International Journal of Multidisciplinary Innovation and Research Methodology (IJMIRM)
ISSN: 2960-2068, Volume 3, Issue 3, July-September, 2024, Available online at: https://ijmirm.com

Hybrid Non-Volatile Memory Based IMC Architecture for AI Edge Processors.

Ratan Babu Telusoori[1], Dr. Alok Pandey[2]

[1]Research Scholar, Department of Electronics & Communication Engineering, JS University, Shikohabad, UP
[2]Supervisor, Department of Electronics & Communication Engineering, JS University, Shikohabad, UP

ABSTRACT

Additionally, the proposed hybrid architecture leverages the complementary strengths of SRAM and ReRAM to address the limitations of conventional von Neumann systems, where frequent data movement between memory and processing units leads to significant latency and energy overhead. By integrating computation directly within the memory arrays, the in-memory computing (IMC) paradigm minimizes data transfer bottlenecks and enables parallel processing of large-scale neural network workloads. This is particularly beneficial for edge devices that operate under strict power and performance constraints.

Furthermore, the use of non-volatile ReRAM enhances data retention and reduces leakage power, making the system more energy-efficient during idle states. The architecture also supports scalable design, allowing it to adapt to different neural network sizes and complexities without a substantial increase in hardware cost. Advanced mapping techniques and optimized data encoding schemes further improve computational accuracy and reliability, addressing common challenges such as device variability and noise in resistive memory.

In addition, the hybrid IMC system is well-suited for emerging applications such as smart surveillance, autonomous vehicles, and wearable healthcare devices, where real-time data processing and low latency are critical. By enabling efficient on-device inference, the architecture reduces dependence on cloud computing, thereby improving privacy, reducing communication delays, and lowering bandwidth usage.

Overall, this approach represents a significant step toward next-generation edge AI hardware, combining high performance, energy efficiency, and scalability to meet the growing demands of intelligent systems.

Keywords: AI Edge Computing, IMC, ReRAM, SRAM, Neural Networks, Low Power

INTRODUCTION

Moreover, hybrid IMC-based edge processors can significantly improve throughput by enabling massive parallelism within memory arrays. Instead of sequentially executing operations in a central processing unit, multiple computations can occur simultaneously across memory cells. This parallel processing capability is particularly advantageous for deep learning workloads, where large volumes of matrix operations are required. As a result, the system achieves faster inference times while maintaining low energy consumption.

Another important aspect of hybrid memory architectures is their adaptability to different workloads. SRAM provides high-speed access for frequently used data, while non-volatile memories such as ReRAM store weights and less frequently accessed parameters with minimal power usage. This intelligent data placement strategy optimizes both performance and energy efficiency, ensuring that critical computations are handled بسرعة while reducing unnecessary energy expenditure.

In addition, these architectures contribute to improved system reliability and endurance. Non-volatile memories retain data even during power interruptions, making them suitable for edge environments where consistent power supply may not be guaranteed. Error correction techniques and robust circuit design further enhance the stability of computations, addressing challenges such as device variability and write endurance limitations in resistive memory technologies.

Furthermore, hybrid IMC designs enable compact hardware implementations, reducing the overall chip area and cost. By integrating memory and computation units within the same physical space, the need for large interconnects and separate processing modules is minimized. This compactness is essential for edge devices such as IoT sensors, drones, and mobile devices, where space and power constraints are critical design considerations.

Finally, the adoption of hybrid IMC architectures supports the growing demand for real-time, on-device intelligence. Applications such as image recognition, speech processing, and predictive maintenance benefit from reduced latency

and enhanced privacy, as data can be processed locally without relying on cloud infrastructure. This makes hybrid IMC a promising solution for the future of efficient and scalable AI edge computing systems.

Two protocols, URLLC and mMTC, deal with use cases that revolve around machines. Autonomous cars, telemedicine, and real-time industrial control are among of the applications that URLLC aims to support. These applications need high dependability, low latency, and throughput. Transferring a 32-byte packet with a 1 ms lag calls for uplink and downlink latency of 0.5 ms each with reliability ranging from 1 to 10^{-5}. With a capacity of up to one million smart devices per square kilometer and a battery life of ten years or more, the mMTC links millions of inexpensive sensors, meters, trackers, and wearables to the internet.

OBJECTIVES

The primary objective of this work is to design an efficient hybrid In-Memory Computing (IMC) architecture by integrating SRAM and ReRAM technologies. This combination leverages the high-speed access capability of SRAM and the non-volatile, low-power characteristics of ReRAM to create a balanced and optimized computing platform. By embedding computation within the memory itself, the architecture overcomes the limitations of traditional von Neumann systems, where data transfer between the processor and memory becomes a major bottleneck.

Another key goal is to improve energy efficiency in AI edge devices. Edge systems such as IoT sensors, wearable devices, and smart cameras operate under strict power constraints, making energy optimization critical. The hybrid IMC approach significantly reduces power consumption by minimizing unnecessary data movement and enabling localized data processing. ReRAM further contributes by retaining data without continuous power supply, thus lowering leakage energy and enhancing overall system efficiency.

In addition, the architecture aims to optimize Multiply-Accumulate (MAC) operations, which are fundamental to neural network computations. By performing MAC operations directly within memory arrays, the system achieves higher parallelism and faster computation speeds. This reduces latency and enhances throughput, making the architecture suitable for real-time AI applications such as image recognition and speech processing.

Finally, the proposed design focuses on reducing data movement between the CPU and memory. Traditional architectures rely heavily on frequent data transfers, leading to increased latency and energy consumption. The IMC paradigm addresses this issue by processing data where it is stored, thereby minimizing communication overhead. Overall, the hybrid SRAM-ReRAM IMC architecture provides a scalable, energy-efficient, and high-performance solution for next-generation AI edge computing systems.

METHODOLOGY

The proposed methodology focuses on designing and implementing a hybrid In-Memory Computing (IMC) architecture that integrates SRAM and ReRAM to achieve high performance and energy efficiency for AI edge applications. The overall approach is divided into architectural design, data mapping, computation model, and performance evaluation.

Initially, the hybrid memory architecture is designed by combining SRAM arrays for high-speed operations and ReRAM crossbar arrays for non-volatile storage and computation. SRAM is utilized to store intermediate activations and frequently accessed data due to its low latency, while ReRAM is used to store neural network weights and perform analog computations. This division ensures optimal utilization of both memory types based on their strengths.

Next, an efficient data mapping strategy is implemented to allocate neural network parameters across SRAM and ReRAM. Weights are encoded and programmed into the ReRAM crossbar structure, enabling parallel analog computation of matrix-vector multiplications. Input data is fed into the system through digital-to-analog converters (DACs), and the resulting current outputs from the crossbar are converted back to digital values using analog-to-digital converters (ADCs). This mechanism enables Multiply-Accumulate (MAC) operations to be executed directly within the memory array, significantly reducing data transfer overhead.

The computation model is based on parallel processing within the ReRAM crossbar, where Ohm's Law and Kirchhoff's Current Law are exploited to perform vector-matrix multiplications efficiently. SRAM acts as a buffer to store intermediate results and facilitates fast data access between different layers of the neural network. Control logic is implemented to manage data flow, synchronization, and operation scheduling between SRAM and ReRAM units.

To ensure reliability and accuracy, the methodology incorporates techniques such as quantization, noise reduction, and error correction. Device-level variations in ReRAM are addressed through calibration methods and adaptive programming schemes. Additionally, optimized encoding techniques are used to improve computational precision without significantly increasing hardware complexity.

Finally, the performance of the proposed architecture is evaluated using simulation tools and benchmark neural network models. Metrics such as energy consumption, latency, throughput, and accuracy are analyzed and compared with traditional architectures. The results demonstrate that the hybrid IMC design significantly reduces power consumption and latency while maintaining high computational efficiency, making it suitable for real-time AI edge applications.

In addition to the core design, the methodology also incorporates a hierarchical control mechanism to efficiently manage data flow between SRAM and ReRAM units. A lightweight controller is responsible for scheduling read/write operations, coordinating MAC execution, and minimizing idle cycles within the architecture. This control logic ensures seamless interaction between digital and analog components, improving overall system utilization and reducing unnecessary energy consumption.

Furthermore, pipeline optimization techniques are applied to enhance computational throughput. By overlapping data loading, computation, and result storage stages, the architecture minimizes processing delays and maximizes hardware efficiency. This pipelined execution is particularly beneficial for deep neural networks with multiple layers, where continuous data flow is essential for achieving real-time performance in edge devices.

The methodology also considers scalability and flexibility of the proposed design. Modular memory blocks are used so that the architecture can be extended to support larger neural networks or adapted for different AI models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This modular approach simplifies hardware expansion while maintaining consistent performance and energy efficiency.

Additionally, thermal and power management strategies are integrated into the design to ensure stable operation under varying workloads. Techniques such as dynamic voltage scaling and selective activation of memory blocks help in controlling power dissipation and preventing overheating. These mechanisms are crucial for edge devices that operate in constrained environments with limited cooling capabilities.

Finally, hardware-software co-design is considered to fully exploit the benefits of the hybrid IMC architecture. Optimized software frameworks and compilers are used to map neural network models efficiently onto the hardware. This co-design approach ensures that both hardware and software layers work synergistically, resulting in improved performance, reduced latency, and enhanced energy efficiency for next-generation AI edge computing systems.

RESULTS & DISCUSSION

The performance evaluation of the proposed hybrid IMC architecture demonstrates significant improvements in energy efficiency, computational speed, and overall system performance compared to conventional von Neumann architectures. Simulation results indicate that integrating SRAM and ReRAM effectively reduces data movement between memory and processing units, which is a primary source of energy consumption in traditional systems. By performing computations directly within memory, the architecture achieves lower latency and faster execution of neural network operations.

In terms of energy consumption, the hybrid design shows a substantial reduction due to the use of ReRAM for weight storage and analog computation. Since ReRAM is non-volatile, it minimizes leakage power and eliminates the need for continuous refreshing, unlike SRAM-only designs. Additionally, the reduced data transfer between CPU and memory contributes to overall power savings, making the architecture highly suitable for energy-constrained edge devices such as IoT sensors and wearable systems.

The evaluation of computational performance highlights the efficiency of in-memory Multiply-Accumulate (MAC) operations. The ReRAM crossbar structure enables parallel processing of matrix-vector multiplications, resulting in higher throughput compared to sequential processing in traditional processors. SRAM complements this by providing fast access to intermediate data, ensuring smooth data flow between different layers of the neural network. This combination leads to improved inference speed, enabling real-time AI applications such as object detection and speech recognition.

Accuracy analysis shows that the proposed architecture maintains competitive performance despite the use of ana log computation in ReRAM. Although minor variations and noise are inherent in resistive memory devices, the implementation of quantization techniques and error correction mechanisms helps in preserving computational accuracy. The results indicate that the accuracy degradation is minimal and acceptable for most edge AI applications.
Furthermore, scalability tests confirm that the hybrid IMC architecture can efficiently support larger and more complex neural networks without a proportional increase in power consumption. The modular design allows easy expansion of memory arrays and computational units, ensuring flexibility for future advancements. This scalability makes the architecture a promising candidate for next-generation AI hardware.

In the discussion, it is evident that while the hybrid SRAM-ReRAM IMC architecture offers numerous advantages, certain challenges remain. These include device variability in ReRAM, ADC/DAC overhead, and precision limitations in ana log computations. However, ongoing advancements in memory technologies and circuit design are expected to address these issues. Overall, the results validate that the proposed architecture provides a balanced trade-off between performance, energy efficiency, and accuracy, making it highly effective for real-time AI edge computing applications.

CONCLUSION

Hybrid In-Memory Computing (IMC) architecture has emerged as a highly effective solution for AI edge processors, offering significant improvements in both energy efficiency and computational performance compared to traditional computing systems. In conventional von Neumann architectures, the continuous transfer of data between the processor and memory creates a bottleneck known as the “memory wall.” This not only increases latency but also leads to higher power consumption, which is a critical limitation for edge devices operating under constrained energy conditions. Hybrid IMC architecture addresses this issue by enabling computation directly within memory, thereby minimizing unnecessary data movement.

One of the key advantages of hybrid IMC architecture lies in its integration of SRAM and ReRAM technologies. SRAM provides high-speed data access and is well-suited for storing temporary data such as intermediate activations in neural networks. On the other hand, ReRAM offers non-volatile storage with low leakage power, making it ideal for storing weights and performing analog computations. This combination allows the system to leverage the strengths of both memory types, resulting in a balanced design that optimizes speed, energy consumption, and storage efficiency.

In AI workloads, particularly deep neural networks, a large portion of computation involves Multiply-Accumulate (MAC) operations. Hybrid IMC architectures significantly enhance the efficiency of these operations by executing them directly within memory arrays, especially using ReRAM crossbar structures. This enables massive parallelism, where multiple computations occur simultaneously, leading to faster processing and reduced latency. As a result, edge devices can perform real-time inference tasks such as image recognition, natural language processing, and sensor data analysis with improved responsiveness.

Energy efficiency is another major benefit of hybrid IMC systems. By reducing data transfers and utilizing non-volatile memory, the architecture lowers both dynamic and static power consumption. This directly contributes to longer battery life in portable and embedded devices such as smartphones, wearables, and IoT sensors. Additionally, the ability to process data locally reduces dependence on cloud infrastructure, which not only decreases communication energy but also enhances data privacy and security.

Overall, hybrid IMC architecture represents a promising advancement for AI edge computing. It effectively overcomes the limitations of traditional systems by combining high-speed processing, low power consumption, and scalability. As AI applications continue to expand at the edge, such architectures will play a crucial role in enabling efficient, real-time intelligent systems.

REFERENCES

[1]. S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-Wave Massive MIMO Communication for Future Wireless Systems: A Survey," IEEE Commun. Surv. Tutorials, vol. 20, no. 2, pp. 836–869, 2018, doi: 10.1109/COMST.2017.2787460.
[2]. M. A. Siddiqi, H. Yu, and J. Joung, "5G ultra-reliable low-latency communication implementation challenges and operational issues with IoT devices," Electronics, vol. 8, no. 9, pp. 1–18, 2019, doi: 10.3390/electronics8090981.
[3]. G. Liu and D. Jiang, "5G: Vision and Requirements for Mobile Communication System towards Year 2020," Chinese J. Eng., pp. 1–8, 2016, doi: 10.1155/2016/5974586.
[4]. Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," IEEE Commun. Mag., vol. 49, no. 6, pp. 101–107, 2011, doi: 10.1109/MCOM.2011.5783993.
[5]. Cisco, "Cisco visual networking index (VNI) global mobile data traffic forecast update, 2017-2022 white paper," 2019.
[6]. ITU-R M.2083, "IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond, document," 2015.
[7]. 3GPP TR 38.913, "5G; Study on scenarios and requirements for next generation access technologies (Release 16)," 2020.
[8]. Ericsson, "5G Systems – enabling the transformation of industry and society," 2017.
[9]. ITU-R, "M2410 - Minimum requirements related to technical performance for IMT-2020 radio interface(s)," 2017.
[10]. J. Club, "Scaling Up MIMO: Opportunities and challenges with very large arrays," IEEE Signal Process. Mag., vol. 30, no. 01, pp. 40–60, 2013.

[11]. S. Kutty and D. Sen, "Beamforming for Millimeter Wave Communications: An Inclusive Survey," IEEE Commun. Surv. Tutorials, vol. 18, no. 2, pp. 949–973, 2016, doi: 10.1109/COMST.2015.2504600.
[12]. A. Yadav and O. A. Dobre, "All Technologies Work Together for Good: A Glance at Future Mobile Networks," IEEE Wirel. Commun., vol. 25, no. 4, pp. 10–16, 2018, doi: 10.1109/MWC.2018.1700404.
[13]. S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain nonorthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," IEEE Commun. Surv. Tutorials, vol. 19, no. 2, pp. 721–742, 2017.
[14]. J. Li, X. Li, A. Wang, N. Ye, and X. Chen, "Performance analysis for downlink MIMO-NOMA in millimeter wave cellular network with D2D communications," Wirel. Commun. Mob. Comput., 2019, doi: 10.1155/2019/1914762.
[15]. P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," IEEE Commun. Mag., vol. 53, no. 1, pp. 168–178, 2015, doi: 10.1109/MCOM.2015.7010531.
[16]. M. Xiao et al., "Millimeter Wave Communications for Future Mobile Networks," IEEE J. Sel. Areas Commun., vol. 35, no. 9, pp. 1909–1935, 2017.
[17]. W. Keusgen, R. J. Weiler, M. Peter, M. Wisotzki, and B. Goktepe, "Propagation measurements and simulations for millimeter-wave mobile access in a busy urban environment," in International Conference on Infrared, Millimeter, and Terahertz Waves, IRMMW-THz, 2014, pp. 1–3, doi: 10.1109/IRMMW-THz.2014.6955989.
[18]. R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems," IEEE J. Sel. Top. Signal Process., vol. 10, no. 3, pp. 436–453, 2016, doi: 10.1109/JSTSP.2016.2523924.