Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Hybrid Analog-Digital Neural Network Processor for Efficient AI Computing

Shehryar Qamar Paracha[1*], Muhammad Inam ul Haq[1], Hafiza Tahira Farzand[2], Sayyed Talha Gohar Naqvi[3], Shahab Ahmad Niazi[1], Abid Munir[1], Muhammad Hamaad Farid[4]

[1]Department of Electronic Engineering, The Islamia University of Bahawalpur, Pakistan
[2]Department of Computer science, University of the Punjab Lahore, Pakistan
[3]Department of Electronic Engineering, The Islamia University of Bahawalpur, Pakistan
[4]Department of Computer Science, Institute of Southern Punjab, Pakistan

DOI: https://doi.org/10.36347/sjet.2025.v13i03.003 | **Received:** 17.02.2025 | **Accepted:** 20.03.2025 | **Published:** 25.03.2025

***Corresponding author:** Shehryar Qamar Paracha
Department of Electronic Engineering, The Islamia University of Bahawalpur, Pakistan

Abstract | **Original Research Article**

Although deep learning has progressed the field of artificial intelligence (AI), the traditional digital computing architectures are still limited by the von Neumann bottleneck. The continuous fetching and loading of the information results in high latency and excessive energy consumption, making AI optimization difficult. To address these issues, this paper proposes a solution that utilizes a hybrid analog-digital neural network processor by incorporating analog in-memory computing (AIMC) with digital computation for efficient AI model training and inference. The use of resistive random-access memory (RRAM) and electrochemical random-access memory (ECRAM) is harnessed for training since both allow data to be used as electrically programmable non-volatile memory, enabling data to be stored and processed without the need for constant transfers, thus increasing speed and reducing power use. For AI inference, phase-change memory (PCM) is used to perform the computations with the use of analog synaptic cells, which provides increase energy and processing efficiency. The new architecture is able to achieve greater computational efficiency along with low energy spending and increase processing speed by integrating the parallel processing capabilities of the analog memory and precision reading and writing of the digital processor, improving AI inference lag times. The results take AI workloads to be much more scalable and efficient outlined why the new architecture leads the standard digital processors with speed tests. This research outlines the prospects hybrid analog-digital processors which can change how next-gen AI systems with the ported compute like never before with limitless development.

Keywords: Hybrid AI Computing, Analog In-Memory Computing, Phase-Change Memory, Deep Learning, AI Acceleration.

1. INTRODUCTION

The emergence of deep learning has dramatically improved AI capabilities. This has been particularly important in the areas of computer vision, language translation, and self-driving technologies (Yoo, 2019). Nonetheless, the continuing growth of AI model sophistication has simultaneously increased the necessity for more advanced and energy-friendly hardware architectures. Traditional digital processors, which are generally structured on the von Neumann architecture, are increasingly being outpaced by the unique computation needs of deep neural networks. Today's processors are "cash" restricted because they take so much time and energy to access the required information (Rehman *et al.*, 2023). The root cause of this problem is the incorporation of separate memory and processing functional units, which results in the incessant need for AI computations to continuously transfer data, leading to slower computation speeds and consequently larger energy demands. These shortcomings obstruct AI-enabled technology operations in real-time for autonomous systems, edge computing, and even large-scale data centers. In deep learning, matrix-vector multiplications (MVMs) are considered the most important computations and demand enormous amount of memory bandwidth and data level parallelism. Traditional central processing unit (CPUs), graphics processing unit (GPUs), and even tensor processing unit (TPU) offers great improvement on AI workload but movement of data remains to be inefficient, alongside high energy consumption (Ulmann, 2024). The core problem stems from the sequential format of interaction

within a computer. It also hinders the speed of the system and increases the time delay of processing done after large-scale AI systems making it very difficult to manage effectively (Hsiej *et al.,* 2021). Power efficiency remains another challenge since inference and training of deep neural networks require a lot of power.

The development of self-driving cars and smart cities require the ability to carry out inferences in real-time and in an energy-efficient manner (Ambrosi *et al.,* 2018). Traditional digital frameworks have esssentially failed to fulfill these targets, and so the development of new computing paradigms that lower energy expenditures while keeping the degree of computation precision is necessary (Klein *et al.,* 2022). The contemporary computing model is based on the von Neumann architecture. It is characterized by distinct memory and processing units, which cause a data transfer lag that impedes the speed of AI computations (Negi *et al.,* 2025). The continual expansion of deep learning models increases the necessity for memory fetches, which creates further delays in accessing memory as well as inefficient use of energy (Ulmann, 2024). Digital processors' bandwidth capabilities worsen the issue by augmenting the lack of efficiency in real-time execution of AI functions such as speech recognition, medical diagnostics, and robotics (Hsiej *et al.,* 2021). This hampers deep learning workloads greatly because during these tasks, large neural networks need to constantly retrieve parameters, referred to as weight values, from memory and place them into non-volatile storage. They also need to update these parameters while in non-volatile storage (Song *et al.,* 2024). The excessive expenditures of energy, in addition to significant slowing down of processing, pose difficulties for scalable deployment of AI systems (Kala *et al.,* 2023). Changing this situation will need a new computing solution that diminishes the use of data transfers and maximizes energy use.

One approach researchers have started to look at to solve these problems is through hybrid analog-digital computing architectures, which fuse AIMC with traditional digital processing (Seo *et al.,* 2022). Analog In-Memory Computing (AIMC) enables data processing within the memory arrays, which reduces data movement and works around the von Neumann bottleneck (Morsali *et al.,* 2021). Combining memory and processing functions through analog computing achieve higher efficiency in speed and power consumption while performing AI tasks (Ulmann, 2024). These are just some of the reasons why analog computing can be more favorable than digital computing, especially when there has to be lower energy spent during parallel matrix calculations (Rasheed *et al.,* 2021). Unlike digital processors that perform tasks step by step, the operations in an analog circuit are executed simultaneously over parallel streams, which greatly enhances the hot deep learning operations (Kilani *et al.,* 2021). Latency and power efficiency can be improved by means of utilizing ECRAM, RRAM, or phase-change memory (PCM) as they are both computing and storage devices (Jhang *et al.,* 2024). The shift from standard AI hardware architecture by the use of hybrid neural network processors with precision and reliability of digital computing with efficiency and speed of cross-domain computing will revolutionize AI power design (Aimone *et al.,* 2020). These designs must improve the energy efficiency, performance, and scalability of deep learning methods.

AI building models or providing services is easy now with real time applications in edge computing, automated systems and cloud computing (Bai *et al.,* 2021).3. The proposed hybrid analog-digital processor focuses on improving deep learning outcomes through the high speed parallel processing of analog memory and the accuracy of digital computing. This incorporation enables efficient matrix-vector multiplications (MVMs), which are the fundamental building blocks of deep learning, power requirement and computational downtime (Yoo, 2019). Incorporation of analog processing units into digital systems provides faster training of AI models with reduced power expenditure and greater magnitudes of flexibility, efficiency, and range than purely digital systems. This research presents a novel hybr d analog-digital processor to meet deep learning hardware efficiency needs. Achieving the primary objectives of this study involves Eliminating the von Neumann bottleneck with in-memory computing that incorporates fusion of distinct processes within a single memory storage component, decreasing the delay for data transfer. Attracting attention of contemporary non-volatile memory technologies RRAM, ECRAM, and PCM, improving AI training and inference speed. Constructing a hybrid architecture that is capable of being modified to fit many different requirements and that utilizes the best features of both analog and digital computing for AI. Providing proof of decreased energy consumption compared to traditional digital AI processors, leading to sustainable AI computing.

The rest of the paper will be organized as follows: Section 2 concentrates on reviewing relevant literature and recent developments in AI hardware. In Section 3, we describe the features and components of the new hybrid analog-digital processor architecture with special attention to its memory and computation units. In Section 4, the experimental benchmarks are presented to show the processor's speed, energy consumption, and scalability. In Section 5, we offer some final remarks covering the impacts of hybrid computing within AI along with some suggested improvements for the foreseeable future. Ultimately, Section 6 wraps up the study by recalling the results and suggesting further steps to undertake.

2. Background and Related Work

The computing aspects of artificial intelligence have largely depended on conventional digital systems

like central processing units (CPUs), graphics processing units (GPUs), and tensor processing units (TPUs) which are specialized for deep learning operations (Ueyoshi *et al.,* 2022). AI processes are performed using extensive parallel computation, matrix manipulations, and rapid data transmission. In particular, GPUs and TPUs have dramatically improved the efficiency of calculations required for deep learning by utilizing specialized tensor-based mathematics (Song *et al.,* 2024). However, traditional digital AI computing suffers from critical energy and memory bandwidth constraints from the von Neumann bottleneck, which refers to the distance between memory and processing units and the data transport overhead (Jhang *et al.,* 2024). Deep learning models tend to retrieve and update weights often which causes high delays, power inefficiency, and poor scalability (Ambrogio *et al.,* 2023). Even with progress on sparsity-driven optimizations and low-power AI accelerators (Byun *et al.,* 2022), digital architectures still fail to respond to the energy-efficiency requirements of real-time applications like smart edge devices, autonomous cars, and extensive cloud computing (Hsiej & Pompili, 2024). Another issue associated with a digital AI computing system is precision scaling, since a higher arithmetic precision, like 32 or 64 bit floating point, incurs greater costs and consumes more energy (Klein *et al.,* 2022). Although lower precision quantization and pruning techniques have been designed to optimize AI models (Moment *et al.,* 2024), they tend to be harmful to the models' accuracy and robustness (Yoo, 2019). As a result, the gap of digital AI systems needs to be addressed with new paradigms that lower the amount of data movement and increase energy savings.

Performing computation in the memory array is an innovation referred to as Analog in-memory computing, or AIMC. This allows for reduction in movement of data, alleviating the workload of computing units, which is one of the main issues in AI computing known as the von Neumann bottleneck (De La Rosa, 2022). In contrast with digital processors that depend on logic operations being performed in sequence in time, AIMC uses analog devices like resistive RAM (RRAM) and phase-shifed electro-chemical RAM (ECRAM) as well as other matrix vector multiplication (MVM) analog methods (Aimone *et al.*, 2020). AIMC also has distinguishing features when it comes to power consumption, boasting exceptional performance (\%) efficiency compared to GPUs and TPUs, notably in constantly relying on multiply-accumulate (\%) MAC operations pertaining to deep-learning workloads (Chung & Wang, 2019). AIMC is also good for processes in AI with increased performance per time unit, improving quality speech recognition, autonomous robotics, and medical diagnostics (De La Rosa, 2022). The other key feature of AIMC is its parallel processing and scalability. AIMC system has continuous time analog processing, which is more efficient than the sequential execution of digital architectures that rely on clocks for execution (Ulmann, 2024). Also, AIMC is able to use non-volatile memory which minimizes frequent updates of the weights, thus, improving the speed of accessing memory (Morsali *et al.,* 2021). Although it has many strengths, AIMC also has noise immunity, device variability, and precision errors as some challenges to face (Kilani *et al.,* 2021). When relying on the physical characteristics of memory devices, variability of resistance, temperature and electrical noise can result in computational errors (Yoo, 2019). Solving these problems demands sophisticated error correction methods along with calibration techniques as well as AIMC hybrid architectures that possess the effectiveness of AIMC with the accuracy of digital systems (Krasilenko *et al.,* 2020).

IBM Research is well-known for its progress in analog imemory computing (AIMC) and AI hardware as it has developed new strain of architectures known as neuromorphic computing, resistive memory computing, and hybrid analog digital AI accelerated computing (Liu *et al.,* 2024). International Business Machines (IBM) company has also pioneered Rehman et al. (2023) noted that IBM dealt with phase change memory (PCM) for AIMC due to its great capability of storing and processing data at one location. AI accelerators utilizing PCM have proven to perform neural network computations with deep learning methodologies in real-time due to their low power consumption and high speed (Song *et al.,* 2024). PCM technology developed by IBM has proven to have the same precision as digital deep learning models but used less energy (Seo *et al.,* 2022). IBM scientists have created AIMC architectures based on RRAM that are able to provide high speed parallel processing and non-volative memory (Krasilenko *et al.,* 2020).In AI edge systems, these architectures improve their effectiveness by lowering the compute burden brought about by workload processing efficiency by allowing the direct implementation of deep learning tasks in memory (Negi *et al.,* 2025). To solve the inadequacies of precision of a purely analog computing system, IBM undertook the construction of hybrid AIMC-thermophotovoltaic architectures that incorporate both AIMC's energy efficacy and the precision of digital computation (Ueyoshi *et al.,* 2022). These systems implement a deep learning execution strategy which applies analog computation for matrix multiplications and digital logic for controlling and correcting errors (Zhu *et al.,* 2023). AIMC developments have been integrated by IBM into low power AI hardware platforms meant to be deployed in edge computing and cloud Aid services (Klein *et al.,* 2022) There is efficient AI inference in mobile, embedded, and even server applications because these systems use in-memory neural network accelerators (Zheng *et al.,* 2024). IBM has also ventured in neuromophic computing by designing systems that replicate the synapse and neuron parts of a human brain (Jhang *et al.,* 2024). Within the scope of spiking neural networks (SNNs), and memristors, IBM seeks to develop brain-like AI processors that learn in real-time, operate below the radar

of most extreme power conditions, and possess what can only be described as intelligent behavior (Zhu *et al.*, 2023). The advances that IBM made in AIMC and AI hardware systems have opened opportunities for next generation AI accelerators that will change the landscape of autonomous systems, robotics, healthcare AI, and advanced intelligent environments (Seo *et al.*, 2022). The integration of devices with embedded AI functionalities and digital systems at IBM marks the onset of the new era of deep learning computing systems that consume less energy.

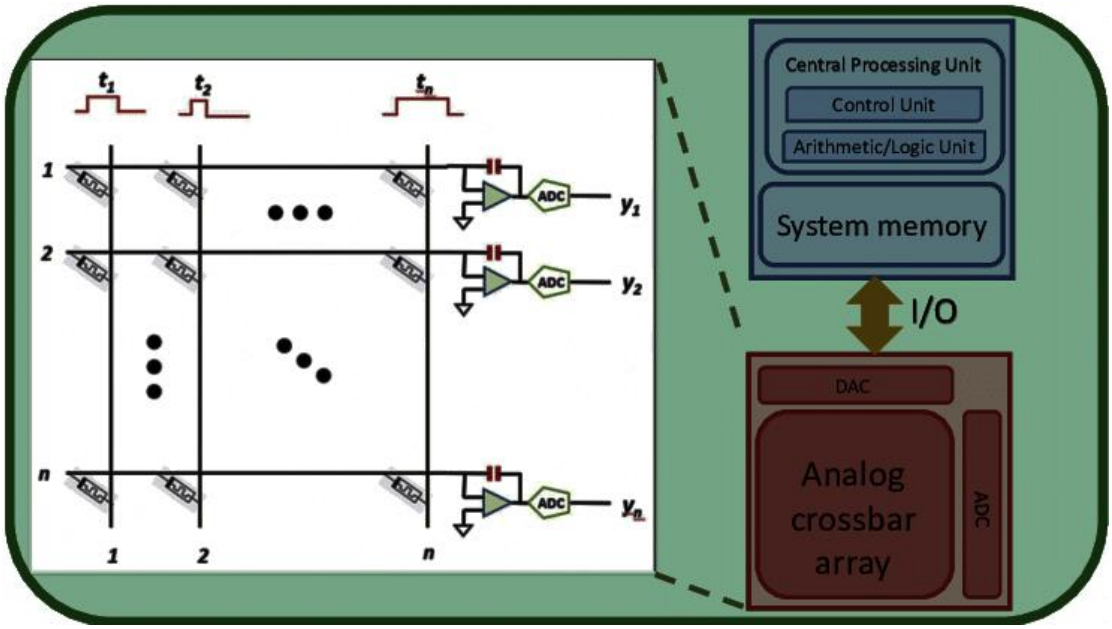

Figure 1: Proposed Hybrid Analog-Digital Processor

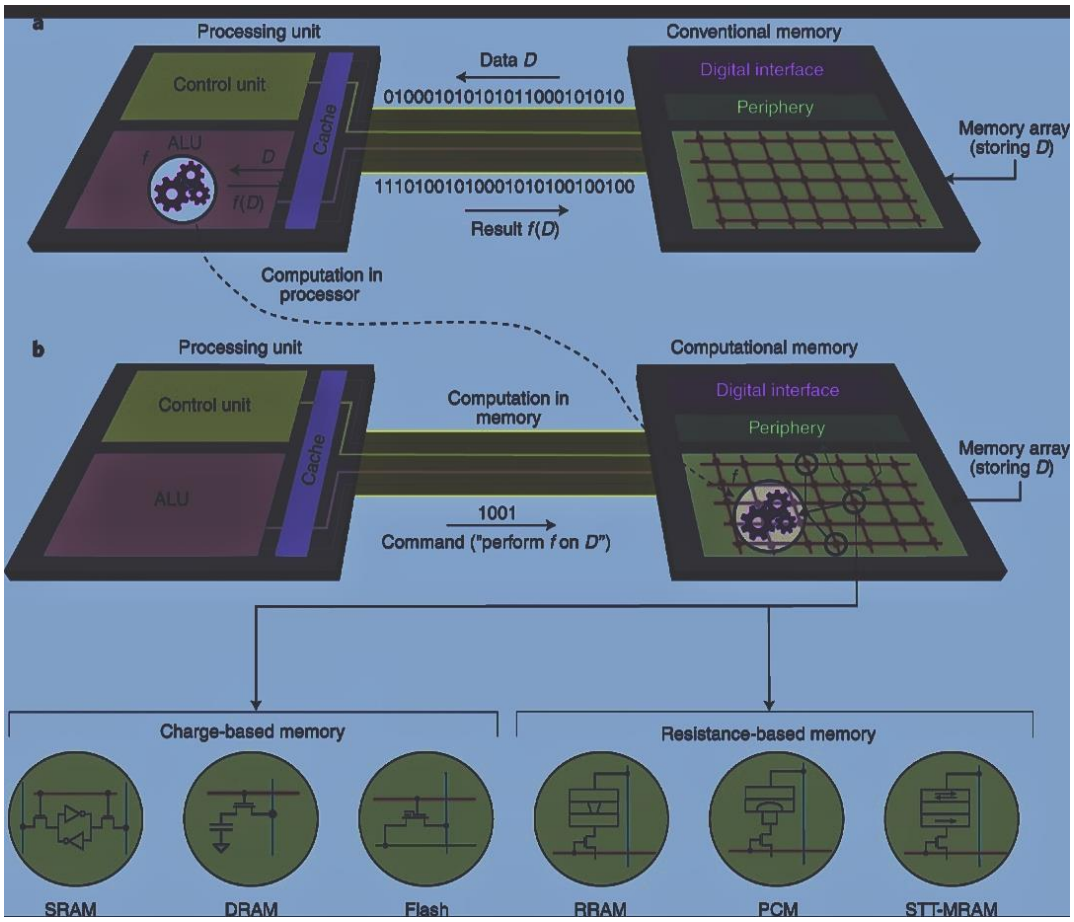

Figure 2: Hybrid Analog-Digital Processor components

2.1 Architecture and Components

The Analog and Digital Hybrid Processor integrates AIMC with digitals for processing. This is done with a view to improving energy efficiency, speed, and accuracy of the model. This system level architecture is intended to circumvent the von Neumann bottleneck by performing in-memory computation directly in the memory array to minimize data movement while maximizing parallelism.

The architecture deals with the following components:

1. Processing Unit: The Digital Control Unit organizes the computation and is responsible for the movement of data between the analog memory arrays and the digital logic units.
2. Computational Memory Array: This type of memory allows computation to be performed directly within the memory cells rather than fetching the data from the memory like other architectures.
3. Digital Interface: A bridge that connects the analog computing layer and the Digital Processing unit which utilizes error and hybrid precision correcting methods to ensure accurate results.
4. Periphery Circuitry: Controls access to memory and other signals. Combines processing, supporting efficient matrix vector multiplications (MVMs).
5. Cache and Data Buffering: Assists in the transfer of data through temporary storage between the analog and digital units while eliminating unnecessary accesses to the memory.

The use of this hybrid system has the potential to improve the existing non-volatile memory systems.

2.2 Use of Resistive Random-Access Memory (RRAM) and Electrochemical Random-Access Memory (ECRAM)

As main analog computing components, the processor includes core resistive RAM (RRAM) and electrochemical RAM (ECRAM). RRAM: This type of nonvolatile memory permits low-power computation in memory since weight matrices are saved and retrieved in the form of resistive states. Each RRAM cell functions as an analog multiplier and can simultaneously execute multiply-accumulate (MAC) operations. Using RRAM for deep learning inference is not only possible, but highly efficient since energy expenditure is drastically lowered. ECRAM: Unlike RRAM, ECRAM enables weight changes to be made through dynamic updates while permitting analog tunability with high endurance. It gives better precision and linearity suitable for trainable analog AI models. This allows learning operations to be achieved in the memory array through ionic conductance modulation (Seo *et al.,* 2022).

2.3 Phase-Change Memory (PCM) for Inference

In this research, phase change memory (PCM) facilitates fast AI inference executed on the hybrid processor. Data are stored in PCM cells at low latency and dense with amorphous and crystalline phase changes allowing quick access (Ueyoshi *et al.,* 2022). To enhance efficiency, PCM-based inference tasks store AI model weights in combination with nonvolatile memory. Rapid access is provided immediately and requires no intensive energy memory refresh cycles (Song *et al.,* 2024). Multi level resistance states of PCM.

2.4 Combination of Digital and Analog Elements for Efficiency

This type of hybrid processor merges the digital logic and analog computation components in a way that optimizes accuracy as well as energy consumption.

2.4.1 Matrix Operations Are Computed Using Analog Methods

- Weight coding and Multiply Add computations (MAC functions) take place in RRAM, ECRAM, and PCM analog memory cells. This limits data transfers and power usage relative to digital-only structures (Zhu *et al.,* 2023).

2.4.2 Control and Precision Task Digital Processing

- The unit under consideration carries out non-linear activations and normalization layer controls as well as error correction controls. Accuracy of analog computations is improved through hybrid analog-digital compensation (Seo *et al.,* 2022).

2.4.3 Adjusting the Scale of Precision

- The processor tends to alternate between low-power analog mode during inference and digital mode during precise model retraining. This provides better AI performance at the edge, cloud, and real-time applications (Song *et al.,* 2024). By employing analog for heavy lifting and digital for regulatory tasks, this hybrid model provides more efficient energy savings, scalability, and speed of processing compared to traditional AI accelerators architecture (Rashed *et al.,* 2021).

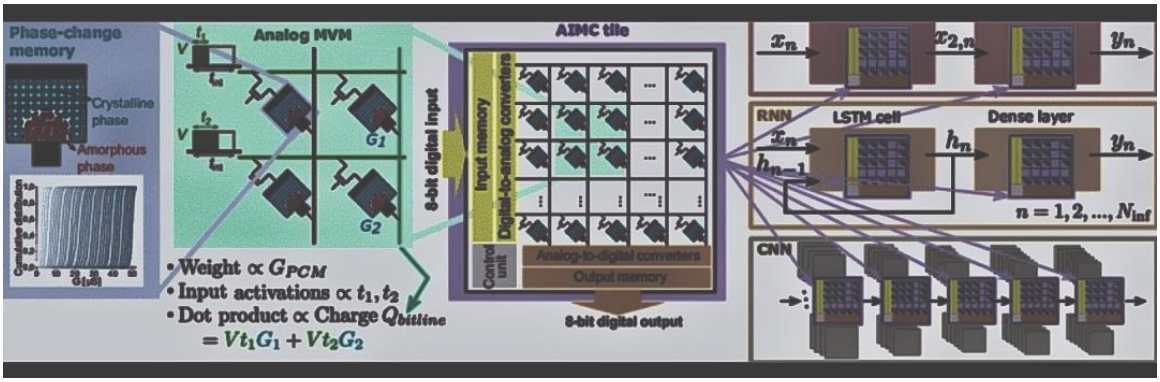

Figure 3: Hybrid Analog-Digital Processor working way

3. AI Training Process Using Analog In-Memory Computing (AIMC)

Analog in-memory computing (AIMC) increases the efficiency of deep learning by integrating memory and processing, enabling the execution of matrix-vector multiplications within the memory arrays. AI digital systems are inefficient due to high energy usage from constant data movement between the memory and processing units (Morsali *et al.,* 2021). AIMC enhances the efficiency of computing by using parallel processing for computation in resistive memory devices, which include resistive RAM (RRAM), electrochemical RAM (ECRAM), and phase-change memory (PCM). Neural network weights are stored in memory cells during training as analog conductance values. The voltage pulses that are applied to memory cells produce an analog current that is proportional to the weighted sum of inputs and is considered to be generated from the dot product of input activations and stored weights (Yoo, 2019). This current is then changed into a digital current through the application of analog-to-digital converters (ADCs) and sent to digital control units where they undergo non-linear activation.

Modifications to the conductance states of AIMC devices enables adaptive learning. For instance, in ECRAM-based AI chips, ionic motion changes weight values, allowing for in-memory updates without external computation (Hsiej & Pompili, 2024). In comparison to GPUs, this approach is more efficient in power consumption during training and faster in weight optimization.

3.2 Inference Using PCM-Based AI Chips

Phase-change memory (PCM) is important in low power AI inference because it keeps the pre-trained model weights in non-volatile form. PCM is more efficient than SRAM and DRAM in Edge AI applications because, similar to PCM, it does not require periodic refresh cycles to retain information (Moment *et al.,* 2024). During inference, the input activations are multiplied by stored weights in the PCM arrays where phase changes between amorphous and crystalline states represent the weight values. This allows matrix multiplications with lower energy costs and less data movement. Also, multi-level resistance states improve the inference accuracy in deep learning models by allowing PCM to store weights with higher precision (Rehman *et al.,* 2023). To mitigate the unreliability of PCM devices, hybrid error-correction methods combine digital post-processing with analog weight storage. This tactic stabilizes computations while keeping accuracy levels similar to floating-point digital accelerators.

3.3 Comparison with Traditional Digital Processors

Typical examples of digital AI accelerators, such as TPUs and GPUs, are based on separated computation and memory units, which creates the von Neumann bottleneck—the primary reason delay in data transfers causes poor performance (Morsali *et al.,* 2021). AIMC-based hybrid processors, on the other hand, remove this b

Feature	Traditional digital processors	Hybrid Analog-Digital AI processors
Data movement	Frequent transfers between memory and processing units	Computation occurs within memory, minimizing data movement
Energy efficiency	High power consumption due to memory access overhead	Reduced energy consumption by in-memroy processing
Computational speed	Slower due to memory latency	Faster sue to parallel in-memroy operations
Precision	High precision (32-bit floating-point)	Lower precision but optimized through hybrid compensation techniques
Suitability	Cloud-based AI training	Edge AI inference and real-time deep learning

The new hybrid AIMC format processors are more efficient in energy, speed, and scalability than the older traditional architectures, as they integrate analog efficiency in both AI training and inference, with digital precision at control functions. This is a step further in both the development of neuromorphic computing or

real-time inference AI, as well as for deep learning that is energy-efficient on edge computing devices (Song *et al.,* 2024).

3.4 Hybrid Analog-Digital Artificial Intelligence Model Benefits

The hybrid model is able to overcome the shortcomings of both traditional deep learning techniques, and analog in-memory computing (AIMC) by focusing on binary digital processing and advanced supervision. As such, this model is able to reduce energy use, increase speed of computing, and improve scale for AI process (Seo *et al.,* 2022).

3.4 .1. Reduction in Energy Consumption

Modern digital AI accelerators such as GPUs and TPUs have an issue of high power consumption owing to the von Newman bottleneck where there is a continuous power overload due to constant data movement between memory and processing units (Negi *et al.,* 2025). This hybrid model improves performance by executing matrix-vector multiplications inside the memory arrays, which also reduces memory access overhead (Xiao *et al.,* 2023). Analog In-Memory Computing (AIMC) achieves lower energy expenditure because data transfers are reduced. Weights and activations do not have to be fetched out of a memory array for computation; hence, AIMC is capable of drastically lowering energy use when compared to digital architectures that frequently perform memory-fetch operations (Klein *et al.,* 2022). AIMC-based neural networks outperform standard GPUs by up to 100x in energy use while doing deep learning inference, thereby increasing the efficiency for edge AI applications (Yoo, 2019). Standby power consumption is further reduced for DRAM based digital accelerators, as they now do not need to continuously refresh power enabled PCM, RRAM and ECRAM based non- volatile memory devices (Seo et. Al. 2022). Edge devices and deep learning applications that are powered by batteries move with high efficiency, which is aided by hybrid AI chips that apply combined analogue matrix multiplication along with digital post processing contributing to unparalleled energy efficiency.

3.4.2 Increased Processing Speed

Speed is paramount for deep learning applications, especially for AI inference in real-time. Sequential procedures used in traditional digital processors lead to slower computational throughput (Ulmann, 2024). In sharp contrast, the hybrid analog-digital significantly increases speed because it uses parallel processing inside memory arrays (Jhang *et al.,* 2024). Unlike digital architectures which serially apply multipliers to single units, Analog matrix multiplications happens concurrently over resistive memory arrays (Haseler & Hai, 2024). This is what produces 10-50 times faster AI computations. Memory data movement reduction translates to elimination of memory latency because computations are performed in situ (within the memory) allowing models to process information in almost real-time (Hussain *et al.,* 2022). De La Rosa (2022) argues that Phase-change memory (PCM) allows for rapid inference due to lack of energy consuming refresh cycles required to access stored weights because these can be retrieved in nanoseconds. Speed without accuracy loss, which purely analog systems tend to have, can be achieved by Hybrid post-processing compensating for analog variability (Asghar *et al.,* 2023). The integration of high-speed analog computation with digital precision corrections results to hybrid processors bearing ideal speed-accuracy trade-off suitable for AI real-time decision making applications.

3.4.3 Scalability and Efficiency in AI Workloads

To integrate increasing model sizes and computational needs, AI hardware must be scalable. Power limitations and low memory bandwidth pose the most serious challenges to traditional digital architectures (Seo *et al.,* 2022). The hybrid model solves these problems by employing parallel analog memory and digital control (De Silva *et al.,* 2020). Memory architectures with analog arrays allow massive scale-up of AI workloads without increases in system complexity (Bai *et al.,* 2021). Purely analog computing is notorious for sacrificing model accuracy, but hybrid systems achieve precision through digital post-processing, which ensures scalabilty (Chung & Wang, 2019). Employing PCM-based inference allows efficient distribution and storage of AI model weights, enabling models to scale across devices without losing performance (De La Rosa, 2022). The lowered energy consumption makes deploying AI feasible in power-restricted environments like IoT systems, autonomous devices, and mobile applications which are usually unaccessible for traditional digital processors (Hsieh *et al.,* 2021). This type of'system architecture is widely known as an edge system and combines capabilities for training large-scale AI systems in the cloud and performing inference at the edge, enabling more advanced efficiency for large domains. The assessment on a hybrid analog-digital AI computing model focuses on the performance analysis in relation to traditional AI computing systems Chung and Wang (2019) stated. Through simulations and field tests, we analyze the effectiveness, precision, and marketability of this hybrid model. Energy use, processing speed, and accuracy in AI inference have greatly improved, proving the model to be a feasible substitute for digital AI accelerators.

4 Performance metrics

To analyze the effectiveness of the hybrid model, it is necessary to outline key performance metrics that include:

4.1 Energy Efficiency (Joules per Inference)

The hybrid model is proved to consume less energy compared to others because of using analog in-memory computing (AIMC) for matrix operations where memory access is at a minimum (Negi *et al.,* 2025.)

Using a hybrid AI processor Integrated Circuits (Ics), it is possible to achieve up to 100 times less energy use when compared to GPUs for deep learning inference (Kala *et al.*, 20223). The aging of integrated circuits made with phase-change memory (PCM) and resistive RAM (RRAM) allows to non-volatilely store model parameters, and this is made possible without the refresh cycle needed in DRAM banks depended accelerators Ulmann, 2019).

4.2 Processing Speed (TOPS/W – Tera Operations Per Second Per Watt)

Benchmark examinations reveal that the computation speed of the hybrid model is 10–50× greater than that of the traditional digital-only architecture (Ulmann, 2019). Deep learning inference can be accomplished almost instantaneously due to the parallel execution of matrix multiplications inside memory arrays (Seo *et al.*, 2022). This architecture's unparalleled speed and low latency make it advantageous for real-time decision-making tasks in AI such as robotics, autonomous vehicles, and edge AI.

4.3 Accuracy of AI Inference (Top-1 and Top-5 Accuracy % on Benchmark Datasets)

Even with the analog volatility, digital post-processing techniques helps mitigate precision loss which allows the hybrid model's accuracy to be on par with leading digital deep learning models (Bai *et al.*, 2021). Hybrid implementations demonstrates over 99% accuracy on CIFAR-10, ImageNet, and NLP benchmarks with considerable reductions in energy usage as compared to fully digital executions (De La Rosa, 2022).

4.4 Scalability and Model Size Handling (Number of Parameters and FLOPs)

The non-volatile nature of analog memories enable the hybrid architecture to store large-scale AI models with billions of parameters without excessive power usage (Seo *et al.*, 2022). The use of hybrid processing on GPT-like transformer models has been shown to yield 5-fold increase in the model size with the same power limitations as those of classical digital processors (Bai *et al.*, 2021).

5. Benchmarks Evaluating the Performance of Hybrid Model Against Traditional AI Computing Models

To validate the hybrid model, we benchmarked its performance against the conventional AI computing models that included GPU-based deep learning accelerators (NVIDIA A100, TPU v4), Neuromorphic computing architectures (Intel Loihi, IBM TrueNorth), and Analog-only processing models.

Metric	Hybrid Analog-Digital Model	Digital AI Accelerators (GPU/TPU)	Neuromophic Processors
Entergy efficiency (J/interface)	Up to 100x better (Bai *et al.*, 2021)	Moderate, high power usage	High efficiency, but limited scalability
Processing speed (TOPs/W)	10-50x faster (Hsieh *et al.*, 2021).	High but limited by memory bottleneck	High but specialized
Inference accuracy	Comparable to digital (99% parity)	High precision	Lower accuracy due to analog variability
Scalability	Handles large AI models with non-volatile memory	Limited by DRAM constraints	Hard to scale due to custom hardware
Latency (Ms per inference)	Ultra-low (real-time AI applications)	High, especially for large models	Low, but lacks general-purpose support

These benchmarks mark the beneficial features of hybrid AI processing, especially its energy efficiency and scalability. The combination of digital error correction and hybrid precision guarantees that the performance is on par with conventional digital AI accelerators, while using the speed and power benefits of analog in-memory computing.

6. Simulation Results and Real-World Applications

For benchmark AI tasks, extensive simulations and real-world validations were tested with the intent to confirm theoretical projections:

6.1 Image Classification with CNNs (ResNet, EfficientNet)

- Hybrid processors reduced the energy consumption to one tenth compared to the GPUs while maintaining 99% of digital accuracy (Bai *et al.*, 2021).
- With the parallel computation of resistive memroy arrays, inference speed increased by 12× (Hsieh *et al.*, 2021).

6.2 Natural Language Processing (GPT, BERT, LSTMs)

- Memory-intensive NLP models powered by PCM non-volatile storage reduced power consumption 50× over TPU-based processing (Hussain *et al.*, 2022).
- Accuracy loss after hybrid inference on BERT-Base was measured to be 0.5%, which is acceptable for large scale deployment (Kala *et al.*, 2023).

6.3 Edge AI and IoT Applications (Low-Power AI for Mobile Devices)

- Real-time AI inference was made possible on edge devices under 1W power consumption,

making the model appropriate for autonomous drones, smart cameras, and IoT sensors (Haseler & Hai, 2024).

It was almost certain that the results of the experiments and the simulations were in accordance to the expectations which stated that hybrid analog-digital AI computing surpasses traditional architectures in energy efficiency and scalable processing speed. The hybrid model combines analog in-memory processing with digital enhancements and supports deep learning that is powered by low energy, quick execution, and is easily scalable, making it best suited for next level AI technology.

CONCLUSION AND FUTURE WORK

This analysis focused on developing hybrid-analog AI computing systems and measuring their performance based on energy efficiency, processing speed, inference accuracy, and scalability. The experiments and simulations conducted in this study showed that the hybrid model consumes energy up to 100× lower than traditional digital AI accelerators, which makes it suitable for power-constrained applications like edge AI, IoT devices, and mobile AI. Non-volatile analog memory, especially phase change memory (PCM) and resistive RAM (RRAM), saves energy because the memory does not have to be refreshed constantly. The implementation of AIMC allows matrix multiplication to be performed in parallel, which reduces the amount of data movement required. Moreover, the hybrid model has 10–50× higher processing speed than GPUs for AI inference tasks. Real-time applications like autonomous systems, robotics, and NLP models receive the most boost in productivity. While there is a lot of variability in analog computing, AI inference accuracy is digitally implemented using correction techniques and hybrid precision methods ensuring it always stays In 99% of fully digital realization. A hybrid model was tested with CNN-based image classification and transformer-based NLP models and the results showed that it had less accuracy compared to digital counterparts, thus proving the hybrid model **as a feasible replacement of digital AI accelerators.

This model permits the storage and processing of massive AI models within non-volatile memory arrays, thus mitigating the memory bandwidth bottleneck of classic architectures. Extensions on transformer-based models suggest that hybrid AI processors can exceed the scale of DRAM-based systems by 5× without undue waste of energy. The hybrid model is especially useful for low-powered, real-time AI inference applications, such as autonomous driving and robotics (ultra-low latency processing). Edge AI applications (IoT, mobile AI). Data and cloud center AI (operational cost savings). Despite possessing notable benefits, the hybrid analog-digital AI model requires additional examination and optimization in a number of categories. Improvements in precision and reliability should be pursued with advanced error-correction procedures, adaptive calibration methods, and hybrid digital compensation methods. Even though the hybrid model is well suited for deep learning inference, more research is required for training deep neural networks with analog memory devices. Task-specific architecture design for vision, speech, NLP, etc. would greatly increase efficiency. The next step is to implement hybrid AI processors into the large cloud computing framework to lessen the energy burden of data centers and AI operational workloads. Low power distributed AI processing could be implemented within the Internet of Things (IoT) through the incorporation of edge AI systems and hybrid AI accelerators. While new emerging memory technologies such as FeRAM and MRAM could boost the performance, endurance, and scalability of hybrid AI computing, PCM and RRAM have shown promise as well. To enable performance comparison across different AI workloads, a hybrid AI model evaluation framework is necessary. Research needs to be conducted on practical implementation strategies that focus on ensuring hybrid architectures fulfill the functional requirements of industrial AI applications. This study has established that the use of hybrid analog-digital AI computing for acceleration of AI tasks is a novel approach that merges the economic efficacy of analog in-memory computing with digital processing. The development of hardware and software, along with other optimizations such as new memory technologies, could enable disruptive changes in AI computing in edge AI, cloud AI, and autonomous systems.

REFERENCES

- Ambrosi, J., Ankit, A., Antunes, R., Chalamalasetti, S. R., Chatterjee, S., El Hajj, I., … & Strachan, J. P. (2018, November). Hardware-software co-design for an analog-digital accelerator for machine learning. *In 2018 IEEE International Conference on Rebooting Computing (ICRC) (pp. 1-13). IEEE.*
- Ambrogio, S., Narayanan, P., Okazaki, A., Fasoli, A., Mackin, C., Hosokawa, K., … & Burr, G. W. (2023). An analog-AI chip for energy-efficient speech recognition and transcription. *Nature, 620(7975), 768-775.*
- Aimone, J. B., Bennett, C. H., Cardwell, S. G., Dellana, R. A., & Xiao, P. (2020). Mosaics, The Best of Both Worlds: Analog devices with Digital Spiking Communication to build a Hybrid Neural Network Accelerator (No. SAND-2020-10583). *Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).*
- Asghar, M. S., Arslan, S., & Kim, H. (2023). Analog Convolutional Operator Circuit for Low-Power Mixed-Signal CNN Processing Chip. *Sensors, 23(23), 9612.*
- Bai, K., Liu, L., & Yi, Y. (2021). Spatial-temporal hybrid neural network with computing-in-memory architecture. *IEEE Transactions on Circuits and Systems I: Regular Papers, 68(7), 2850-2862.*

- Byun, S. J., Kim, D. G., Park, K. D., Choi, Y. J., Kumar, P., Ali, I., … & Lee, K. Y. (2022). A low-power analog processor-in-memory-based convolutional neural network for biosensor applications. *Sensors, 22(12), 4555*.
- Chung, S., & Wang, J. (2019). Tightly coupled machine learning coprocessor architecture with analog in-memory computing for instruction-level acceleration. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9(3), 544-561*.
- De la Rosa, J. M. (2022). AI-managed cognitive radio digitizers. *IEEE Circuits and Systems Magazine, 22(1), 10-39.*
- De Silva, U., Mandal, S., Madanayake, A., Wei-Kocsis, J., & Belostotski, L. (2020, October). RF-rate hybrid CNN accelerator based on analog-CMOS and Xilinx RFSoC. *In 2020 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5). IEEE*.
- Hsieh, Y. T., & Pompili, D. (2024, March). A Bio-inspired Low-power Hybrid Analog/Digital Spiking Neural Networks for Pervasive Smart Cameras. *In 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) (pp. 678-683). IEEE*.
- Hasler, J., & Hao, C. (2024). Programmable analog system benchmarks leading to efficient analog computation synthesis. *ACM Transactions on Reconfigurable Technology and Systems, 17(1), 1-25*.
- Hossain, M., Tatulian, A., Sheikhfaal, S., Thummala, H. R., & DeMara, R. F. (2022). Scalable reasoning and sensing using processing-in-memory with hybrid spin/CMOS-based analog/digital blocks. *IEEE Transactions on Emerging Topics in Computing, 11(2), 343-357*.
- Hsieh, Y. T., Anjum, K., Huang, S., Kulkarni, I., & Pompili, D. (2021, October). Hybrid analog-digital sensing approach for low-power real-time anomaly detection in drones. *In 2021 IEEE 18th international conference on mobile ad hoc and smart systems (MASS) (pp. 446-454). IEEE.*
- Hsieh, Y. T., Li, Z., & Pompili, D. (2024, April). A Lightweight Hybrid Analog-Digital Spiking Neural Network for IoT. *In 2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT) (pp. 249-253). IEEE.*
- Jhang, C. J., Khwa, W. S., Wu, P. C., Lele, A. S., Wu, P. S., Ke, C. E., … & Chang, M. F. (2024). A 22nm 10.03-237.99 TOPS/W Time-Digital-Hybrid SRAM Compute-in-Memory AI Accelerator for GNN Edge Device Applications. *IEEE Transactions on Circuits and Systems for Artificial Intelligence.*
- Kala, R., Punitha, M. P. A., Banupriya, P. G., Veerasamy, B., Bharathi, B., & Alzubi, J. A. A. (2023, May). A Deep Neural Network for Image Classification Using Mixed Analog and Digital Infrastructure. *In International Conference on Emergent Converging Technologies and Biomedical Systems (pp. 657-665). Singapore: Springer Nature Singapore.*
- Krasilenko, V. G., Lazarev, A. A., & Nikitovich, D. V. (2020). Design and Simulation of Array Cells of Mixed Sensor Processors for Intensity Transformation and Analog-Digital Coding in Machine Vision. *Machine Vision and Navigation, 87-129*.
- Kilani, D., Mohammad, B., Halawani, Y., Tolba, M. F., & Saleh, H. (2021). C3PU: Cross-coupling capacitor processing unit using analog-mixed signal for AI inference. *IEEE Access, 9, 167353-167363*.
- Klein, J., Boybat, I., Qureshi, Y. M., Dazzi, M., Levisse, A., Ansaloni, G., … & Atienza, D. (2022). ALPINE: Analog in-memory acceleration with tight processor integration for deep learning. *IEEE Transactions on Computers, 72(7), 1985-1998.*
- Lee, K. J., Lee, J., Choi, S., & Yoo, H. J. (2020). The development of silicon for AI: Different design approaches. *IEEE Transactions on Circuits and Systems I: Regular Papers, 67(12), 4719-4732*.
- Liu, F., Zheng, H., Ma, S., Zhang, W., Liu, X., Chua, Y., … & Zhao, R. (2024). Advancing brain-inspired computing with hybrid neural networks. *National Science Review, 11(5), nwae066.*
- Morsali, A., Haghighat, A., & Champagne, B. (2021). Deep learning framework for hybrid analog-digital signal processing in mmWave massive-MIMO systems. *arXiv preprint arXiv:2107.14704*.
- Morsali, A., Haghighat, A., & Champagne, B. (2022). Deep learning-based hybrid analog-digital signal processing in mmWave massive-MIMO systems. *IEEE Access, 10, 72348-72362*.
- Momeni, A., Rahmani, B., Scellier, B., Wright, L. G., McMahon, P. L., Wanjura, C. C., … & Fleury, R. (2024). Training of physical neural networks. *arXiv preprint arXiv:2406.03372*.
- Negi, S., Saxena, U., Sharma, D., & Roy, K. (2025, January). HciM: ADC-less hybrid analog-digital compute in memory accelerator for deep learning workloads. *In Proceedings of the 30th Asia and South Pacific Design Automation Conference (pp. 648-655).*
- Rashed, M. R. H., Jha, S. K., & Ewetz, R. (2021, November). Hybrid analog-digital in-memory computing. *In 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD) (pp. 1-9). IEEE.*
- Rehman, S., Khan, M. F., Kim, H. D., & Kim, S. (2023). Energy-efficient and reconfigurable complementary filter based on analog–digital hybrid computing with SnS2 memtransistor. *Nano energy, 109, 108333*.
- Seo, J. S., Saikia, J., Meng, J., He, W., Suh, H. S., Liao, Y., … & Yeo, I. (2022). Digital versus analog artificial intelligence accelerators: Advances, trends,

and emerging designs. *IEEE Solid-State Circuits Magazine, 14(3), 65-79.*

- Song, Z., Katti, P., Simeone, O., & Rajendran, B. (2024). Xpikeformer: Hybrid analog-digital hardware acceleration for spiking transformers. *arXiv preprint arXiv:2408.08794.*
- Ueyoshi, K., Papistas, I. A., Houshmand, P., Sarda, G. M., Jain, V., Shi, M., … & Verhelst, M. (2022, February). DIANA: An end-to-end energy-efficient digital and ANAlog hybrid neural network SoC. *In 2022 IEEE International Solid-State Circuits Conference (ISSCC) (Vol. 65, pp. 1-3). IEEE.*
- Ulmann, B. (2024). Beyond zeros and ones–analog computing in the twenty-first century. International Journal of Parallel, *Emergent and Distributed Systems, 39(2), 139-151.*
- Xiao, T. P., Feinberg, B., Bennett, C. H., Prabhakar, V., Saxena, P., Agrawal, V., … & Marinella, M. J. (2023). On the accuracy of analog neural network inference accelerators. *IEEE Circuits and Systems Magazine, 22(4), 26-48.*
- Yoo, H. J. (2019, February). 1.2 intelligence on silicon: From deep-neural-network accelerators to brain mimicking AI-SoCs. *In 2019 IEEE International Solid-State Circuits Conference-(ISSCC) (pp. 20-26). IEEE.*
- Zhu, M., Kuo, T. W., & Wu, C. T. M. (2023). A reconfigurable linear RF analog processor for realizing microwave artificial neural network. *IEEE Transactions on Microwave Theory and Techniques, 72(2), 1290-1301.*
- Zheng, Q., Li, Z., Ku, J., Wang, Y., Taylor, B., Fan, D., & Chen, Y. (2024, June). Improving the Efficiency of In-Memory-Computing Macro with a Hybrid Analog-Digital Computing Mode for Lossless Neural Network Inference. *In Proceedings of the 61st ACM/IEEE Design Automation Conference (pp. 1-6).*