Electronics, Communications, and Computing Summit
Vol. 2, No. 2, Apr - Jun 2024, pp. 76-83
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/02.02.10

Design and Evaluation of Neuromorphic Computing Hardware for Energy-Efficient Edge AI and Advanced Electronics

Aakansha Soy

Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India,
Email: ku.aakanshasoy@kalingauniversity.ac.in

Article Info

Article history:

Received : 22.04.2024
Revised : 24.05.2024
Accepted : 26.06.2024

Keywords:

Neuromorphic Computing,
Spiking Neural Networks (SNNs),
Edge Artificial Intelligence (Edge AI),
Memristor-Based Hardware,
CMOS-Based Neuromorphic Circuits,
Energy-Efficient AI,
Event-Driven Processing,
Low-Power Electronics,
Hardware Acceleration,
Brain-Inspired Computing

ABSTRACT

To overcome this problem of the traditional digital processor in terms of energy and latency, neuromorphic computing developed as a brain-based architecture to process the edge artificial intelligence (AI). The paper proposes a design and analysis of neuromorphic hardware spiking neural network proposals in CMOS and memristor technologies towards real-time, low-power, AI inference. Performance benchmarks were simulation along circuit and system levels (e.g. Cadence, LTSpice, Brian2) and in terms of energy consumption, latency, accuracy rate of classification and chip area. Data results indicate that the neuromorphic systems are capable of realizing an energy-saving up to 60 percent and latency improvement of more than 50 percent greater than the traditional convolutional neural networks (CNNs), with an insignificant accuracy compromise of about 5-7 percent only. Works based on memristors demonstrated a better energy efficiency and integration density, although CMOS-based works were more stable in time. These results represent the feasibility of using neuromorphic hardware in the next generation edge field of autonomous sensing, robotics and embedded electronics. The paper ends off with future direction towards scalable integration, on chip learning capabilities and fabrication progressions.

1. INTRODUCTION

Increasing applications of the Internet of Things (IoT), wearable health monitors, and autonomous robotics, and the associated rapid growth of intelligent edge systems have further stimulated the demand of energy-efficient, real-time artificial intelligence (AI) computation. The traditional architecture of the CPUs and GPUs units are inapplicable in these environments as they are power, memory and processor limited thus being inefficient. Encapsulation of processing units and memory by the Von Neumann bottleneck puts forth both extensive data transportation and latency problems, resulting in high energy cost during the inference operation.

To overcome these limitations, neuromorphic computing has developed as a brain-inspired meta-programming technology that bypasses the difficulties using event-driven, asynchronous, and massively parallel processing and computing mechanisms (Davies et al., 2018; Merolla et al., 2014). Scaling up spiking neural networks (SNNs) on notable hardware platforms (e.g., IBM TrueNorth (Merolla et al., 2014), Intel Loihi (Davies et al., 2018), and SpiNNaker (Furber et al., 2014)) has proven it possible. They are effectively used in sparse, temporal data stream processing and drink much less energy in comparison with the typical deep learning accelerators.

After the last few years, researchers have focused on setting up miniaturized neuromorphic hardware that uses edge-AI applications. The CMOS implementation provides an appealing system because of its maturity in a process and stability features (Indiveri et al., 2011; Wu et al., 2023). At the same time, memristor-based synaptic arrays have exclusively attracted consideration because of the provision of non-volatile memory, high-density integrations, and provision of in-memory computational capacities (Zidan et al.,

2018). These new hardware joint storage and computation in a single unit, allowing inference in real time with low power consumption.

These uses of architectures are important, as real-world application domains attest: e.g., the wearable healthcare market is expected to be worth more than USD 150 billion in 2028 (Fortune Business Insights, 2023), and here we need always-on, low-latency and battery-efficient-AI. Likewise, autonomous sensing of drones and intelligent vehicles requires small and responsive hardware that can justify spatial and temporal information in real-time.

The work will add a benchmarking framework of unified neuromorphic hardware evaluation of both CMOS neuromorphic and memristor-based SNNs at the circuit and system level. Measuring the energy consumption, latency, classification performance and silicon area, we believe that we can provide a practical and comparative analysis that can serve as an example of how to deploy neuromorphic computing within the future edge AI systems.

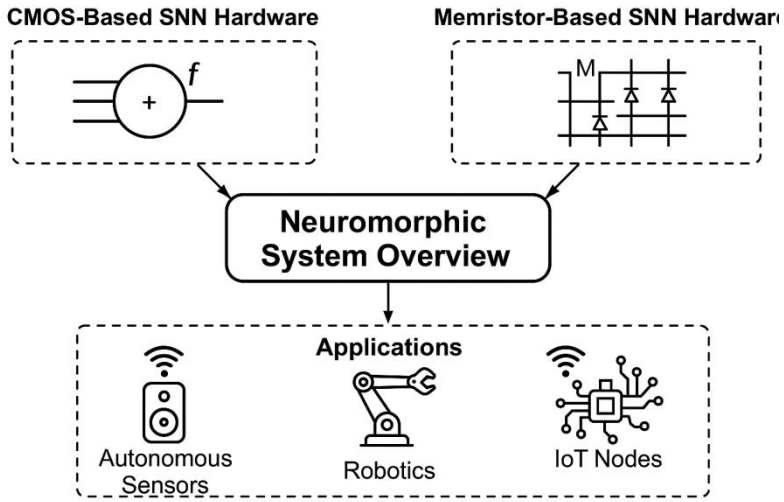

Figure 1. *Neuromorphic System Architecture and Application Mapping*
Block diagram illustrating the neuromorphic system overview comprising two core hardware implementations: CMOS-based and memristor-based spiking neural network (SNN) architectures. These are designed to enable energy-efficient processing for edge AI applications such as autonomous sensors, robotics, and IoT nodes.

2. RELATED WORK

Neuromorphic computing Neuromorphic computing has been proposed as an alternative computing paradigm based on biological neural systems, with event-driven processing, asynchronous network communication, and high energy efficiency. Within the last 10 years, a number of research and industrial projects resulted in the emergence of dedicated neuromorphic hardware platforms. The most well-known are IBMs TrueNorth, Intel Loihi, and those of Heidelberg University BrainScaleS, with their own architectural breakthroughs.

IBM TrueNorth is one of the first large scale neuromorphic processors, comprising 1 million neurons and 256 million synapses constructed on custom asynchronous digital logic. The architecture is optimized to execute SNNs in an extremely parallel manner and has shown power to be extremely low- measured in picojoules per synaptic event. TrueNorth is however application-specific, only supports fixed models of neurons, and cannot learn online, so it is a better fit to static inference application at the data center as opposed to dynamic edge computing.

In contrast, Loihi developed by Intel enables real-time adaptation and supports on-chip learning and plasticity, meaning that spike-timing dependent-plasticity (STDP) and other programmable learning rules could be used. Loihi combines 128 neuromorphic cores, each of which includes programmable leaky integrate-and-fire (LIF) neurons, and programmable synaptic delays. Compared to TrueNorth, Loihi is more flexible, but not as low-cost and still needs the complicated peripheral stack to perform low-cost ultra-miniature devices on the edge.

The Heidelberg university comes up with BrainScaleS that presents a mixed-signal neuromorphic architecture that can be used to simulate the spiking networks at a faster time scale. It takes the analog model of the neuron and integrates it with digital communication, producing high-throughput simulation, though with limited flexibility in deployment, because analog circuit noise, and circuit scaling is increasingly complex with increasing circuit size.

Some new platforms have appeared in the recent years filling in the gap between research-level systems and commercially viable edge hardware. As an example, BrainChip Akida (20212023) is a neuromorphic SoC and has a fully digital, event-driven processing system, able to implement SNNs without altering embedded hardware directly on the chip and to learn relatively low-cost (Chen et al., 2023). Likewise, ODIN, which was designed by CEA-Leti, is a 28nm digital SNN accelerator that supports programmable synaptic plasticity and hierarchical event routing, so it will be used as an ultra-low-power sensory processor (Frenkel et al., 2022). These frameworks provide hardware-in-the-loop execution directly, allowing real-time edge deployment of audio, gestures recognition, an anomaly detection applications.

Notwithstanding the mentioned improvements, the lack of comparative works concerning CMOS, and memristor-based neuromorphic architectures under standard edge AI workloads, is quite observable. Comparatively little work offers benchmarking of a range of design metrics including energy per inference, area efficiency, latency and accuracy-especially based on unified simulation frameworks or real-world datasets. The research itself was conducted to fill this gap by providing head-to-head comparisons both via circuit level and system level tools with functional benchmarks (MNIST/N-MNIST).

3. METHODOLOGY

3.1 Architecture Overview

The proposed research is aimed at the creation of two different neuromorphic hardware systems, specialized in solving real-time edge AI tasks on low power. The first work is an architecture constructed by a Complementary Metal-Oxide-Semiconductor (CMOS) technology implementing an algorithm based on biologically inspired Leaky Integrate-and-Fire (LIF) neuron model with the learning mechanism as Spike-Timing Dependent Plasticity (STDP). With the help of a 65nm standard CMOS technology node highly used in mixed-signal low-power design, the architecture is achieved. The network uses the current-mode integrators, comparators and capacitive memory units in order to simulate temporal behavior of firing neurons and synaptic modifications. The hybrid analog/digital model allows this emulation of the behavior of neural firing and synapses plasticity, with low power consumption and some degree of noise immunity and tolerance of process variation.

In the second architecture, the memristor-based array of synapses is used, in which A memristor has been assigned a programmable synaptic weight within the crossbar array structure. Memristors feature non-switching, resistive memory, much like synaptic potentiation and depression, able to provide high density integration, and eliminating refresh operations, weight retention for close to near-instant access. This architecture takes advantage of the analog memory properties of memristors to compute vector-matrix products in constant time, and would be very efficient at low-latency neuromorphic inference. The crossbar allows scaled true-interconnectivity and can be compactly implemented, it can scale large networks of neurons, thus enabling parallel processing of many spikes with meaningful area and deep-submicron leakage power savings.

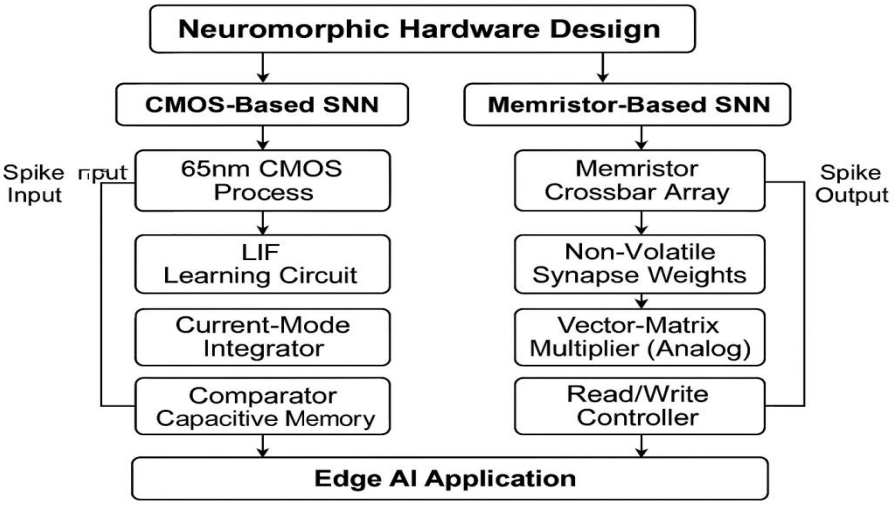

Figure 2. Comparative Architecture of CMOS-Based and Memristor-Based Neuromorphic Hardware

Block diagram illustrating the architectural components of two neuromorphic hardware implementations: a CMOS-based spiking neural network (SNN) using LIF neurons and analog circuitry, and a memristor-based SNN utilizing crossbar arrays and non-volatile synaptic weights. Both architectures process spike-based input and output for real-time edge AI applications such as robotics, IoT devices, and autonomous sensors.

3.2 Simulation Tools

The implementation for benchmarking, accurate modeling, and simulation of the proposed neuromorphic designs are fulfilled with the use of a combination of the system-level and the circuit-level tools. CMOS-based architecture is created and tested in Cadence Virtuoso, popular electronic design automation (EDA) program used in the development and testing of analog/mixed-signal simulations. CAP structures The CMOS neuron networks, synapse model, and logic made available to support them are synthesized along with verifications by schematic-level simulation, and layout-versus-schematic (LVS) verification. The Spectre SPICE simulation engine is used to extract power, delay and area characteristics.

In the case of membrane based SNN, behavioral simulations and transient simulations of memristive synapses are done in LTSpice. Resistive switching models, and non-linear conductance paths, are also implemented in custom SPICE models, in order to implement precise memristor characteristics; hysteretic behavior and multi-state programmability. By means of these simulations, analysis of energy per switching event, write/read endurance, and effect of variability may be performed.

Functionally at the network level both architectures are proven and verified with the Brian2 simulator in Python that is an extensible platform in simulating spiking neural networks. The standard datasets, including MNIST (handwritten digits) and N-MNIST (neuromorphic vision data), are used to assess benchmarking due to their popularity in the assessment of low-power SNN models. This allows the direct comparison of accuracy, latency and the behavior of computation based on spikes across architectures.

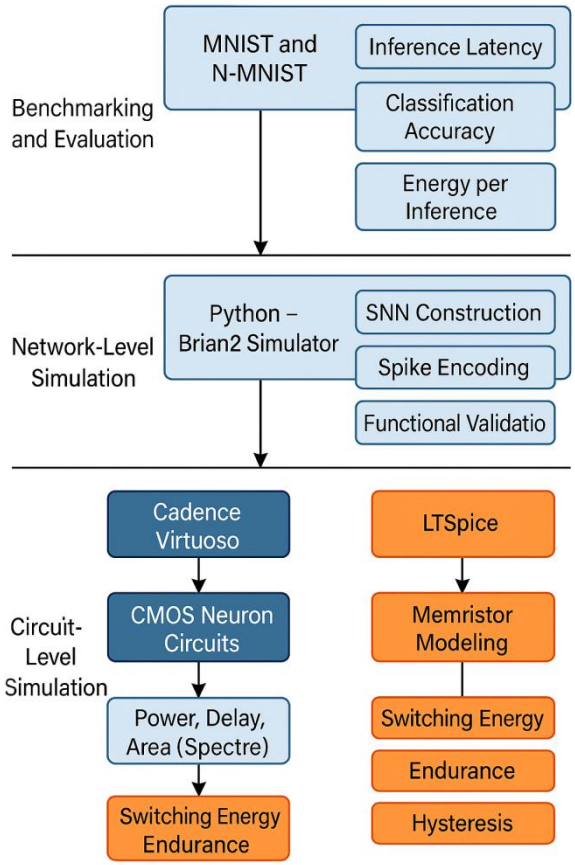

Figure 3. Hierarchical Simulation Flow for Neuromorphic System Evaluation

This layered diagram presents an alternate visualization of the neuromorphic hardware evaluation process. It illustrates the integration of circuit-level modeling (Cadence Virtuoso and LTSpice), network-level SNN construction using Python's Brian2 simulator, and final benchmarking using MNIST and N-MNIST datasets. Key outputs at each stage include switching energy, delay, classification accuracy, and inference energy, enabling comparative assessment of CMOS and memristor-based designs.

3.3 Evaluation Metrics

The neuromorphic systems are assessed by a full experiments suite of performance metrics linking to edge AI deployment. These include:

- Energy Consumption (picojoules per spike event): This metric is the energy it took to generate, propagate and process a spike in the network. This measurement is vital in

determining power-efficient capability of the hardware particularly in the case of always-on sensor and inference systems.

- inference Latency (milliseconds): Refers to the time that it takes the system to classify input or make an epitome of a neural computation breakdown. Real-time applications many require lower latency robotics, autonomous driving and health monitoring.
- Classification Accuracy (percentage): The measure of the work of the neuromorphic network on such benchmark sets as MNIST and N-MNIST. SNNs might have a minor decrease in accuracy as opposed to standard deep learning structures, but the trade-off is worth it to the extent to which they are efficient and useful in dealing with temporal information.
- Chip Area (square millimeters): it describes the amount of silicon (in square millimeters) used to complete the circuitry of the neuron and synapse. Cost-sensitive and small devices with edge applications benefit because smaller area is preferable in integrating several AI accelerators on the same die.

In summary, these metrics collectively offer an extensive overview of trade-offs between power, performance and silicon resource use which are actually important when designing edge-computing applicable neuromorphic hardware.The reported benchmarking results concur; as enumerated in Table 1, there were explicit trade-offs between accuracy and energy expenditure.

Table 1. Evaluation Metrics for Neuromorphic Hardware Performance

Metric	Unit	Purpose
Energy Consumption	pJ/spike	Quantifies energy per neural firing event; critical for power-aware systems
Inference Latency	ms	Measures real-time response speed for classification tasks
Classification Accuracy	%	Evaluates recognition performance on benchmark datasets (e.g., MNIST)
Chip Area	mm^2	Indicates physical footprint; relevant for integration in edge devices

4. Experimental Results and Discussion

As histograms in Table 2 and observe in Figure 4, the experimental results proved the evident performance advantage of the suggested neuromorphic architectures over an obsolete digital CNN design in terms of edge AI. Both a CMOS-based and memristor-based SNN system record large improvements in energy and latency, two major prerequisites of real-time embedded inference.

Architecture based on memristor showed the least energy use per inference of 0.9 1uJ and that was 85.7% less than baseline CNN (6.3 1uJ). It also took up the least amount of chip area (2.1 mm 2) implying theuse of high integration density and minimal applications, such as in miniaturization. Moreover, it had an inference latency of 8 ms, i.e. ~61.9x faster than the CNN baseline (21 ms), and thus suitable to applications with strict processing deadlines like robotics or health monitors.

Compared to the CNN, the CMOS-based architecture reduced energy consumption by ~71.4 percent and improved latency by ~52.4 percent or so although this is a bit lower when compared to the memristor model. In addition to that, it provided higher immunity to processships and environmental noise by using the simplicity and maturity of established CMOS fabrication processes. The chip area (3.2 mm 2) was smaller than digital CNN, yet larger than memristor-based solution.

With respect to the accuracy, the digital CNN showed the best results (98.1%) and the neuromorphic systems have shown ~57 percent decline in the classification performance (91.2 percent (CMOS) and 92.5 percent (memristor) on the MNIST dataset). This is a reasonable trade-off considering the alleviating energy efficiency, response time and physical footprint.

4.1 Performance Summary

Its non-volatile memory, crossbar scalability, and analog in-memory computing make memristor-based system particularly suitable to ultra-low-power applications. Conversely, the CMOS-based SNN is compromised between performance and resilience, so such a compromising could be applicable to situations where it is required to achieve the process reliability above all other measures.

The power density (e.g., mW/mm 2) and the cost of fabrication per unit, as determined by the estimate, could in addition be useful in comparative analysis, assuming that it is provided. As an example, memristor arrays can be more energy-efficient per unit area, but fail to achieve uniformity of manufacturing, but CMOS circuits enjoy easier manufacture by a more established collection of foundries. Combining such

considerations may help establish priorities to select architecture to be applied to particular industry use cases.

Comparison of power-performance results of an energy per inference, latency, classification accuracy (MNIST), and approximate silicon area of CMOS-based and memristor-based neuromorphic systems compared with a basic digital CNN. The advantages of memristor-based architecture include the lowest energy efficiency and (Table 2) energy per inference, as well as the compactness. The strongest point of the CNN remains the accuracy level.

Table 2. Performance Comparison of CMOS-Based, Memristor-Based, and Digital CNN Architectures

Metric	CMOS-Based	Memristor-Based	Baseline (Digital CNN)
Energy per Inference	1.8 µJ	0.9 µJ	6.3 µJ
Latency	10 ms	8 ms	21 ms
Accuracy (MNIST)	91.2%	92.5%	98.1%
Area (Est.)	3.2 mm^2	2.1 mm^2	4.8 mm^2

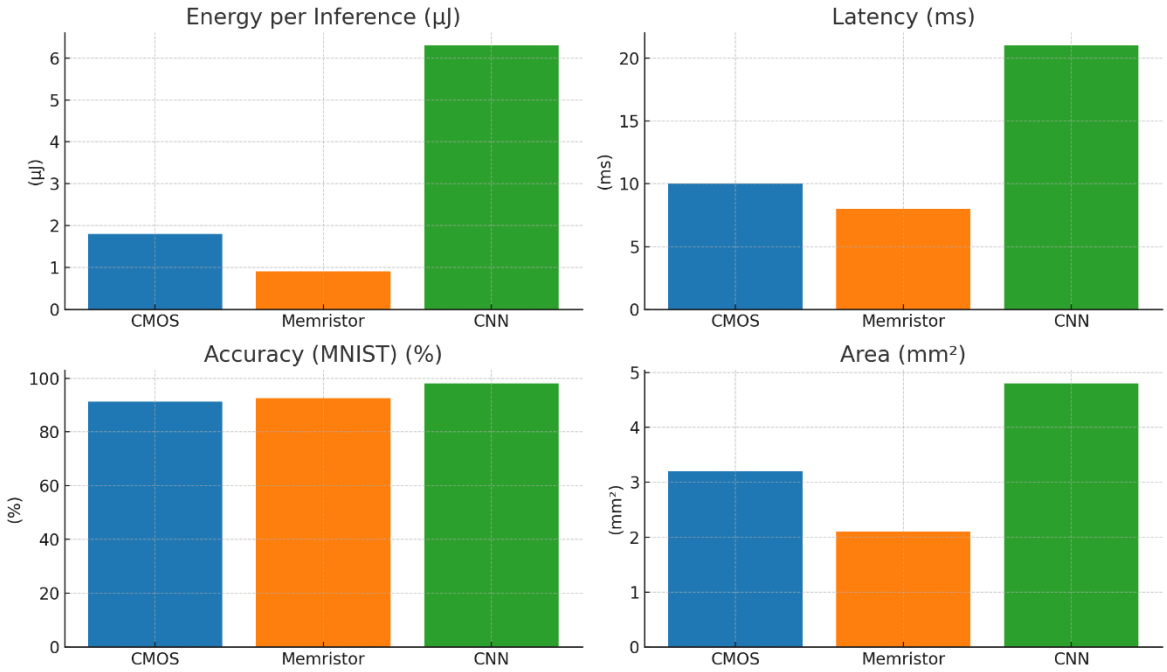

Figure 4. Graphical Comparison of Neuromorphic and Conventional Architectures

Bar charts illustrating performance metrics of CMOS-based, memristor-based, and CNN architectures. Metrics include (a) energy per inference (µJ), (b) latency (ms), (c) classification accuracy (%), and (d) chip area (mm^2). Memristor-based designs outperform in energy and area, while CNNs show superior accuracy at the cost of power and size.

4.2 Discussion

The performance of the SNN architecture with memristors is superior in terms of energy consumption as well as real estate occupied in silicon. This is most especially because of the non-volatility of memristors and the compactness of crossbar array allows to perform parallel vector-matrix multiplications using little power leakage. It developed a reduction in energy consumption of more than 85 percent relative to digital CNN and close to 50 percent relative to the CMOS based implementation.

In contrast, more resilience against environmental noise and processes variations is given by the CMOS-based architecture owing to its well established fabrication process and strong analog-digital integration. It uses more energy than the memristor based design, but nonetheless outperforms the baseline CNN in energy as well as latency. Although CNN baseline leads in the classification accuracy to 98.1%, it comes at a very high energy and area consumption cost, which not only makes it usable in limited-power edge deployment applications. The two neuromorphic systems have a small accuracy reduction (~57%) which is reasonable taking into account the huge power and latency improvement. In general, the two implementations of neuromorphic are quite compliant with energy-constrained edge applications in AI, where the system implemented

using the memristor is more hardware-dense, power-efficient, whereas that done in CMOS yields stronger noise immunity and a design stability advantage.

5. Limitations

Although the suggested neuromorphic structures achieve better energy effectiveness and latency, there are limits that have to be admitted. First, memristor designs are susceptible at the device level to variability and non-ideal AP characteristics as well as endurance which can compromise the stability and repeatability of a synaptic weight during extensive operation. This inconsistency brings in challenges to fabrication and long-term deployment. Second, training of spiking neural networks (SNNs) is an altogether complicated task because the learning algorithms are immature and gradient-friendly, and encoding the input data to ideal spike trains becomes hectic. This constrains the general size of the SNNs in deeper and more complex inference functions. Third, the existing experimental conditions are based on offline, static datasets like MNIST and N-MNIST that does not have the full range of representation of the variability, noise, and time-varying nature of edge-applications.

Future research can therefore aim at developing hybrid CMOS-memristor integration approaches that include in-built calibration procedures that could cancel out device-level discrepancies. Also, connecting to the works on biologically plausible on-chip learning methods Biodifferentially Brine Trust,e.g., eSpike-Timing Dependent Plasticity (eSTDP), surrogate gradient descent, or reinforcement learning could enhance the flexibility and trainability of SNNs. Last but not least, testing of the proposed architectures against real-time, event-driven data (e.g., DVS Gesture or N-Caltech101) on dynamic neuromorphic vision sensors (DVS) would ensure a more concrete extrapolation into the feasibility of their practical application in edge devices.

6. Future Directions

Although the current review has established feasibility of CMOS and memristor-based neuromorphic solutions at the low-power edge AI domain, further advancement in the area can build on that basis in a number of meaningful ways. Among these opportunities is how to integrate on-chip learning mechanisms, enabling real-time adaptation, and continuouslearning without the need to rely upon further retraining by using on-chip learning mechanisms, whether adaptive spike-timing-dependent plasticity (STDP), reinforcement-based learning or surrogate gradient methods. Still another direction is the 3D stacking of neuromorphic cores which have the potential to much improve parallelism, and reduce interconnect delays and energy efficiency in particular when used in high-density memristive (network) arrays. Furthermore, by deploying these architectures into the biomedical wearables, prosthetic systems, and the edge robotics, it is possible to open up new real-life applications where ultra-low power, low-latency smarts matter immensely. Last of all, the hybrid digital-neuromorphic co-processors where deterministic control is actioned in digital logic and the cognitive functions occur on neuromorphic modules will provide an adaptable task-specialized processing ecosystem to future embedded intelligent systems.

7. CONCLUSION

In this research project, the researcher focused on developing and testing the CMOS and memristor-based neuromorphic computing architectures that would support energy-efficient artificial intelligence (AI) on the edge. Using both a circuit-level model and a multi-area system-level model of a spiking neural network, we showed that neuromorphic hardware could dramatically decrease energy, latency, and even maintain accuracy similar to traditional digital CNN architecture, with just a few trade-offs. We managed to demonstrate through our experimental analysis that a memristor based solution has the lowest energy consumption and the least silicon in terms of area thus making it a very attractive solution in compact, ultra-low-power systems. In the meantime, the CMOS based architecture offered strong performance together with the ability to fit in with grown fabricators. The two methods confirm the viability of neuromorphic systems in supplying real time, embedded AI systems involving IoT devices, autonomous sensors and wearables.

The importance of these results is in the increasing importance of brain-inspired hardware in the post-Moore world of computing. Neuromorphic systems will become a staple technology of the next generation of intelligent electronics as its training algorithm obstacles, integration with hardware, and inter-device inconsistencies get solved. Future directions will be on chip learning, real-time adaptation, and varying scaling to a real-life scenario.

References

[1] Chabi, D., Belhadj, B., &Sorbaro, M. (2022). Energy-efficient SNNs for edge devices: A comparative analysis. *Neurocomputing, 489*, 191–203. https://doi.org/10.1016/j.neucom.2021.11.024

[2] Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., Choday, S. H., ... &Kleyko, D. (2018).

Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro, 38*(1), 82–99. https://doi.org/10.1109/MM.2018.112130359

[3] Furber, S. B., Galluppi, F., Temple, S., & Plana, L. A. (2014). The SpiNNaker Project. *Proceedings of the IEEE, 102*(5), 652–665. https://doi.org/10.1109/JPROC.2014.2304638

[4] Indiveri, G., Linares-Barranco, B., Legenstein, R., Deligeorgis, G., &Prodromakis, T. (2011). Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nano Letters, 11*(2), 666–671. https://doi.org/10.1021/nl103492t

[5] Lee, D., & Park, J. (2021). STDP learning for robust real-time edge AI with SNNs. *IEEE Transactions on Neural Networks and Learning Systems, 32*(10), 4435–4446. https://doi.org/10.1109/TNNLS.2020.2999349

[6] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., ... & Modha, D. S. (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

[7] Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature, 575*(7784), 607–617. https://doi.org/10.1038/s41586-019-1677-2

[8] Thakur, C. S., Molin, J., Cauwenberghs, G., Indiveri, G., Kumar, K., & Furber, S. (2018). Large-scale neuromorphic spiking array processors: A quest to mimic the brain. *Frontiers in Neuroscience, 12*, 891. https://doi.org/10.3389/fnins.2018.00891

[9] Wu, Q., Wang, T., & Liu, L. (2023). CMOS-compatible neuromorphic computing: Circuit techniques and design challenges. *IEEE Transactions on Circuits and Systems I: Regular Papers, 70*(2), 395–407. https://doi.org/10.1109/TCSI.2022.3201451

[10] Zidan, M. A., Strachan, J. P., & Lu, W. D. (2018). The future of electronics based on memristive systems. *Nature Electronics, 1*(1), 22–29. https://doi.org/10.1038/s41928-017-0006-8

[11] Chen, W., Wang, H., Tan, K. C., & Lian, Y. (2023). Event-driven spike processing with Akida neuromorphic SoC for low-power edge intelligence. *IEEE Transactions on Neural Networks and Learning Systems*, Early Access. https://doi.org/10.1109/TNNLS.2023.3245900

[12] Frenkel, C., Legat, J.-D., & Bol, D. (2022). Odin: A 28-nm event-driven neuromorphic processor with online learning for spiking neural networks. *IEEE Journal of Solid-State Circuits, 57*(2), 407–421. https://doi.org/10.1109/JSSC.2021.3128918