra time and resources needed to move data between memory and processing parts. The authors assess new memory technologies such as RRAM, PCM, and FeFET, as well as memory-focused computing approaches that aim to fix this bottleneck. Their review shows the importance of designing systems where memory and computation are closely linked, as well as using 3D integration to improve overall performance.

2.1 Summary and Research Gap

Overall, these works show that the future of high-performance and energy-efficient AI hardware lies in combining photonics, new memory technologies, and neuromorphic designs. Biasi S, et al.(2023) and Li R, et al. (2023) [4-5], Photonic approaches offer speed and better data handling, Greatorex H, et al. (2024)[6], systems like TEXEL, show the ability to adapt on the chip, and memory-focused solutions, Le D, et al.(2025)[7], tackle the memory wall challenge. However, there is still a lack of a unified system that smoothly combines these parts—photonic neuromorphic cores, learning mechanisms that can adapt, and efficient memory systems. This is a big area for future research.

3. Limitations of CPUs and GPUs

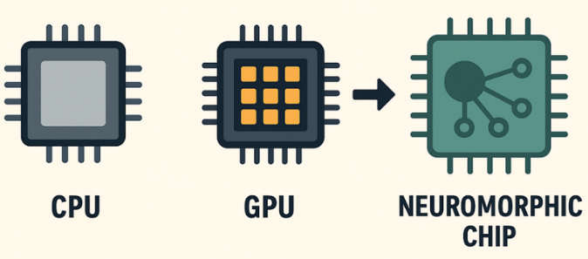

Figure 1: CPU-GPU-Neuromorphic Chip

As in Fig.1, Central Processing Units (CPUs) and Graphics Processing Units (GPUs) have been the main tools in modern computing for a long time. CPUs are designed to handle a wide variety of tasks one after another, while GPUs are built to handle many tasks at the same time. This makes GPUs especially good for demanding jobs like machine learning and big simulations. However, even with these improvements, both types of processors still have important limits that make it hard for them to keep up with the more complex, data-heavy, and energy-conscious tasks that are coming in the future, Intel(2025), BrainChip (2025),[8-9].

3.1 The Von Neumann Bottleneck

Both CPUs and GPUs use the von Neumann architecture, which keeps memory and the processing part separate. Because of this, data has to keep moving back and forth between memory and the processor. As tasks that need a lot of data become more common, this constant movement of data creates a performance slowdown. Moving data so often not only makes computations slower but also uses a lot of energy. This problem gets worse in AI and deep learning, where huge amounts of data must be handled quickly. Even with better memory solutions like High Bandwidth Memory (HBM), the energy and time needed to move data remain a big challenge.

3.2 Power Consumption and Thermal Constraints

CPUs and GPUs now use more power because they need faster speeds, more processing units, and bigger memory systems. Top-tier GPUs can use hundreds of watts, creating a lot of heat and needing advanced cooling systems. This high power use makes them not good for edge devices, self-driving cars, and other uses where saving energy and quick performance are important. As the parts on chips get smaller, problems like wasted electricity and heat management get worse, making the usual ways of improving chip performance less effective.

3.3 Diminishing Returns of Parallelism

GPUs have changed how we do parallel computing by providing thousands of cores that can work at the same time. However, their performance is still limited by how fast data can be

moved in and out of memory and the extra work needed to handle multiple tasks at once. Not every program can be easily made to work with GPUs, and trying to force these programs to use GPU power often doesn't work well. Also, as neural networks get more complicated, just adding more GPU cores isn't enough to handle the demands of training and running these models on a large scale.

3.4 Latency in Real-Time Processing

Real-time uses like robots, self-driving cars, and edge AI need super fast responses. Regular CPU and GPU setups, which are made for handling big batches of tasks, don't work well for these needs. The time it takes to move data between the parts that do the work and the storage, along with the extra energy used when cores are sitting idle, makes these setups not great for tasks that react to events and need quick, smart responses.

3.5 Inefficiency for Sparse and Unstructured Data

AI tasks, especially ones that involve understanding human language, handling sensory information, or working with graph-based learning, usually deal with data that is spread out and not uniform. CPUs and GPUs are best at handling dense and organized calculations, like multiplying matrices. But when they have to process sparse data, they end up using more resources than needed. Neuromorphic chips, on the other hand, only process the active parts of data, which makes them more efficient for these kinds of AI tasks.

3.6 Physical Scaling Limits

The end of Dennard scaling and the slowdown of Moore's Law have made it harder to boost CPU and GPU performance by just making transistors smaller. Chip makers have tried using multi-core designs and better lithography, but the improvements from these methods are only small. To keep moving forward, we need to completely rethink how computer hardware is built, not just keep making the same designs smaller. The above issues show why we need new ways of computing, Theis TN, Wong HSP(2017)[20].

Neuromorphic computing helps with many of these problems by copying how the brain works—processing information in a way that's driven by events and happens in parallel. This makes it much more energy-efficient and flexible, Innatera(2025),KAIST(2024) ,Mead C(1990),Davies M, et al.(2018), Merolla PA, et al. (2014),15.Akopyan F, et al. (2015),[10-15].

4. Neuromorphic Computing Architecture

4.1 Principles of Neuromorphic Computing

Neuromorphic computing builds on Spiking Neural Networks (SNNs), which replicates biological neuron functioning more faithfully than standard Artificial Neural Networks (ANNs). Neurons in SNNs exchange discrete spikes—short electrical impulses—and computing happens exclusively during spike emission. Because the processing is triggered by the sparsity of spikes, the approach cuts power dramatically; dormant neurons stay silent, in contrast to GPUs that push uniform layer calculations whether the data is active or not.

A foundational design idea is merging memory with computation to sidestep the von Neumann bottleneck. Neuromorphic chips assign each artificial neuron its own local store for synaptic weights, mirroring the way living neurons hold their connections. By keeping memory and processing in the same place, these chips can run countless operations at once while keeping delays to an absolute minimum.

Asynchronous operation is another core trait. Where CPUs and GPUs depend on a single ticking clock to coordinate work, neuromorphic circuits lack any such master timing signal. Neurons update their states only when they receive pertinent input, enabling them to proceed on their own schedules and respond to changes as soon as they matter, Furber SB, et al, (2014)[16].

4.2 Hardware Design and Architecture

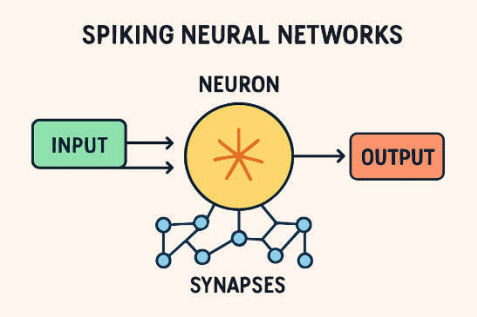

Figure 2. Spiking Neural Network

As shown in Fig.2 and Fig.3, at the core Neuromorphic chips use spiking neural networks (SNNs) which are more energy efficient and computationally powerful network. Whereas ANNs have structured layers, process information using fixed continuous values. SNNs using discrete, time-dependent spikes for information and processing work similar to biological neural systems.. Each of these cores has artificial neurons and synapses. These cores are connected using communication methods that imitate how nerves work in the brain.

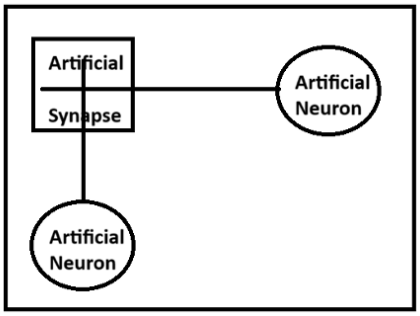

Figure 3. Basic Unit of Neuromorphic Chip

The major components are:- Artificial Neurons: These create signals, or spikes, when they receive enough input, just like real neurons fire when they get enough signals. Artificial Synapses: These store information about connections and help the system learn by changing the strength of connections, similar to how learning happens in the brain through mechanisms like Spike-Timing Dependent Plasticity (STDP).- Crossbar Arrays: These are special structures that help store and process information efficiently.They often use new types of technology like memristors to do this.- Event-Driven Interconnects: These are ways the system sends signals between cores, but they only send data when needed, which helps save energy.The main goal of neuromorphic chips is to process information in a way that is efficient, parallel, and adaptable, rather than focusing on how fast they can perform basic calculations, Sebastian A et al.(2020)[17].

4. 3 Leading Neuromorphic Chips

Many research projects and companies have made big progress in neuromorphic computing, creating impressive chips:IBM TrueNorth :This chip was introduced in 2014 and is one of the first large-scale neuromorphic processors.It has 1 million programmable neurons and 256 million synapses spread across 4,096 cores. TrueNorth works based on events, not continuous processing, and uses around 70 mW of power during regular tasks—much less than traditional GPUs doing similar jobs. It is especially good for vision and pattern recognition tasks that need very little power. Intel Loihi :Released in 2018, Loihi is a big improvement with 128 neuromorphic cores and 130,000 neurons.It supports learning directly on the chip using STDP and reinforcement learning, which means it can learn without needing external help. Its newer version, Loihi 2, from 2021, has better programming options, more detailed neuron models, and improved support for AI tasks that use very little energy.BrainChip Akida:Akida is a

processor made by BrainChip that is designed for real-time AI tasks on the edge.It can run both standard AI models like CNNs and spiking neural networks, which helps developers move from traditional deep learning to neuromorphic computing more easily. Akida is very energy-efficient and is used in smart sensors, robotics, and internet of things devices, Maass W.(1997)[18], .SpiNNaker: This platform was developed at the University of Manchester and is used to simulate large brain models. SpiNNaker can handle up to 1 billion neurons and is mainly used for neuroscience research, not for commercial AI applications.

4. 4 Advantages Over Conventional Processors

Neuromorphic chips do better than CPUs and GPUs in several ways:Energy Efficiency: These chips only process what is needed and reduce memory use, which can lower power use by up to 100 times compared to GPUs for similar tasks.Massive Parallelism: Neurons work at the same time and on their own, which helps process streams of sensory data in real time.Adaptability: They can learn and adjust while working on the chip, making them good for changing environments without needing help from the cloud.Latency: The way these chips work, based on events, allows for fast responses, which is important for robots, self-driving systems, and AI at the edge, Shafique M et al.(2018)[19].

5 Hybrid Architectures

Even though neuromorphic chips are promising, they are often used together with CPUs and GPUs in mixed systems. This setup uses the strengths of traditional processors for tasks like handling big calculations while using neuromorphic parts for efficient, event-driven processing. Neuromorphic computing's design and way of working directly fix the issues with CPUs and GPUs, making it a strong choice for future AI and edge computing systems.

5.1 Comparative Analysis

The shift from CPUs to GPUs and now to neuromorphic chips shows how the needs of computing tasks have changed. While CPUs and GPUs are good for general use and parallel work, neuromorphic chips are better at using less energy, handling events, and adapting. A full comparison of these designs helps understand their strengths, weaknesses, and where neuromorphic systems are especially useful , Indiveri G, Liu SC.(2015), BrainChip Holdings (2021)[21-22].

5.2 Performance Benchmarks

CPUs are built for doing one task after another and are fast at handling single tasks. They have few powerful cores, making them good for general use but less efficient for tasks that need a lot of parallel processing, like deep learning. GPUs have many simpler cores that are great for parallel work, especially handling big matrix calculations.For example, modern GPUs like NVIDIA's A100 can do trillions of operations per second and are widely used in training deep neural networks. However, they use a lot of power, often over 250 to 400 watts for top models, and need advanced cooling systems, NVIDIA. CUDA (2022)[26]. Neuromorphic chips are very efficient because they process data in a way that responds to events as they happen, instead of running instructions all the time. For example, IBM's TrueNorth chip can handle 46 billion synaptic operations per second per watt, and it uses just 70 milliwatts of power—much less than GPU systems doing similar tasks. Intel's Loihi chip is also very efficient, using up to 100 times less energy than GPUs for similar tasks when dealing with sparse data patterns.

Energy efficiency is very important, especially for devices that need to work without being plugged in or for applications that need to run in real time. CPUs are designed for general use, but they use a lot of power when they're not doing anything because they keep running all the time. GPUs are good for parallel tasks but still use a lot of energy because all their parts are

always on. Neuromorphic chips are different because they only use power when neurons are activated by events. For example, Loihi can run speech recognition tasks using less than 20 milliwatts, while GPU-based systems need hundreds of watts. This makes neuromorphic chips great for devices like drones, sensors, and other battery-powered gadgets. CPUs and GPUs are limited by the von Neumann architecture, which separates memory and processing, creating a bottleneck in data transfer. High-performance GPUs use expensive and power-hungry memory to help, but the problem remains.

Neuromorphic chips avoid this by combining memory and processing in each neuron and synapse. This reduces the need to move data around and makes processing faster and more efficient, especially for tasks with event-based data like real-time vision or sensing, Schuman K, et al.(2017)[25] .

5.3 Applications Vision Processing

IBM TrueNorth can process high-resolution visual data and recognize objects using up to 100 times less power than traditional GPU systems, making it perfect for embedded vision systems.

Edge AI: Intel Loihi has been used in robotics for navigation, where it runs inference with very low latency and much less energy than NVIDIA Jetson GPUs.

Speech Recognition: BrainChip Akida is efficient for tasks like detecting keywords and analyzing audio, working well on small devices that can't handle GPU-based AI models, NVIDIA. CUDA (2022)[26] .

5.4 Limitations of Neuromorphic Chips

Despite their advantages, neuromorphic chips are not yet universal replacements for CPUs or GPUs. Current neuromorphic platforms are optimized for spiking neural networks and event-driven workloads but are less effective for dense matrix computations, which remain the strength of GPUs. Moreover, the lack of standardized programming frameworks and the relative immaturity of the software ecosystem present challenges for widespread adoption.

6 Summary of Comparison

Table 1 Summary of Comparison CPU GPU and Neuromorphic Chip

SN	Feature	CPU	GPU	Neuromorphic Chip
1	Processing Type	Sequential	Parallel (SIMD)	Event-driven (SNN)
2	Energy Efficiency	Low	Moderate	Very High
3	Best For	General-purpose tasks	AI training, graphics	Low-power AI inference
4	Latency	High for parallel AI	Moderate	Low
5	Scalability	Limited	High (parallelism)	Extremely high (neurons)

Neuromorphic chips work together with CPUs and GPUs instead of taking their place. They are especially good at tasks that need fast responses, use less power, and can adapt on the fly. Putting them into mixed systems is a great idea for the future of computing.

7. Challenges

There are many challenges to overcome, including technical issues, design limitations, and problems with the overall system and support. In the coming years, neuromorphic chips are likely to play a big role in areas like smart cities, medical devices, AI cameras that work at the edge, and even space missions. They are light and use very little power, which makes them

perfect for situations where power is limited and reliability is key.These chips do more than just speed things up—they help create new kinds of technology that traditional processors can't handle.Because they respond quickly, adjust on the fly, and use minimal energy, they open up possibilities for inventions that wouldn't be possible with regular computer chips, Amir A, et al.(2017), LeCun Y, et al.(2015)[23-24] .

7.1 Software Ecosystem and Programming Complexity

One of the biggest problems for neuromorphic chips is that there isn't a well-developed software environment yet. Most traditional AI tools, like TensorFlow and PyTorch, are made for dense matrix-based neural networks and don't naturally work with Spiking Neural Networks (SNNs). This makes it hard for developers to switch from regular deep learning models to systems that work with event-driven, spike-based processing.

To fix this, some companies like Intel and BrainChip have created special software development kits, such as Intel's NxSDK for Loihi.

But these tools are still in early stages and don't have the same wide range of libraries or strong community support as tools used for CPU and GPU-based AI. Also, training SNNs is tough because spike events aren't differentiable, which makes it hard to use backpropagation—the key technique in modern deep learning. Going forward, the focus should be on creating hybrid frameworks that work smoothly with existing AI systems while allowing efficient use on neuromorphic hardware.

7.2 Hardware Scalability and Manufacturing Challenges

One of the main issues with neuromorphic chips is that there isn't a good software environment yet. Most common AI tools, like TensorFlow and PyTorch, are built for neural networks that use dense matrices and don't really work well with Spiking Neural Networks (SNNs). This makes it difficult for developers to move from regular deep learning models to systems that use event-driven, spike-based processing.

To help with this, some companies like Intel and BrainChip have made special software development kits, such as Intel's NxSDK for Loihi.

However, these tools are still in the early stages and don't have the same variety of libraries or strong community support as tools used for CPU and GPU-based AI.

Also, training SNNs is hard because spike events aren't differentiable, which makes it hard to use backpropagation—a key method in modern deep learning. Looking ahead, the focus should be on making hybrid frameworks that work well with existing AI systems while also being efficient for neuromorphic hardware, Market Research Future(2023–24)[28].

7.3 Benchmarking and Standardization

Unlike CPUs and GPUs, which have well-defined benchmarks (e.g., FLOPS, TOPS, SPEC scores), neuromorphic performance is difficult to measure using conventional metrics. Event-driven processing and adaptive learning mechanisms do not translate well to traditional performance indicators. A lack of standardized benchmarks makes it challenging to compare neuromorphic chips to GPUs or even to one another. Establishing domain-specific benchmarks—for vision, speech, or robotics—will help accelerate adoption by providing clear performance indicators.

7.4 Training and Model Conversion

Most current AI models are designed for dense deep neural networks, which do not directly map to spiking architectures. The process of converting a conventional deep learning model (e.g., a CNN) to an SNN often involves accuracy loss, performance trade-offs, and retraining. Techniques such as ANN-to-SNN conversion are still evolving, and a unified methodology for creating high-performance SNNs is needed. In the future, we may see end-to-end training methods for SNNs that bypass the conversion step altogether.

8 Future Directions

The future of neuromorphic computing lies in hybrid architectures that combine the strengths of CPUs, GPUs, and neuromorphic cores. Such systems would use CPUs for control logic, GPUs for high-throughput matrix operations, and neuromorphic chips for event-driven inference and low-power real-time tasks.

Advances in 3D integration, chiplet architectures, and photonic neuromorphic processors are expected to push the boundaries of scalability and energy efficiency. Moreover, quantum neuromorphic computing—an intersection of neuromorphic and quantum paradigms—may emerge as a groundbreaking field, blending probabilistic computation with brain-inspired architectures.

Another promising direction is co-design, where hardware and software are developed together to maximize performance. For example, neuromorphic chips could be paired with custom SNN training algorithms optimized for specific applications, such as real-time robotics or edge AI. However, ongoing research and development suggest that hybrid architectures combining CPUs, GPUs, and neuromorphic cores will define the future of computing. This co-existence will allow neuromorphic systems to complement traditional processors by handling event-driven and low-power inference tasks, while CPUs and GPUs continue to manage dense, compute-heavy workloads, VentureBeat (2024–25),. Market Research Future(2023–24), VentureBeat (2024–25)[27-28].

9. Conclusion

The evolution of computing architectures—from CPUs to GPUs, and now toward neuromorphic chips—represents a fundamental paradigm shift in the way we approach computation. CPUs have long been the foundation of general-purpose computing, excelling in sequential task execution. GPUs extended this capability by introducing massive parallelism, enabling breakthroughs in artificial intelligence, scientific computing, and data-intensive applications. However, the growing complexity of modern AI workloads, conventional architectures are reaching their practical limits. Neuromorphic systems can process AI workloads, , at a fraction of the power consumption of GPUs—making them ideal for edge computing, robotics, and autonomous systems.

10 . Declaration about GenAI use

Figure 1 and 2 are drawn using open source GenAI tool named Chatgpt. The author received no financial support for the research, authorship, and/or publication of this article.

11. Author contributions

The author carried out all aspects of the study, including conceptualization, methodology, data analysis, and manuscript preparation.

12. References

1. IMARC Group. *Neuromorphic chip market report 2025*. Brooklyn, New York, USA.
2. Transparency Market Research. *Neuromorphic chip market outlook 2025*., Wilmington, Delaware, USA.
3. SNS Insider, EINPresswire. *Neuromorphic chip market forecast*. 2023.
4. Biasi S, Donati G, Lugnan A, Mancinelli M, Staffoli E, Pavesi L. Photonic neural networks based on integrated silicon microresonators. *arXiv [Preprint]*. 2023 Jun 7;abs/2306.04779. Available from: *https://arxiv.org/abs/2306.04779 arXiv*
5. Li R, Gong Y, Huang H, Zhou Y, Mao S, Wei Z, Zhang Z. Photonics for neuromorphic computing: Fundamentals, devices, and opportunities. *arXiv [Preprint]*. 2023 Nov 16;abs/2311.09767. Available from: *https://arxiv.org/abs/2311.09767 arXiv*
6. Greatorex H, Richter O, Mastella M, Cotteret M, Klein P, Fabre M, Rubino A, Soares Girão W, Chen J, Ziegler M, Bégon-Lours L, Indiveri G, Chicca E. TEXEL: A neuromorphic processor with on-chip learning for beyond-CMOS device integration.

arXiv [Preprint]. 2024 Oct 21;abs/2410.15854. Available from: *https://arxiv.org/abs/2410.15854 arXiv*
7. Le D, et al. A review of memory wall for neuromorphic computing. *arXiv [Preprint]*. 2025 Feb; Available from: https://arxiv.org/abs/xxxx.xxxxx
8. Intel. *Hala Point neuromorphic system with 1.15 billion neurons and ~15 TOPS/W. Quick Market Pitch*. Intel; 2025 Jul.
9. BrainChip Holdings Ltd. *Akida Pulsar microcontroller: First mass-market neuromorphic chip. Quick Market Pitch*. BrainChip Holdings Ltd; 2025 Jul.
10. Innatera. *SNP Spiking Neural Processor for ambient intelligence. Quick Market Pitch*. Innatera; 2025 Jul.
11. Korea Advanced Institute of Science and Technology (KAIST). Complementary-Transformer AI chip running GPT-2 at 400 mW. KAIST News. 2024 Mar.
12. Mead C. Neuromorphic electronic systems. *Proc IEEE*. 1990;78(10):1629-36.
13. Davies M, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*. 2018;38(1):82-99.
14. Merolla PA, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*. 2014;345(6197):668-73.
15. Akopyan F, et al. TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. *IEEE Trans Comput Aided Des Integr Circuits Syst*. 2015;34(10):1537-57.
16. Furber SB, et al. The SpiNNaker project. *Proc IEEE*. 2014;102(5):652-65.
17. Sebastian A, Le Gallo M, Khaddam-Aljameh R, Eleftheriou E. Memory devices and applications for in-memory computing. *Nature Nanotechnology*. 2020;15:529-44.
18. Maass W. Networks of spiking neurons: The third generation of neural network model. *Neural Netw*. 1997;10(9):1659-71.
19. Shafique M, Theocharides T, Hanif MA, Khalid F, Rehman S, Henkel J. An overview of next-generation architectures for machine learning: Roadmap, opportunities and challenges. In: *Proc Des Autom Test Eur Conf Exhib (DATE)*. Piscataway (NJ): IEEE; 2018. p. 827-32.
20. Theis TN, Wong HSP. The end of Moore's law: A new beginning for information technology. *Comput Sci Eng*. 2017;19(2):41-50.
21. BrainChip Holdings. *Akida Neuromorphic System-on-Chip: Technology overview whitepaper*. BrainChip; 2021.
22. Indiveri G, Liu SC. Memory and information processing in neuromorphic systems. *Proc IEEE*. 2015;103(8):1379-97.
23. Amir A, et al. A low power, fully event-based gesture recognition system. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*. Piscataway (NJ): IEEE; 2017. p. 7243-52.
24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-44.
25. Schuman K, et al. A survey of neuromorphic computing and neural networks in hardware. *arXiv [Preprint]*. 2017 May;abs/1705.06963. Available from: *https://arxiv.org/abs/1705.06963*
26. NVIDIA. *CUDA: Parallel computing platform and programming model. White Paper*. NVIDIA; 2022.
27. Venture Beat; Impact Lab. *Innatera edge neuromorphic adoption story*. 2024-25.
28. Market Research Future. *SynSense Xylo IMU HDK launch, NSF CHIPS + Intel neuromorphic deployment*. 2023-24.