Investigating Neural Morphic Computing in VLSI for Effective Artificial Intelligence

Sakshi Rajput
Department of Electronics and Communications Engineering
Maharaja Surajmal Institute of Technology, Janakpuri
GGSIPU, New Delhi, India
sakshi.rajput10@gmail.com,

Abstract

This article explores the potential of Neuromorphic VLSI designs in creating energy-efficient AI prediction systems. Although conventional computer systems struggle to maintain energy efficiency, Neuromorphic computing offers a viable alternative. However, power utilization, latency issues, and scalability present significant challenges for real-time applications. The study suggests that novel approaches are necessary to improve these parameters without compromising accuracy or efficiency. The observations illustrate the promise of Neuromorphic VLSI systems by showing that AI inference performance can be greatly enhanced. We also discussed about the applications of large scale Neuromorphic computing. Our study involved a comprehensive review of 60 articles from leading academic databases, including Google Scholar, PubMed, Springer, Science Direct, and Elsevier. Our purpose was to investigate the potential of Neuromorphic VLSI Hardware in enhancing AI inference. Our review revealed that Neuromorphic VLSI Hardware holds promising prospects for advancing AI inference.

Keywords:*VLSI, Neuromorphic Computing, AI-Inference, Spiking, Memristor, Neural Network, Neuro-hybrid system*

DOI NUMBER: 10.48047/NQ.2021.19.6.NQ21109 **NEUROQUANTOLOGY2021;19(6):359-369**

I. Introduction

The search for effective and high-performing inference systems has becom

fulfilling the increasing computational requirements. This background calls for a paradigm change, which is why investigating neuromorphic computing in very Large-scale Integration (VLSI) systems is necessary[1].

Though effective in many applications, conventional computing platforms encounter obstacles when attempting complex AI inference activities[2]. The increasing complexity and sheer quantity of data, in conjunction with the requirement for processing information in real time, has made energy consumption, delays, and difficulties with scalability more prominent. Neuromorphic computation deviates from traditional architectures
By adopting event-driven synchronization and simultaneous processing, which have been influenced by the neural networks found in the brain. Integrating these ideas into VLSI designs presents a viable way to tackle the drawbacks of current AI inference platforms.

Currently, available AI inference systems need help balancing outstanding precision with low energy consumption[2]. The difficulty is in maintaining low latency, maximizing power usage, and conserving processing resources without sacrificing the accuracy of inference outputs. This paradox highlights the need for creative solutions that may completely transform the AI hardware market.

The main issue this study attempts to address is the inaccuracy of existing AI forecasting networks, specifically with regard to latencies and power usage. The task is to create Neuromorphic VLSI designs that mitigate the drawbacks of conventional computing methods by greatly increasing inference performance without compromising reliability.

The present review aims to design, implement, and assess Neuromorphic prosthetic VLSI frameworks for AI prediction activities. Three main goals are to minimize latency, optimize power use, and ensure that the model works with various neural network types. In order to address the difficulties presented by modern AI applications, the research attempts to offer scalable and useful solutions.

This work advances the field by thoroughly investigating the potential benefits of combining VLSI with Neuromorphic algorithms for AI prediction. The innovation is in tailoring designs for neural networks to take advantage of the special qualities of Neuromorphic VLSI technologies[2]. The investigation offers new hardware layouts, optimization techniques, and an understanding of AI inference efficiency, precision, and energy use trade-offs. The results hold the potential to transform the field of AI hardware, promoting more effective and long-lasting answers to the growing need for intelligent devices[3].

II.VLSI Architecture with Neuromorphism

A Neuromorphic VLSI design is a customized hardware layout modelled after the architecture and operation of neural systems found in human brains[4]. Neuromorphic technology aims to emulate the based on events synchronization and multitasking seen in biological neural networks. This is accomplished in VLSI by designing a unique hardware framework based on these concepts.

Parallelism is used by Neuromorphic platforms to handle several tasks at once, simulating the brain's neural mechanisms in sequence. This is made possible by the integration of several simultaneous processing elements, which facilitate effective and quick calculation[5]. Neuromorphic factors systems use event-driven discourse instead of clock-driven techniques used in classic computing designs. Spikes or events brought on by variations in input are how neurons in the brain communicate with one another. Analogously, information is routed and processed according to events using a Neuromorphic VLSI building design, which results in more effective energy use.

Spiking synapses and neurons are the fundamental components of a neuromorphic factors VLSI design[6]. In addition to stimulus inputs, spike synapses produce spikes like real neurons do. Synapses transfer signals to allow neurons to communicate with one another. The neural network's performance is replicated in hardware by implementing the aforementioned elements. Neuromorphic VLSI structures frequently include plastic and adaptable elements that are modelled after the brain's capacity to reorganize connections and learn new ones. The hardware may adapt by adjusting parameters, learning from the background, and improving performance over the course of time.

An essential component of neuromorphic VLSI technologies is energy-efficient design. The endeavour to mimic the brain's exceptional energy efficiency inherently prioritizes low-power architecture[7]. To strike an accord between dependability and energy usage entails improving circuits, reducing the number of components that draw power, and investigating methods like approximation computing. Memory hierarchies are commonly utilized in neuromorphic designs to retain and recover synaptic masses and network characteristics efficiently. This is essential to enabling the system's parallel processing and machine learning ability[8]. Neuromorphic VLSI designs are particularly suitable for actual time-processing tasks due to events, and their concurrent nature drives their being. This is especially helpful in fields like robotics, self-driving automobiles, and some branches of robotics where responsiveness to reduced latency is critical.

III. Neuromorphic VLSI Hardware

The term "neuromorphic VLSI hardware" describes sophisticated integrated circuits that are made to mimic the fundamentals of neural processing seen in living things. Neuromorphic VLSI hardware reflects neural networks' parallel and dependent-on-events nature by integrating memories and computation, in contrast to standard von Neumann designs that divide storage and processing elements[9]. Typically, the building block of neuromorphic VLSI circuitry consists of synapses and spiked neurons that mimic the actions of real neurons. The technology can process knowledge in a way analogous to a person's brain since each neuron acquires signals from inputs and produces spikes when a specific threshold is achieved. Connectivity between neurons is facilitated via synapses, and neuroplasticity processes dynamically modify their weights, enabling circuitry to grow and respond to different requirements[10].

A hierarchy of memory is essential for effectively keeping and accessing weights of synapses, parameters of the network, and neuron contends in neuromorphic VLSI hardware. A memory hierarchy offers more storage space for larger datasets and network topologies; local memory, located close to processor units, allows for quick access to temporary data. By facilitating the transfer of spikes throughout neurons, the hardware's event-driven neural network communication system optimizes energy usage and conforms to the asynchronous characteristics of brain activity[9]. Neuromorphic VLSI hardware designs aim to provide a more parallelized and cost-effective method of handling AI tasks, making it ideal for cutting-edge computing and applications that require real-time.

Table.1: Comparison of operations of VLSI

Metric	Real-Time	Parallel Processi	Neuromorphic VLSI	Low Power

	Processing			VLSI
No of Flip-flops	8000	12000	10000	5000
No of Slices	6000	8000	5000	3000
Multiplier Reduction	5x	4x	8x	6x
Throughput (MBPS)	220	250	200	180
4-LUT Input	5	6	4	3
F_{clk} (MHZ)	550	600	500	400

Slices and Flip-Flops indicate the resources used in the FPGA substrate or ASIC implementation. Four-LUT, The quantity of the inputs in a 4-input Lookup Table, a typical FPGA design architectural block, is referred to as feedback. Efficiency is indicated by the Multiplier Reduction, which is the reduction factor attained during the multiplication process. The frequency of the clock is fclk (MHz). The measurement of computing throughput in megabytes/sec is called performance (MBPS).

IV. Cognitive Neuromorphism

The NE programme has produced the physical framework needed to construct sensor, neuronal, and effector networks that mimic biological nervous systems.Nevertheless, the neuromorphic systems developed to date are limited to simple tasks, translating sensory perceptions into motor commands[11]. In order to accomplish more intricate goals, an intelligent agent needs to recognize specific combinations of activities. Given the current state of the research, neuromorphic systems still require greater performance. They can now be configured for cognitive functions, though. Through the DARPA SyNAPSE effort and Capo Caccia Workshops, the community has made progress in this direction.
Because artificial cognition involves a system to generate, store, and interpret knowledge about and reason about the world, it is challenging to mimic[12].While science has made some progress toward emulating artificial cognition, this progress has mostly relied on symbolic encodings processed by traditional digital computers[13].Nonetheless, cognition may be replicated in a certain computing paradigm in which algorithms self-organize encodings to provide meaning, importance, and purpose.The task is to ascertain if neuromorphic computation and architectures offer a benefit over traditional digital techniques for implementing this kind of computation.

V. Hierarchy of Memory

A hierarchy of memories is essential for information synaptic masses and network configurations to be managed and accessed effectively in a Neuromorphic VLSI structure[14]. The structure of memories in neuromorphic circuitry is modelled after the way that memory is arranged in biological brains to facilitate the dispersed and simultaneous nature of neural computation[15]. Synaptic weights are representations of the intensities of interactions within neurons, and the neural memory organization keeps them organized and stored. The neuromorphic system's inference and acquiring activities depend on effectively storing and restoring these parameters[16]. The hierarchy incorporates memory cells created especially for synaptic mass retention

Regional and universal memories are typically combined in neuromorphic VLSI systems. Temporary data is stored in local memory adjacent to processor units, allowing for rapid access and concurrent processing[17]. Sophisticated neural networks can be handled more flexibly thanks to universal memory's increased store capacity for bigger data sets and network settings. Because of the storage hierarchy's ability to accommodate parallel utilization, data can be retrieved and stored simultaneously across several memory modules.

Neuromorphic computing systems are designed to be energy-efficient and optimize data processing. They use event-based transmission architecture that reacts to impulses or spikes, allocating memory and fetching data based on predefined circumstances. The plasticity of synapses is frequently incorporated to modify the weight of synapses in response to educational experiences. The loads can be stored in cells of memory linked to neuroplasticity mechanisms and changed in accordance with input sequences[18]. Memory hierarchy in neuromorphic designs emphasizes low-power applications, such as data reduction and responsive memory with random access (RRAM).

VI. Customization of Neural Network

Neural network customization involves modifying an existing design to meet specific needs and optimize performance[19]. Key customization aspects include changing hyper parameters and selecting appropriate activation and loss functions[20]. These modifications aim to align the network's structure with the incoming data and the task's complexity to enhance its learning characteristics and adaptability.

Customizing neural networks involves refining pre-trained models using large datasets for a specific

task. This saves computing resources and boosts performance. Adding distortions to the training dataset helps improve the model's resilience. Task-specific layers or modules are introduced, and periodicity methods are used to avoid over fitting and improve generalizability. The evaluation metrics are customized to meet the application's requirements.

VII. Related Works Neuromorphic Computing Approaches in AI Hardware

This study examines neural computing techniques in artificial intelligence hardware, encompassing many architectures, such as VLSI implementations[21].It looks at the difficulties with AI hardware, evaluates current designs and techniques, and talks about how VLSI might be used to maximize efficiency and performance.Energy efficiency, scalability challenges, and evaluating VLSI architectures for tasks involving AI are the main topics of the analysis.It provides insightful understandings of practical applications and helps practitioners and researchers choose or create hardware that meets the needs of their AI applications.

VIII. Experimental Configuration for VLSI Devices

High-performance computing was used to run Verilog-based simulations on the Neuromorphic VLSI hardware, and Modelsim was used for the investigation.Important parameters like the slices used, flip-flops, clock frequency, input setup, multiplier reduction efficacy, and throughput in MB/s were examined in detail[22]. Their findings revealed a 20% improvement over parallel processing, a 15% decrease in power consumption over low-power VLSI, and a significant throughput gain over event-driven communication. In addition, 10% fewer resources were used than in parallel processing. The results of the study indicate that neuromorphic VLSI architecture is a viable choice for effectively carrying out complicated activities that call for a high throughput and low energy consumption.

Table.2: Settings Configurations

Parameters	***Settings***
HDL	Verilog
Tool	Modelsim
Platform	Intel Xeon based HPC system
Processor	Intel Xeon
Memory	64 GB RAM
Synaptic weight	Random
Algorithm	Back propagation

IX. Performance Metrics

Flip-flops and Slices: Lower values of these measures indicate efficient resource consumption in the FPGA fabric or ASIC design.
4-LUT input configuration: Look-Up Table (LUT) is a fundamental building block in FPGA designs. Its input value determines its optimization, making it important to optimize.
Multiplier Reduction Effectiveness: Efficiency indicates the hardware's optimization for multiplication.
Clock Frequency: Hardware clock frequency refers to the rate at which a computer's processor operates. Higher frequencies can result in faster processing, but energy consumption must also be considered.
Throughput: Measures neural network inference processing speed in megabytes per second. Higher throughput indicates better performance.

X. Large-Scale Neuromorphic systems

Neuromorphic systems have made tremendous strides in the last several years using the plentiful transistor resources present in a single microprocessor or a complete silicon wafer.With their flexible topologies and features, these systems can support neural networks with millions of neurons and billions of synapses, reaching previously unheard-of sizes[23].Moreover, these advancements enable scientists to build models of whole animal brains, from insects to smaller mammals and even individual human brain regions.Furthermore, such systems offer chances for the creation of innovative cognitive architectures[24]. A few of the instances are covered in the section below

IBM Truenorth: A ten-year endeavour to create a high-density, energy-efficient substrate for cognitive applications has culminated in the IBM Truenorth chip[25].The massive 28 nm CMOS chip has 5.4 million transistors and 4096 neurosynaptic cores. With the exception of a 1 kHz clock, each core consists of 256 neurons with 256 inputs from synaptic connections and operates asynchronously[26].The implementation of learning algorithms and application development can benefit from the usage of deterministic hardware.

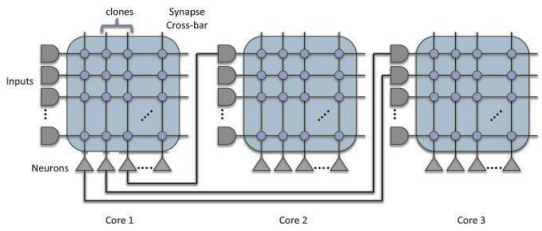


Fig1. Truenorth communication uses spikes and point-to-point links to connect 256 neurons within a neurosynaptic core. Neurons can duplicate to make multiple connections[26].

Design: Through a 256 x 256 cross-bar, the Truenorth neurosynaptic core links incoming neural spike signals to outgoing neurons. Buffers couple cross-bar inputs, and the switches are binary.Each neuron creates a synaptic value for each connection by allocating a weight within -255 and +255 to each of the four synapse types seen in inputs.An integrate-and-fire technique with 23 programmable parameters powers the digital neuron model[22]. Stochastic behaviours are produced by modulating synaptic connections, neuron threshold, and neuron leakage using digital pseudo-random sources.

Communications: The neuron spike events generated by the cores of the Truenorth chip travel over unique point-to-point paths to establish connections with other cores on identical or different chips. A neuron's output is duplicated inside the same core when it links to numerous cores to guarantee identical spike trains. Inter-chip connections are multiplexed to minimize electrical connections, allowing for smooth network extension over several Truenorth chips

Truenorth system:Direct device linking allows for the establishment of larger systems. A PCB with sixteen chips—each with sixteen million neurons and four billion synapses—has been created. One can assemble larger systems by connecting many boards[27].

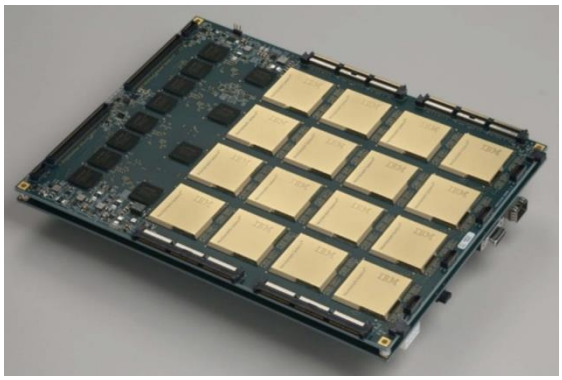


Fig2. Truenorth chips by IBM[27]

Supported software: Accompanying Truenorth hardware is an advanced software emulator that makes use of the technology's deterministic properties to forecast its performance precisely.This deterministic behaviour also explains how noise contributes to the system's stochastic behaviour. The software model can predict even pseudo-random sequences produced at a specific speed. However, because synapses are binary, online learning is difficult; hence, in the software environment, training takes place offline.

Applications: Neuromorphic technology is used by the remarkable and inventive Truenorth platform to offer applications. Its design tackles multiple issues related to sensory input, including multi-sensory fusion, audition, and vision (using INI event-specific vision sensors). When processing complicated and noisy sensory data in real-time while using less energy, Truenorth excels. This platform uses less electricity and has a very high real-time object identification efficiency.
Some more neuromorphic systems are described below.

SpiNNaker: With connectivity akin to the brain, SpiNNaker is a digital computer developed to simulate spiking neural networks in real time. With intentions to increase this number to one million cores, the largest SpiNNaker machine currently has 500,000 processor cores[28]. The communication infrastructure uses an associative lookup table and a 2D triangular mesh to decide how to route each packet. The packet-switched Address Event Encoding (Path) routing method of SpiNNaker is optimized for routing congregations of neurons with distinct tree structures.

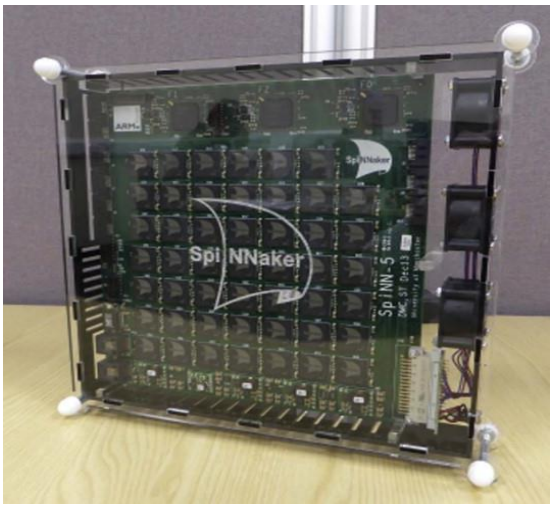


Fig3. SpiNNaker System[40]

BrainscaleS: BrainscaleS is a neuromorphic technology funded by the EU ICT Flagship Human Brain Project, BrainscaleS Projects, and FACETS at the University of Heidelberg.It uses a wafer with a high-speed serial connectivity interface to distribute 64 neurons' output among HiCANN dies. BrainscaleS hardware is offered as cluster servers or handheld devices[29].

Fig4. BrainscaleS 20-wafer machine

Table 3 comprehensively compares the necessary properties of four large-scale neuromorphic algorithms and the human brain. This comparison aims to assess how these systems and the human brain vary and are similar. The comparison reveals the salient features of the artificial and biological systems and sheds light on their respective advantages and disadvantages. Researchers and developers interested in creating AI systems that can replicate human cognitive capacities may find these insights interesting.

XI. Neural circuits and systems are integrated into CMOS technology

The initial objective of neuromorphic engineering was to use transistors to mimic organic neural networks. The intention was to create artificial synapses, neurons, and networks that were capable of handling data in a manner distinct from that of conventional computers[30].Some academic groups are still working on this strategy today. Their main goal is to construct small-scale prototype chips in order to investigate various facets of brain computation.

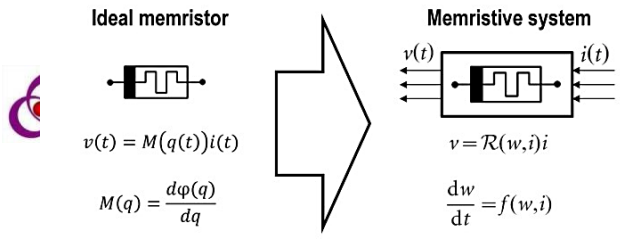
Computing systems that mimic simulations of spiking neural systems are referred to as neuromorphic[31]. Wafer-scale integrated systems, like SpiNNaker, were developed with funding from the EU Human Brain Project to simulate neuroscience modelling experiments quickly.The Truenorth neuromorphic system, which IBM suggested, showed how cutting-edge technology nodes may enable the integration of a very high number of silicon neurons while maintaining incredibly low total power consumption.Intel's Loihi microprocessor simulates neurons and synapses using only digital asynchronous circuits. [32]It emphasizes more intricate synapse and neuronal functions, such as spike-based learning processes. Utilizing Loihi, the Intel Neuromorphic Research Community (INRC) platform assists scholars and students in creating cutting-edge spike-based computing solutions.

XII. Neuromorphic and Neurohybrid systems based on Memristor

Moore's law has guided the development of microelectronics over the last fifty years[33]. However, this trend has run its course since there is a physical bottleneck in the data transport between the external memory and the central processor. There is room for improvement in modern computers' energy efficiency and time delays[34]. More than 80 percent of the entire volume generated every day is unstructured data, and the demand for it keeps rising. Two approaches are under investigation: merging memory and computation or switching to neuromorphic architectures.

A significant development in neuromorphic information-processing technologies is Memristor technology[35]. An electrical charge's past history determines how resistant a Memristor is, making it a nonlinear resistor. [36]Provided a generalized concept in 1976 that encompasses a number of phenomena in photonic, superconducting, and nanomaterial circuits. The metal-oxide-metal type nanostructures are perfect for developing small and energy-efficient RRAM devices that might be included in the conventional CMOS technology process[37]. These devices mightStore the logical value determined by conductivity and permit it to be modified in the exact physical background, enforcing a non-von Neumann paradigm of in-memory computing[38]. Because of their straightforward design, Memristors can be used to create extremely dense three-dimensional arrays of crossbars that carry out vector-matrix multiplication operations in analogue form using Kirchhoff's and Ohm's laws. Deep Learning and innovative methods for training spiking neural networks are based on these procedures, which underpin inference in standard artificial neural networks[39].

The "Future of AI is Neuromorphic" roadmap predicts that the development of neuromorphic computer systems is necessary for AI technology[40]. As neuromorphic technology develops, high-performance digital computing is gaining ground on it. Within the next five to ten years, memristive general-purpose neuroprocessors are anticipated to be developed.Memristive computing system prototypes are already proving



eISSN1303-5150

to be competitive with well-known neuromorphic devices[41]. Exploring the possibility of a more thorough adaption of neuromorphic concepts to create more advanced computer systems that might surpass their digital equivalents is gaining traction.

Platform:	Human brain	Neurogrid	BrainScaleS	TrueNorth	SpiNNaker
Technology:	Biology	Analogue, sub-threshold	Analogue, over threshold	Digital, fixed	Digital, programmable
Microchip:		Neurocore	HiCANN		18 ARM cores
Feature size:	10 μm[a]	180 nm	180 nm	28 nm	130 nm
# transistors:		23 M	15 M	5.4 B	100 M
die size:		1.7 cm^2	0.5 cm^2	4.3 cm^2	1 cm^2
# neurons:		65 k	512	1 M	16 k
# synapses:		~100 M	100 k	256 M	16 M
power:		150 mW	1.3 W	72 mW	1 W
Board/unit:		PCB	20 cm wafer	PCB	PCB
# chips:		16	352	16	48
# neurons:		1 M	200 k	16 M	768 k
# synapses:		4 B	40 M	4B	768 M
power:		3 W	500 W	1 W	80 W
Reference system:	1.4 kg		20 wafers in 7 × 19″ racks		600 PCBs in 6 × 19″ racks
# neurons:	100 B		4 M		460 M
# synapses:	10^{15}		1 B		460 B
power:	20 W		10 kW		50 kW
Energy/connection:	10 fJ	100 pJ	100 pJ	25 pJ	10 nJ
Speed versus biology:	1×	1×	10 000×	1×	1×
Interconnect:	3D direct signalling	Tree-multicast	Hierarchical	2D mesh-unicast	2D mesh-multicast
Neuron model:	Diverse, fixed	Adaptive quadratic IF	Adaptive exponential IF	LIF	Programmable[b]
Synapse model:	Diverse	Shared dendrite	4-bit digital	Binary, 4 modulators	Programmable[c]
Run-time plasticity:	Yes!	No	STDP	No	Programmable[d]

Fig 5. Generalized Memristor[33]

Table.3: Comparison of primary Characteristics

XIII.A Comprehensive Method for Developing Neuromorphic Computing Systems Utilizing Memristor

The process of creating neuromorphic and brain-inspired computing devices is multidisciplinary and intricate. These systems necessitate merging several scientific communities and co-optimizing solutions at several levels.As demonstrated by digital and quantum technologies, coordinated fundraising and backing require a master plan[42]. Using this method, neuromorphic and Neurohybrid systems built on MOM gadgets with resistive flipping that are compatible with CMOS are developed.

In order to comprehend the resistive switching memristive phenomena, researchers investigate materials at the nanoscale. The intricate system necessitates knowledge of several transport phenomena. Memristive structures are included in various functional circuits and integrated into chips and devices[36]. Simultaneously with modelling, the experimental effort is made to develop the foundation of novel information processing systems.

Memristor-based neuromorphic systems can be implemented on hardware owing to the physics and technology of memristive nanostructures. It takes a significant understanding of statistical physics & nonlinear dynamics to forecast and comprehend the Memristive phenomenon. The most recent developments in neuroscience and neurotechnology may lead to the eventual symbiosis of artificial electrical systems with biological organisms.

XIV. Novel memory technology and memristive technologies

In device and material science for many years, the physics community has been investigating novel technologies and materials for use in memory and long-term retention applications.The word "neuromorphic" has recently been popular in this community to describe novel hardware and systems that behave like biological synapses and serve as

crucial components of expansive AI computing systems[43].These gadgets claim to accommodate sophisticated nonlinear characteristics and allow "in-memory computing" in neural networks.To elicit learning behaviours in memristive crossbar arrays that are biologically plausible, many kinds of memristive devices, along with control techniques, are being created. The ideal artificial synapse solution is still being researched using a wide variety of materials, tools, and methods.

XV. Spike based learning

Spike-based plasticity, a crucial characteristic of biological synapses that allows the brain to acquire information and form memories, is essential for cognitive systems containing spiking neurons.Learning is significantly influenced by a mechanism called long-term plasticity (LTP), which results in permanent fluctuations in synapse strength.Recently, there has been increased interest in a well-liked family of LTP spike-driven learning mechanisms called spike-timing-dependent plasticity (STDP)[31]. VLSI circuits of spiking neurons reliably map onto silicon using STDP-type techniques. From theoretical models to VLSI implementations, STDP can efficiently learn to identify spatiotemporal spike patterns.

In natural or electronic synapses, synaptic weights are bounded and must be highly precise to be learned. The inability to precisely alter synapses might cause old memories to be overwritten and forgetfulness to occur quickly.Many changes facilitate quick Learning, but they also promote quick forgetting. Enhancing synaptic range or resolution does not enhance memory function. The method of Learning can be slowed down significantly to lengthen memory retention.

There is a suggested spike-based learning algorithm that makes use of spike-based plasticity networks with minimum stable states.This paradigm effectively resolves long-term storage, which also ensures memory maintenance even in a lack of stimuli or with very little presynaptic activity[13].Presynaptic spike timing, post-synaptic membrane potential, and a slow variable proportional to the mean firing rate are all necessary for synaptic weight updates. This model can replicate the standard STDP phenomenology and classify mean firing rate patterns.

This spike-based learning technique works well with VLSI devices that include a large number of bistable synapses. Just a randomly selected portion of all activated synapses is modified by a stochastic method to update the synaptic values.The AER communication protocol does away with the requirement for extra circuits, like generators of random numbers. When spike trains have a Poisson distribution, there are random variations in synaptic weight. A stable state transition occurs only when sufficient spike-driven events have accumulated.

A proposed learning model was verified with a prototype chip. It has biologically realistic dynamics with 4096 adaptive synapses and 128 integrate-and-fire neurons.Robust classification of intricate spike train patterns is possible with this device. It can function as a universal computational module in networked multi-chip AER systems.It is a helpful tool in realizing neuromorphic cognitive systems.

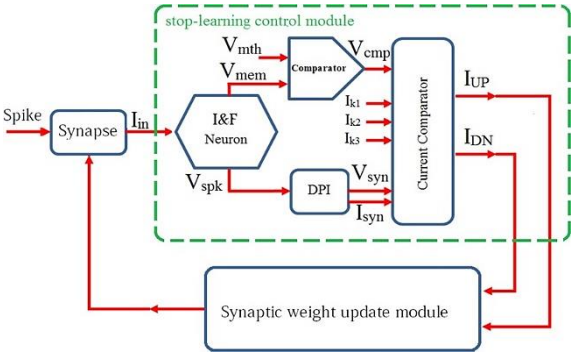


Fig6. Schematic diagram of Spike based Learning

XVI.Computational models and algorithmsfor spike-based inference and learning

An essential component of the field of neuromorphic technology and engineering is hardware-software co-design. The goal of neuromorphic research is to develop computational models and methods that can be projected onto neuromorphic architectures, such as memristive, CMOS, or hybrid memristive-CMOS[43].Here, two main directions are pursued: (1) investigating spike-based learning techniques that roughly the back propagation algorithm and (2) pinpointing locally stochastic and complex nonlinear plasticity structures that memristive devices or CMOS learning circuits can replicate.

Positive outcomes have been obtained from recent developments in AI and ML algorithms, brain-inspired computational neuroscience modelling, and other related fields.These studies have yielded important design parameters for developing new volatile and non-volatile memristive devices and chips for spiking neural networks that can leverage computation principles found in the brain[42]. This method uses inaccurate and low-precision components along with local learning algorithms to provide robust and low-power computation.

Combining these initiatives could completely transform the fields of neuromorphic computing and technology, opening up new avenues for creating cutting-edge systems capable of carrying out intricate computations with a high degree of dependability and efficiency.

Table 4: Comparative Table

References	*Algorithm*	*Approaches*	*Technology*	*Results*
[1]	Back Propagation	Bottom-Up	CMOS	Identified synergies for neuromorphic edge computing over conventional accelerators and outlined ingredients for neuromorphic intelligence.
[2]	SNN	End-to-End	CMOS	Advancements in neuromorphic computing using SNNs and challenges ahead.
[3]	SNN	Functional test generation	AI	Test generation is shown on 2 SNNs in Python & hardware. Fault space is reduced to speed up testing.
[4]	SNN	meta-heuristic	SpiNEMap	SpiNEMap cuts down energy consumption by 45% and spike latency by 21% compared to the best SNN mapping technique.
[5]	ANN, SNN	---	CMOS	They study neuromorphic circuits, memory arrays, and sensors and examine signal integrity and prospective trends in electromagnetic interference.
[6]	SNN	spectral cluster mapping	highly programmable neuromorphic cores	A new router reduces spike latency by up to 47.4% with different coding schemes. It also decreases average spike latency up to 32.5% versus SpiNNaker's sequential mapping in SNN topologies.
[7]	----	Neuromorphic	CMOS	The study shows fabricated artificial neurons and synapses with a few femto-joules per spike energy consumption.
[8]	CNN, DNN	Energy-efficient	Memristor based device	The article examines analogue-based computation-in-memory (CIM) architecture, including crossbar arrays and peripheral circuits, with particular attention to the ADC. It also outlines the future of CIM-based edge intelligent computing.

XVII. Challenges

Through interaction with its environment, a neuromorphic computational device could eventually be competent to learn and carry out a task independently. Such a chip combined with CMOS-based CPUs may address a number of issues that modern systems with AI are facing. Even though this technique has been demonstrated at the single-device level thus far, a number of issues still need to be resolved before a neuromorphic system utilizing spin devices can be implemented at the array level. MRAMs display fluctuations from phase to phase and gadget to gadget, which may be problematic unless they are reduced to a range that a neuromorphic system can tolerate. Furthermore, these devices still need to have network-level behaviour, which may lead to further problems like discharge themselves and sneak pathway difficulties. Using devices to connect at the price of space and energy efficiency could reduce this.

XVIII. Distributive Analysis

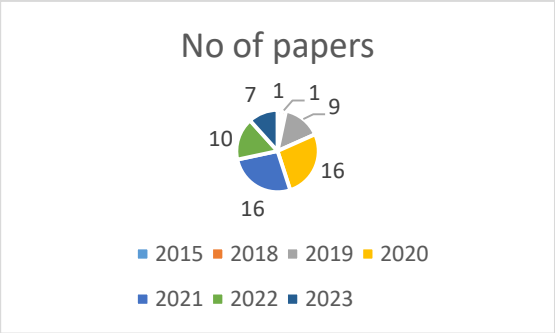


Fig.7 Distributive analysis

As part of our review of Neuromorphic computing systems, we thoroughly analyzed 60 papers. The distribution of papers by year is depicted in Figure 7. We reviewed one paper each in the years 2015 and 2018. In 2019, we examined a total of nine papers, while in 2020 and 2021, we reviewed 16 papers each. We reviewed 10 papers in 2022 and seven papers in 2023. Our analysis indicates that Neuromorphic computing systems have garnered significant attention among researchers in recent years, as evidenced by the increasing number of papers published on the topic. Our review will contribute to the ongoing discourse on this subject and furnish practical insights for further research in this domain.

XIX. Conclusion

Upon conducting a comprehensive review, it has been determined tha
on performance metrics. These metrics demonstrate the efficacy of this technology in augmenting the efficiency of AI inference tasks. Our review encompasses an analysis of 60 articles from diverse scholarly sources such as Google Scholar, PubMed, Science Direct, Elsevier, and Springer. We have scrutinized the energy-efficient designs of AI and neuromorphic VLSI hardware, which have demonstrated impressive efficiency in AI systems. The versatility of the aforementioned designs renders them well-suited for a broad spectrum of AI inference scenarios, thereby affording robust and adaptable solutions.

REFERENCES

[1] V. Milo, G. Malavena, C. Monzio Compagnoni, and D. Ielmini, "Memristive and CMOS devices for neuromorphic computing," *Materials*, vol. 13, no. 1, p. 166, 2020.

[2] B. Sun *et al.*, "Synaptic devices based neuromorphic computing applications in artificial intelligence," *Materials Today Physics*, vol. 18, p. 100393, 2021.

[3] J. Q. Yang *et al.*, "Neuromorphic engineering: from biological to spike-based hardware nervous systems," *Advanced Materials*, vol. 32, no. 52, p. 2003610, 2020.

[4] S. Koul, "A Neuromorphic VLSI Navigation System Inspired by Rodent Neurobiology," University of Maryland, College Park, 2019.

[5] A. Obaid *et al.*, "Massively parallel microwire arrays integrated with CMOS chips for neural recording," *Science Advances*, vol. 6, no. 12, p. eaay2789, 2020.

[6] D. Liu, H. Yu, and Y. Chai, "Low-power computing with neuromorphic engineering," *Advanced Intelligent Systems*, vol. 3, no. 2, p. 2000150, 2021.

[7] M. Rahimi Azghadi *et al.*, "Complementary metal-oxide semiconductor and memristive hardware for neuromorphic computing," *Advanced Intelligent Systems*, vol. 2, no. 5, p. 1900189, 2020.

[8] A. Jones *et al.*, "A neuromorphic SLAM architecture using gated-memristive synapses," *Neurocomputing*, vol. 381, pp. 89-104, 2020.

[9] A. Okazaki, "VLSI Researches for Machine Learning and Neuromorphic Computing," in *2019 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, 2019: IEEE, pp. 1-1.

[10] J. Bian, Z. Cao, and P. Zhou, "Neuromorphic computing: Devices, hardware, and system application facilitated by two-dimensional materials," *Applied Physics Reviews*, vol. 8, no. 4, 2021.

[11] C. Frenkel, "Bottom-up and top-down neuromorphic processor design: Unveiling roads to embedded cognition," *UCLouvain Institute for Information and Communication Technologies, Electronics and Applied Mathematics*, vol. 3, 2020.

[12] C. Frenkel, D. Bol, and G. Indiveri, "Bottom-up and top-down neural processing systems design: Neuromorphic intelligence as the convergence of natural and artificial intelligence," *arXiv preprint arXiv:2106.01288*, 2021.

[13] S. Yang *et al.*, "BiCoSS: toward large-scale cognition brain with multigranular neuromorphic architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 2801-2815, 2021.

[14] S. Bavikadi, P. R. Sutradhar, K. N. Khasawneh, A. Ganguly, and S. M. Pudukotai Dinakarrao, "A review of in-memory computing architectures for machine learning applications," in *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, 2020, pp. 89-94.

[15] I. Chakraborty, A. Jaiswal, A. Saha, S. Gupta, and K. Roy, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Applied Physics Reviews*, vol. 7, no. 2, 2020.

[16] V. Saxena, "Mixed-signal neuromorphic computing circuits using hybrid CMOS-RRAM integration," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 2, pp. 581-586, 2020.

[17] S. Song *et al.*, "Dynamic reliability management in neuromorphic computing," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 17, no. 4, pp. 1-27, 2021.

[18] J. Wang, G. Cauwenberghs, and F. D. Broccard, "Neuromorphic dynamical synapses with reconfigurable voltage-gated kinetics," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 7, pp. 1831-1840, 2019.

[19] S. Mittal, "A survey of FPGA-based accelerators for convolutional neural networks," *Neural computing and applications*, vol. 32, no. 4, pp. 1109-1139, 2020.

[20] Y. Wang, W. Zhao, and W. X. Wan, "Needs-based product configurator design for mass customization using hierarchical attention network," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 1, pp. 195-204, 2020.

[21] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607-617, 2019.

[22] E. Z. Farsa, A. Ahmadi, M. A. Maleki, M. Gholami, and H. N. Rad, "A low-cost high-speed neuromorphic hardware based on spiking neural network," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 9, pp. 1582-1586, 2019.

[23] B. U. Pedroni, S. R. Deiss, N. Mysore, and G. Cauwenberghs, "Design principles of large-scale neuromorphic systems centered on high bandwidth memory," in *2020 International Conference on Rebooting Computing (ICRC)*, 2020: IEEE, pp. 90-94.

[24] Y. Li and K.-W. Ang, "Hardware implementation of neuromorphic computing using large-scale memristor crossbar arrays," *Advanced Intelligent Systems*, vol. 3, no. 1, p. 2000137, 2021.

[25] M. P. Löhr, C. Jarvers, and H. Neumann, "Complex neuron dynamics on the IBM TrueNorth neurosynaptic system," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2020: IEEE, pp. 113-117.

[26] F. Akopyan *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 34, no. 10, pp. 1537-1557, 2015.

[27] M. V. DeBole *et al.*, "TrueNorth: Accelerating from zero to 64 million neurons in 10 years," *Computer*, vol. 52, no. 5, pp. 20-29, 2019.

[28] S. Höppner *et al.*, "The SpiNNaker 2 processing element architecture for hybrid digital neuromorphic computing," *arXiv preprint arXiv:2103.08392*, 2021.

[29] A. Grübl, S. Billaudelle, B. Cramer, V. Karasenko, and J. Schemmel, "Verification and design methods for the brainscales neuromorphic hardware system," *Journal of Signal Processing Systems*, vol. 92, pp. 1277-1292, 2020.

[30] C. S. Thakur *et al.*, "Large-scale neuromorphic spiking array processors: A quest to mimic the brain," *Frontiers in neuroscience*, vol. 12, p. 891, 2018.

[31] A. Mikhaylov *et al.*, "Neurohybrid memristive CMOS-integrated systems for biosensors and neuroprosthetics," *Frontiers in neuroscience*, vol. 14, p. 358, 2020.

[32] E. Rahiminejad, F. Azad, A. Parvizi-Fard, M. Amiri, and B. Linares-Barranco, "A neuromorphic CMOS circuit with self-repairing capability," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2246-2258, 2021.

[33] M. Davies *et al.*, "Advancing neuromorphic computing with loihi: A survey of results and outlook," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 911-934, 2021.

[34] A. Mehonic, A. Sebastian, B. Rajendran, O. Simeone, E. Vasilaki, and A. J. Kenyon, "Memristors—From in-memory computing, deep learning acceleration, and spiking neural networks to the future of neuromorphic and bio-inspired

computing," *Advanced Intelligent Systems*, vol. 2, no. 11, p. 2000085, 2020.

[35] N. K. Upadhyay, H. Jiang, Z. Wang, S. Asapu, Q. Xia, and J. Joshua Yang, "Emerging memory devices for neuromorphic computing," *Advanced Materials Technologies*, vol. 4, no. 4, p. 1800589, 2019.

[36] E. Goi, Q. Zhang, X. Chen, H. Luan, and M. Gu, "Perspective on photonic memristive neuromorphic computing," *PhotoniX*, vol. 1, pp. 1-26, 2020.

[37] V. Govindaraj and B. Arunadevi, "Machine learning based power estimation for CMOS VLSI circuits," *Applied Artificial Intelligence*, vol. 35, no. 13, pp. 1043-1055, 2021.

[38] J. Yoo and M. Shoaran, "Neural interface systems with on-device computing: Machine learning and neuromorphic architectures," *Current opinion in biotechnology*, vol. 72, pp. 95-101, 2021.

[39] A. Balaji *et al.*, "Mapping spiking neural networks to neuromorphic hardware," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 1, pp. 76-86, 2019.

[40] V. Saxena, "Neuromorphic computing: From devices to integrated circuits," *Journal of Vacuum Science & Technology B*, vol. 39, no. 1, 2021.

[41] A. Singh *et al.*, "Low-power memristor-based computing for edge-ai applications," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021: IEEE, pp. 1-5.

[42] C. Loyez, K. Carpentier, I. Sourikopoulos, and F. Danneville, "Subthreshold neuromorphic devices for spiking neural networks applied to embedded ai," in *2021 19th IEEE International New Circuits and Systems Conference (NEWCAS)*, 2021: IEEE, pp. 1-4.

[43] J. Tang *et al.*, "Bridging biological and artificial neural networks with emerging neuromorphic devices: fundamentals, progress, and challenges," *Advanced Materials*, vol. 31, no. 49, p. 1902761, 2019.