SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

NEUROMORPHIC IN-MEMORY COMPUTING FOR ENERGY-EFFICIENT AI: A COMPREHENSIVE REVIEW

Pournima P R
Department of ECE, B E S Institute of Technology, Bengaluru, India

Article DOI: https://doi.org/10.36713/epra23011
DOI No: 10.36713/epra23011

ABSTRACT

As the demand for artificial intelligence continues to grow, there is a need for specialized hardware that can efficiently process and power AI algorithms. This is where VLSI (Very Large-Scale Integration) design comes into play. By integrating millions, or even billions, of transistors onto a single chip, VLSI design enables the creation of powerful AI-integrated circuits.

However, designing hardware specifically for artificial intelligence poses its own set of challenges. The complexity of AI algorithms and the need for parallel processing capabilities demand innovative solutions. VLSI designers must carefully consider factors such as power consumption, heat dissipation, and the integration of specialized neural network accelerators.

In addition to these considerations, VLSI designers must also address the scaling limitations of semiconductor technology. As AI algorithms become more complex, the demand for advanced VLSI architecture increases. Designers must find ways to enhance transistor density and performance while overcoming the limitations imposed by the physical laws governing semiconductor manufacturing.

In-Memory Computing (IMC) is an emerging paradigm designed to overcome the limitations of the traditional Von Neumann architecture, particularly for energy-efficient AI applications.

Traditional computing systems suffer from the "memory wall"—a bottleneck caused by the constant transfer of data between the CPU and memory. This leads to inefficiencies in latency, bandwidth, and energy consumption, especially for AI tasks which involve huge volumes of matrix operations and data movement.

KEYWORDS – *Very Large-Scale Integration (VLSI), Artificial Intelligence (AI), In-memory computing (IMC), Neuromorphic, multiply-and-accumulate (MAC), RRAM (Resistive RAM), MRAM (Magnetoresistive RAM), PCM (Phase-Change Memory), FeFET (Ferroelectric FET), spiking neural networks (SNNs)*

I. INTRODUCTION

Human brains are vastly more energy efficient at interpreting the world visually or understanding speech than any CMOS based computer system of the same size. Neuromorphic computing can perform human-like cognitive computing, such as vision, classification, and inference. The fundamental computing units of artificial neural network are the neurons that connect to each other and external stimuli through programmable connections called synapses. The basic operation of an artificial neuron is summing the N weighted inputs and passing the result through a transfer (activation) function. Such neuron and synapse functions can be efficiently implemented using different emerging post-CMOS device technologies. Our research in this area include:

- Physical modelling of nanoscale emerging devices for potential neuron or synapse applications, such as resistive RAM, spin-transfer torque devices, spin-orbit torque devices, magnetic domain wall motion devices, magnetic skyrmion, etc.
- Exploration of various neuromorphic computing models, such as Deep Learning Convolutional Neural Network, Spiking Neural Network, Hierarchical Temporal Memory, Oscillatory Neural Network, etc
- Cross-layer (device/ circuit/ architecture) co-design for implementing complex machine learning tasks, such as pattern/ speech recognition, semantic reasoning, robotic control and motion detection

Neuromorphic engineering is an emerging field that aims to create AI chips based on the principles of the human brain. By mimicking the structure and functionality of neural networks, neuromorphic chips can enhance the performance and energy efficiency of AI applications. This trend in VLSI Design for AI opens up exciting possibilities for creating hardware that can learn and adapt in real time.

Neuromorphic VLSI circuits, inspired by biological neural networks, have emerged as a promising solution to address the growing demand for energy-efficient and high-performance AI/ML applications. Traditional computing architectures face limitations in power consumption, scalability, and real-time processing, especially for complex, data-intensive tasks. In this paper, we propose the design and implementation of neuromorphic VLSI circuits that mimic the structure and functionality of biological neurons and synapses. By leveraging event-driven, asynchronous spiking neural networks (SNNs), our circuits are able to process information in a parallel and distributed manner, significantly reducing power consumption while improving computation speed. The proposed

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

neuromorphic circuits integrate in-memory computing, which eliminates the energy bottlenecks associated with data transfer between memory and processing units in conventional systems. This paper highlights the architectural advancements in VLSI design that enable real-time learning and adaptation, making these circuits highly suited for AI/ML tasks such as image recognition, natural language processing, and autonomous systems. Simulation results of [4] demonstrate that the neuromorphic VLSI circuits achieve superior energy efficiency and performance compared to traditional AI hardware. This research opens new avenues for developing low-power, scalable AI solutions in edge computing and other energy-constrained environments.

The groundbreaking method of neuromorphic computing, which is inspired by the neural architecture of the human brain, is generating waves in the field of computer science as well as in the design of Very Large-Scale Integration (VLSI). This cutting-edge technology has the potential to change a variety of fields, including artificial intelligence, robotics, healthcare, and other fields as well. In this paper, the investigation of the advent of neuromorphic computing and look into the implications it has for very large-scale integrated circuit design (VLSI design) has been done.

Challenges faced will be such as process node scaling, memory integration, and thermal management in AI hardware design are identified, with a forward-looking analysis on future trends like quantum computing, 3D VLSI integration, and emerging technologies such as RRAM and photonic computing.

In-Memory Computing (IMC) embeds computation capabilities directly within or near the memory, allowing data to be processed where it's stored.

In IMC, the main idea is to perform operations like matrix-vector multiplication (common in neural networks) directly in memory arrays—often using resistive memory technologies such as:

- RRAM (Resistive RAM)
- MRAM (Magnetoresistive RAM)
- PCM (Phase-Change Memory)
- FeFET (Ferroelectric FET)

Instead of fetching weights and inputs separately to a processor, these memory cells can execute operations like multiply-and-accumulate (MAC) in-place using analog or digital methods.

Feature	Energy-Efficiency Impact
✅ Reduced Data Movement	Eliminates the back-and-forth between processor and memory, which accounts for up to 90% of energy in AI workloads.
✅ Parallelism	Many computations can be done in parallel at the memory array level.
✅ Lower Latency	On-chip processing shortens the inference time.
✅ Compact Hardware	Allows for tighter integration and edge deployment.

Table 1: Key points of AI in IMC

❖ **Applications**

1. Edge AI Devices (drones, wearables, smart sensors)
2. Neuromorphic Systems (brain-inspired architectures)
3. Low-Power Vision and Audio Processing
4. Real-Time Inference at the Edge

❖ **Research and Industry Efforts**

- IBM: Analog IMC chips for DNN inference
- Intel Loihi: Neuromorphic chip integrating memory and compute
- MIT, Stanford, UC Berkeley: Pioneering in analog and mixed-signal IMC for deep learning
- Startups: Mythic, Rain Neuromorphics, Syntiant are building analog or mixed-signal IMC chips

❖ **Challenges**

✓ **Precision and Noise:** Analog computation can introduce errors.
✓ Scalability: Difficult to scale for very large models.
✓ Programmability: Need new compilers and frameworks.
✓ Limited Training Capability: IMC is mostly useful for inference, not training (yet).

AI compute workloads and fixed power budgets are forcing chip and system architects to take a much harder look at compute in memory (CIM), which until recently was considered little more than a science project.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

CIM solves two problems. First, it takes more energy to move data back and forth between memory and processor than to actually process it. And second, there is so much data being collected through sensors and other sources and parked in memory, that it's faster to pre-process at least some of that data where it is being stored. Or looked at differently, the majority of data is worthless, but compute resources are valuable, so anything that can be done to reduce the volume of data is a good thing.

- **Opportunities and Obstacles Involved in VLSI Design**
- ✓ Building Neural Networks with Physical Components: Investigating the difficulties that can arise when attempting to implement neural networks in VLSI designs.

The factors of optimization and efficiency that should be considered while designing neural network circuits which need to be considered are

- ✓ The Number of Circuits Per Unit Area And Their Interconnectivity: Taking into account the difficulties presented by the high circuit density of neuromorphic designs.
- ✓ The Integration of Differently Focused Components: Having a conversation on the process of integrating specialized components such as synaptic connections and memristors.

- **Implications for both Artificial Intelligence and Machine Learning**
- ✓ Improving the Speed of AI and ML Algorithms: Investigating the potential for neuromorphic computing to speed up AI and machine learning algorithms.
- ✓ Recent Achievements in the Field of Pattern Recognition: The ability of neuromorphic systems to excel at tasks such as pattern recognition and anomaly detection.

- **Neuromorphic computing in applications that take place in the real world**
- ✓ Robotics and Self-Driving Systems: Having a conversation about the application of neuromorphic processors in robotics for the purpose of real-time sensor fusion and adaptive learning.

- **Applications in Healthcare and the Biomedical Sciences**: Investigating the application of neuromorphic computing to the study of real-time medical data and the early diagnosis of disease.
- ✓ Internet of Things with Energy Management: Investigating the potential implications of neuromorphic systems for intelligent energy management and grids.

- **Future Paths to Take and Obstacles to Overcome**
- ✓ Adaptability to Varying Demands and Standardization: The difficulties that arise when trying to scale up neuromorphic systems for use in real applications are discussed.

- **Challenges Presented by Algorithm Training and Optimization:** Taking up the current scientific problems of improving algorithms for neuromorphic systems is the focus of this article.

- **Inter- and multidisciplinary research collaborations and efforts:** Bringing attention to how important it is for VLSI designers, computer scientists, and neuroscientists to work together.

Designing efficient CIM architectures is non-trivial, though. In work presented at this year's VLSI Symposium, researcher Yuhao Ju and colleagues at Northwestern University considered AI-related tasks for robotics applications[6]. Here, general-purpose computing accounts for more than 75% of the total workload, including such tasks as trajectory tracking and camera localization. In recommendation engines, the large pool of possibilities identified through table lookups still needs to be filtered against the user query. Even when neural network calculations are clearly identifiable as a limiting factor, the exact algorithms involved differ. Neural network research is advancing faster than integrated circuit design cycles. A hardware accelerator designed for a particular algorithm might be obsolete by the time it's actually realized in silicon.

One possible solution, seen in designs like Samsung's LPDDR-PIM accelerator module, relies on a simple, but general-purpose calculation module, optimized for matrix multiplication or some other arithmetic operation. Software tools designed to manage memory-coupled computing assume the job of effectively partitioning the workload.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

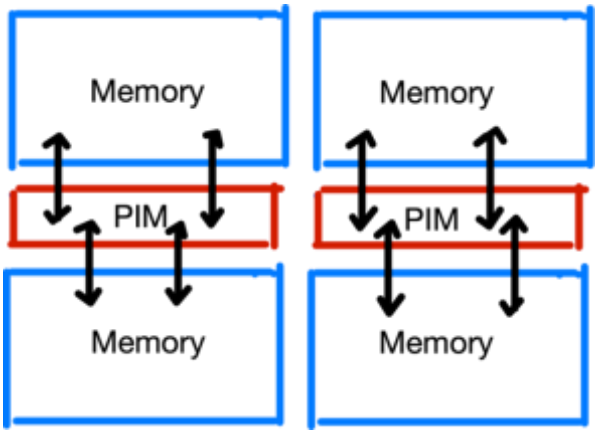


Fig. 1: One possible PIM architecture places a floating-point module between memory banks. Source: K. Derbyshire/Semiconductor Engineering [5]

However, using software to assign tasks adds overhead. If the same data is sent to multiple accelerators, one for each stage in an algorithm, the advantage of CIM may be lost. Another approach, proposed by the Northwestern University group, integrates a CNN accelerator with conventional general-purpose logic. The CPU writes to a memory array, which the accelerator treats as one layer of a CNN. The results are written to an output array, which the CPU treats as an input cache. This approach reduced end-to-end latency by as much as 56%.

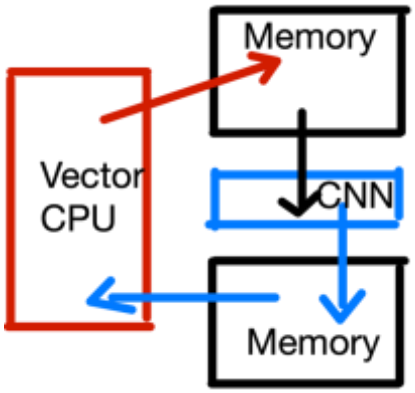


Fig. 2: A unified architecture can boost both neural network and vector calculations. Source: K. Derbyshire/Semiconductor Engineering [5]

These solutions are feasible with current device technology. They depend on conventional CMOS logic and DRAM memory circuits, where the data path is fixed once the circuit is fabricated. In the future, fast non-volatile memories could enable reconfigurable logic arrays, potentially blurring the line between "software" and "hardware."

There is a pressing need for energy-efficient non-von Neumann computing systems for highly data-centric artificial intelligence related applications. An approach that efficiently performs a wide range of machine learning tasks such as compressed sensing, unsupervised learning, solving systems of linear equations, and deep learning has been developed [5].

The next generation computational architectures require embedding deep learning, machine learning and artificial intelligence onto computationally efficient hardware and devices with low energy footprints. This is necessitated by an ever-increasing demand in computing power, which, coupled with data explosion is pushing the boundaries of conventional computational architectures. One such hybrid architecture based on neuromorphic population coding, where encoding of information is carried out by the activity in an ensemble of neurons such as in the olfactory, motor, and visual cortex has been proposed. The proposed hybrid architecture utilizes a CMOS-based silicon neuron as a basic computing element, and molybdenum disulphide (MoS_2) based two-dimensional synaptic memtransistor as an analogue memory, instead of conventional digital memories, to avoid bottleneck in terms of power and area. A combination of silicon technology with 2D materials can surpass the current technology limitation and can offer remarkable advancements in circuit architecture and device performance at atomic levels.

In this regard, the propose a biologically inspired wake-up system [9], shown in Fig 3, with embedded intelligence and energy efficient footprint, that can be integrated with existing edge computing devices to improve their energy and performance. This system exploits the inherent mismatches present in lower silicon technology nodes to reduce computational complexity and simultaneously uses multi-state memtransistive device for low power computation and reduced computational complexity. In the current hybrid framework, this synaptic memtransistor memory provides two functions simultaneously, one a substitute for digital

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

memories as an adaptable multi-state memductance and the other is the execution of an inherent multiplication operation by Kirchhoff's current law (KCL), thus offering a low power computational paradigm. From the promising results for demonstrated regression and classification tasks, this framework proves to be a step closer for designing a low power, reduced dimensions, fault-tolerant, and robust architecture. These characteristics enable the framework to be employed in onsite processing of data such as in IOT devices, edge devices, energy- and area-constrained devices or devices with low computational resources.

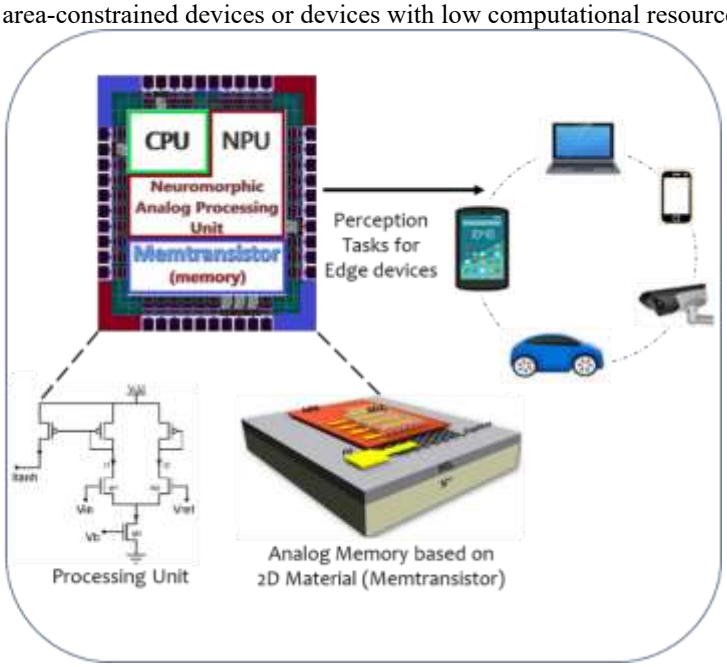


Fig. 3: Chip architecture with functional blocks and applications [9]

The next era of computing for artificial intelligence (AI) requires efficient hardware with increased computational performances and ultra-low energy footprint. This requires a disruptive approach in architecture design, unlike the traditional von Neumann approach which is serial in nature. Presently, the computations are achieved by CPUs, GPUs and TPUs which utilises von Neumann architectures to perform multiply and accumulate (MAC) operations in parallel. Such architectures have large energy and memory footprints. In contrast, the neuromorphic architectures are inspired by neurobiological systems and have characteristics such as in-memory computing, stochastic computing and massive network fan-outs. They are parallel and distributed architectures which exhibit low precision along with adaptive learning. A significant improvement in energy and other performance footprints can be obtained by implementing such architectures in analogue. But the analogue implementation of such architectures comes with its own challenges such as the designed hardware is sensitive to device mismatch and other non-linear effects. Furthermore, the designed architecture depends on the physical responses of the MOS transistors operating in different biasing regimes and hence requires distinct circuit architecture for different computational operations. These challenges along with other trade-offs limit the flexibility in design parameters and put performance and power restrictions on the designed circuit. The proposed hardware software-based co-design platform [10 & 11] can achieve a significant boost in performance, power and accuracy.

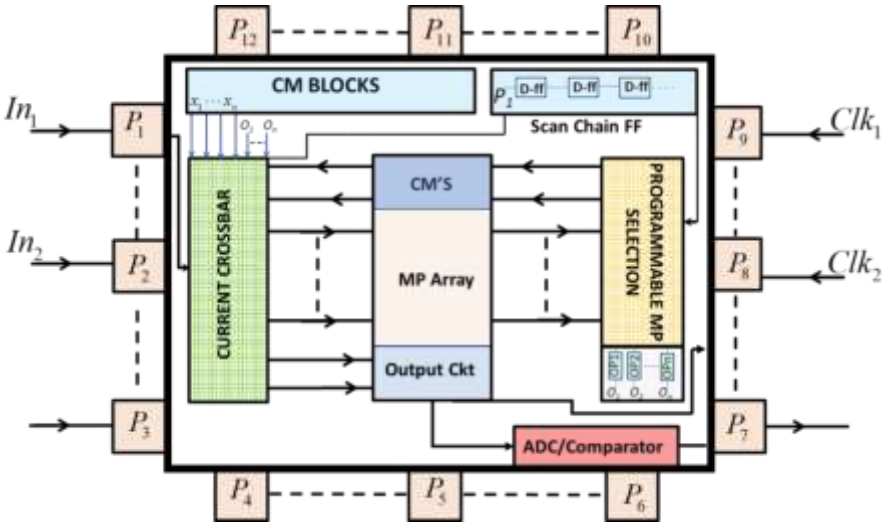


Fig. 4: Top Level Chip Architectural Block [10 & 11]

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

In a conventional computer, the processing and memory units are physically separated. Consequently, a significant amount of data needs to be shuttled back and forth during computation, which creates a performance bottleneck commonly referred to as the "von Neumann Bottleneck" (see Figure 5). This physical separation and associated data transfers are arguably one of the main hurdles of traditional computers, as a memory access typically consumes 100 to 1000 times more energy than a processor operation. In in-memory computing, computation is performed in place by exploiting the physical attributes of memory devices organised as a "computational memory" unit. For example, if data A is stored in a computational memory unit and if we would like to perform f(A), then A is not required to be brought to the processing unit (see Figure 5). This is more energy and time efficient than the process performed by a conventional computing system.

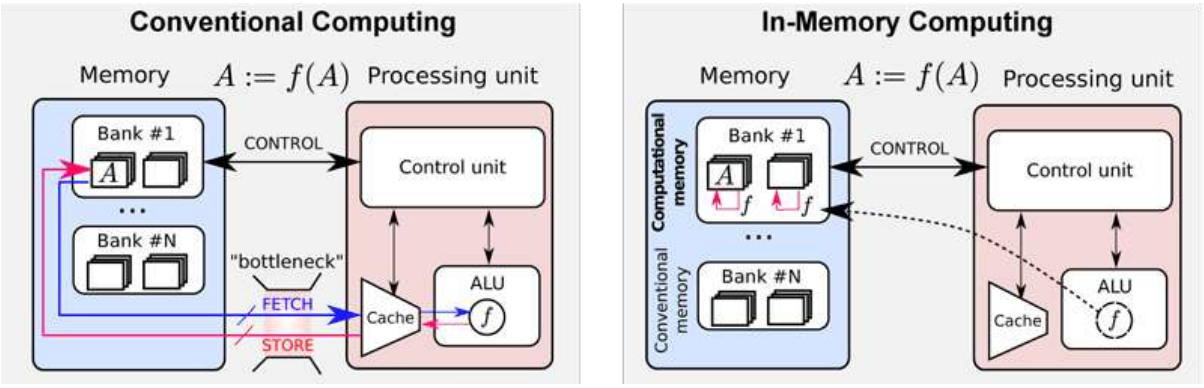


Figure 5: Comparison between a conventional computing architecture (left) and in-memory computing (right). Adapted from [4].

However, there are significant challenges that need to be overcome to enable the practical use of in-memory computing for AI algorithms. It is only possible to perform a limited set of operations in the memory units, as opposed to the general-purpose processors that are used in conventional computing systems which can perform any type of computation. Moreover, in-memory computing can only offer limited precision due to the analog nature of the operations performed in the memory units, as opposed to the usual digital computing which can offer arbitrarily high precision (typically 64-bit in conventional computers). These aspects imply a radical rethink in the way an algorithm needs to be designed to solve a certain problem efficiently using in-memory computing.

In the last few years, working towards overcoming those challenges by designing algorithms that can take advantage of in-memory computing hardware to efficiently solve AI related tasks. A prototype in-memory computing hardware based on phase-change memory comprising one million memory cells and aimed at understanding what type of algorithm can be implemented with the set of operations and precision of computation that are achievable with this chip has been developed. With this hardware, they were able to successfully demonstrate compression and reconstruction of images [1], unsupervised learning of temporal correlations [2], and solving systems of linear equations with arbitrarily high accuracy [3]. In each of these three applications, they have estimated the achieved energy savings of at least one order of magnitude with respect to using a conventional computer for the same tasks.

One additional application that has been of utmost interest in recent years is the training of neural networks. Deep artificial neural networks have shown remarkable human-like performance in tasks such as image processing and voice recognition. Deep neural networks are loosely inspired by biological neural networks. Parallel processing units called neurons are interconnected by plastic synapses. By tuning the weights of these interconnections, these networks are able to solve certain problems remarkably well. However, because of the need to repeatedly show very large datasets to very large neural networks, it can take multiple days or weeks to train state-of-the-art networks on conventional computing systems. In-memory computing could greatly accelerate the training of neural networks by eliminating the need to move the weight data back and forth between memory and processor.

The key idea in this approach is to encode the synaptic weights as the conductance values of phase-change memory devices organised in a computational memory unit and use it to perform the forward and backward propagation, while the weight changes are accumulated in high precision (see Figure 6). This mixed-precision approach enables training the network to reach high classification accuracy, while performing the bulk of the computation as in-memory computing. Figure 6 shows a simulation result of training a multi-layer perceptron to recognise handwritten digits. The accuracy achieved with their approach is less than one percent lower than that obtained using a conventional computer. Most importantly, the trained weights will be retained in the computational memory for many months or even years without the need to supply any power, thanks to the non-volatility of the phase-change memory devices. A chip trained in this way can be used for inference tasks within sensor devices at a fraction (<1%) of the power that would be used in a conventional computer. More details can be found in our recent paper presented at the ISCAS 2018 conference [4].

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

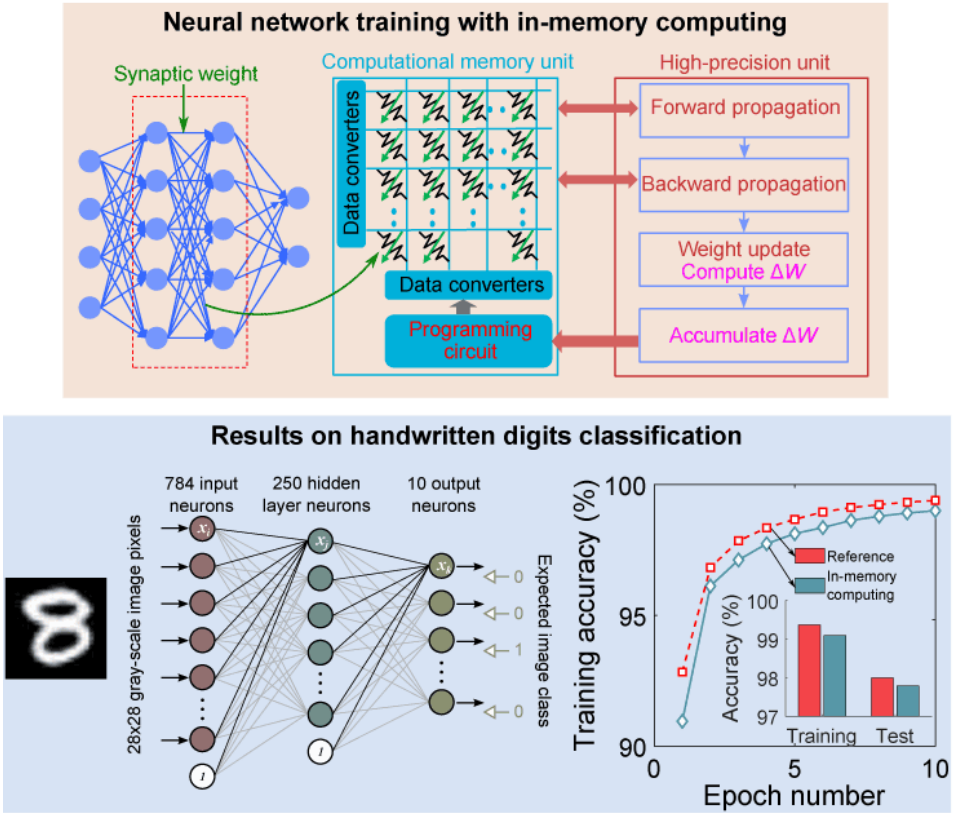


Figure 6: Neural network training algorithm using in-memory computing (top), and simulation results of training a multi-layer perceptron to recognise handwritten digits (bottom). Adapted from [4].

II. CONCLUSION

The rise of neuromorphic computing implies a paradigm shift in the design of very large-scale integrated circuits (VLSI), and this shift has substantial ramifications for a variety of different businesses. This novel technique, which was motivated by the workings of the human brain, possesses a significant amount of potential for the development of real-time, energy-efficient, and adaptable computing systems. Despite the fact that there are obstacles to overcome, there are enormous opportunities for VLSI designers to contribute to the development of neuromorphic computing. We may anticipate substantial discoveries in artificial intelligence (AI), robots, healthcare, and other fields as this sector continues to expand, which will revolutionize the way in which we engage with technology and benefit from it.

In-memory computing (IMC) represents a transformative approach to addressing the growing energy demands of artificial intelligence workloads. By bringing computation directly to the location of data storage, IMC significantly reduces data movement — a primary source of energy consumption in conventional von Neumann architectures. This paradigm shift enables orders-of-magnitude improvements in energy efficiency, latency, and throughput, making it particularly well-suited for AI applications at the edge and in data centres alike. As AI models continue to grow in complexity and deployment scales up, IMC stands out as a critical technology for achieving sustainable, high-performance AI. Continued advancements in materials, circuit design, and algorithm co-optimization will be key to unlocking the full potential of in-memory computing in the AI era.

III. FUTURE OUTLOOK

As we move toward more advanced forms of general artificial intelligence (General AI), the limitations of traditional computing architectures become increasingly apparent. General AI systems require massive amounts of data processing, adaptability, and real-time learning capabilities — all of which place significant demands on energy efficiency and hardware performance.

In-memory computing (IMC) offers a compelling solution to these challenges. By minimizing data movement and integrating computation directly within memory units, IMC architectures can dramatically reduce energy consumption while enabling faster, parallel processing — both of which are essential for the scalability and responsiveness of General AI.

Looking ahead, the integration of IMC with neuromorphic computing, online learning algorithms, and brain-inspired architectures could accelerate the development of truly adaptive and autonomous AI systems. Emerging memory technologies such as ReRAM, PCM, and FeFETs will play a pivotal role in building flexible and efficient hardware capable of supporting the dynamic workloads of General AI.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

To fully realize this potential, future efforts must focus on algorithm-hardware co-design, robust system integration, and the creation of standardized development frameworks. If successful, IMC could become a foundational technology in the pursuit of sustainable, scalable, and intelligent general-purpose AI systems.

As AI models grow and edge AI expands, IMC is positioned as a key technology to achieve orders-of-magnitude improvements in energy efficiency. Especially with the rising demands of on-device intelligence, IMC could enable real-time AI without cloud dependency.

REFERENCES

1. *M. Le Gallo et al.: "Compressed sensing recovery using computational memory", in Proc. of the IEEE International Electron Devices Meeting (IEDM) 2017.*
2. *Sebastian et al.: "Temporal correlation detection using computational phase-change memory", Nature Communications 8, 1115, 2017.*
3. *M. Le Gallo et al.: "Mixed-precision in-memory computing", Nature Electronics 1, 246-253, 2018.*
4. *S.R. Nandakumar et al.: "Mixed-precision architecture based on computational memory for training deep neural networks", in Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS), 1-5, 2018.*
5. *https://semiengineering.com/increasing-ai-energy-efficiency-with-compute-in-memory/*
6. *Ju, et al., "A General-Purpose Compute-in-Memory Processor Combining CPU and Deep Learning with Elevated CPU Efficiency and Enhanced Data Locality," 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2023, pp. 1-2, doi: 10.23919/VLSITechnologyandCir57934.2023.10185311.*
7. *https://vlsifirst.com/blog/ascent-of-neuromorphic-computing-and-its-implications-for-vlsi-design*
8. *Designing of Neuromorphic VLSI Circuits based on Biological Neural Networks to Improve the Energy Efficiency and Performance of AI/ML Applications by S. China Venkateswarlu, Krishna Dharavath, M. Koti Reddy, Narsimha Reddy Kuppireddy, Rajendar Sandiri, Srinivsarao Gajula, Nagarjuna Malladhi, Sreevani Menda, Vallabhuni Vijay in "Communications on Applied Nonlinear Analysis" Volume 32 No. 10(2025).*
9. *Gupta S, Kumar P, Paul T, van Schaik A, Ghosh A, Thakur C S, (2019) "Low Power, CMOS-MoS2 Memtransistor based Neuromorphic Hybrid Architecture for Wake-Up Systems". Scientific Reports (Nature Group)*
10. *S.Gupta, P. Kumar, K. Kumar, S. Chakraborty, Thakur, C. S, "Low Power Neuromorphic Analog System based on Sub-Threshold Current Mode Circuits".IEEE International Symposium on Circuits and Systems (ISCAS), 2019.*
11. *Nazreen P.M., S. Chakrabartty S., Thakur C.S. "Multiplierless and Sparse Machine Learning based on Margin Propagation Networks". arXiv:1910.*
12. *https://vlsiweb.com/vlsi-design-for-ai/*