Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

Modeling ADC Energy and Area Trade-offs for Compute-In-Memory Accelerators

Dr. Ankit Rai
Department of Electrical Engineering, University of Middlesex, UK
a.rai@mdx.ac.uk

Dr. Priya Kaur
Department of Computer Science, University of East London, UK
p.kaur@uel.ac.uk

Abstract:

The need for energy-efficient and high-performance computational systems has led to the exploration of specialized architectures like Compute-in-Memory (CIM). CIM accelerators are designed to process data directly within memory arrays, significantly reducing data movement and improving energy efficiency. A critical component of CIM systems is the Analog-to-Digital Converter (ADC), which converts analog signals from memory into digital data for processing. This paper investigates the energy and area trade-offs of ADCs used in CIM accelerators. By exploring various ADC architectures and their integration with memory-based computing, we highlight the challenges and strategies for optimizing energy consumption and area efficiency in these systems.

Keywords: Compute-in-Memory (CIM), Analog-to-Digital Converter (ADC), Energy Efficiency, Area Efficiency, Accelerator Design, Memory Computing, Hardware Architecture, Low-Power Design

I. Introduction

The evolving computational landscape is driven by the need for efficient, high-performance systems capable of handling the ever-growing data and processing demands in areas like artificial intelligence (AI), machine learning (ML), and big data analytics. Traditional Von Neumann architectures face significant performance bottlenecks due to the limited bandwidth

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

between memory and processing units [1]. Compute-in-Memory (CIM) accelerators present an alternative architecture that reduces these bottlenecks by performing computations directly within the memory arrays. In such systems, data is processed within the memory cells, and the need for frequent data transfers to a central processor is minimized. This approach offers significant improvements in energy efficiency and speed. However, to bridge the gap between analog memory elements and digital processors, an essential component is the Analog-to-Digital Converter (ADC). ADCs perform the critical function of converting the analog values stored in memory into digital data that can be processed by digital circuits. The design of ADCs plays a pivotal role in determining the overall performance of CIM systems. Efficient ADCs are required to minimize both the energy consumption and the area footprint of CIM accelerators. Thus, the development of energy-efficient and area-optimized ADCs is fundamental to achieving the full potential of CIM architectures. Integrating Deep Neural Networks (DNNs) with emerging hardware technologies such as Compute-in-Memory (CIM)**,** Analog-to-Digital Converters (ADC)**,** and advanced energy-efficient and area-efficient designs is a key area of research and innovation aimed at addressing the challenges of traditional deep learning systems. These challenges include power consumption**,** computational complexity**,** and latency**,** especially when deploying DNNs in real-world applications like edge computing, autonomous systems, and Internet of Things (IoT) devices [2]. The need for efficient hardware architectures arises from the high computational demands of DNNs, particularly in terms of large-scale matrix multiplications, data transfer bottlenecks, and significant power consumption. Hardware solutions that incorporate CIM, ADC, and low-power designs aim to make DNN operations more efficient by reducing the energy cost and space requirements while enhancing computational throughput. The figure below shows the layout of Memory Architectures in N Accelerators.

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

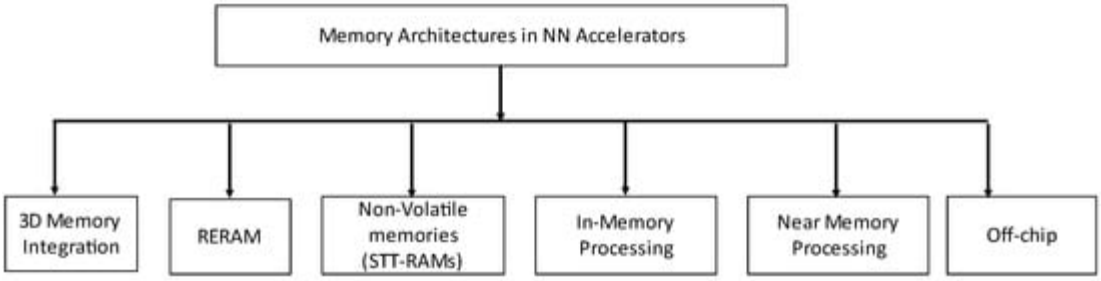

Figure 1 Classification of the memory architectures

This paper delves into the energy and area trade-offs of ADCs in the context of CIM accelerators, exploring how different ADC architectures can impact system-level performance. It presents an analysis of the various factors influencing ADC efficiency, such as resolution, speed, and power consumption [3]. Furthermore, it examines the challenges and opportunities associated with designing ADCs that are optimized for low-power and compactness in CIM applications. The paper aims to provide a comprehensive understanding of the key design considerations for ADCs in CIM accelerators. It also proposes strategies to achieve a balance between energy efficiency and area efficiency, contributing to the development of more effective and sustainable computing systems. This research highlights the importance of ADC design in advancing CIM-based architectures for future high-performance computing applications.

II. Importance of ADCs in CIM Accelerators

The Analog-to-Digital Converter (ADC) is a critical component in CIM accelerators, as it ensures the conversion of analog data from memory cells into digital form that can be processed by digital circuits. In a typical CIM architecture, memory cells store data in an analog form, which enables the computation to be performed directly on the stored values. However, most digital processors, which operate on binary data, cannot directly process these analog signals. Hence, ADCs are used to convert the analog data into digital signals, making it possible to integrate memory and processing in a more efficient manner. The efficiency of ADCs in CIM systems has a direct impact on both the speed and power consumption of the overall system. For example, high-speed ADCs are required when the CIM system needs to handle fast data processing. Similarly, high-resolution ADCs are necessary when precision is critical [4].

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

However, both high-speed and high-resolution ADCs typically come with increased power consumption and larger area requirements, which can be detrimental to the overall efficiency of the system.

Therefore, optimizing ADCs for energy efficiency and area consumption is vital for ensuring the success of CIM accelerators. One of the key challenges in ADC design is to balance the trade-offs between speed, resolution, power consumption, and area. Achieving an optimal design requires careful consideration of the application's specific requirements and the constraints imposed by the physical hardware. For instance, in applications where speed is prioritized, faster ADCs may be required which will consume more power and occupy more area. In contrast, in applications where power efficiency is a primary concern, low-power ADCs with reduced resolution might be preferred. Another critical aspect of ADC performance in CIM accelerators is the sampling rate, which determines how frequently the ADC captures and converts the analog signals [5]. A high sampling rate allows for faster data processing but results in higher power consumption and greater area usage due to the need for additional components to handle the faster sampling. Conversely, a lower sampling rate can help reduce power consumption and area requirements, but it may limit the performance of the system, especially in high-speed applications.

In the context of CIM accelerators, ADCs must also be compatible with the memory architecture. The integration of ADCs with memory elements can pose design challenges related to noise, signal integrity, and timing. The ADC must be able to accurately sample and convert the analog data stored in memory without introducing errors that could compromise the overall computation. Additionally, the ADC's interface with the memory must be optimized to minimize delays and ensure that the conversion process does not become a bottleneck in the system. Furthermore, the scalability of ADC designs is another important consideration. As CIM accelerators are used in increasingly complex applications, the need for higher-resolution and faster ADCs will continue to grow. Therefore, ADCs must be scalable to meet the evolving demands of CIM architectures. This requires innovations in ADC design that allow for improved performance without proportionally increasing power consumption and area.

III. Energy Consumption in ADCs

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

Energy consumption is a key concern in the design of ADCs for CIM accelerators. Since ADCs are involved in the critical process of converting analog signals to digital form, their energy consumption directly impacts the overall efficiency of CIM systems. The total energy consumption of an ADC is influenced by several factors, including the resolution, speed, and conversion method [6]. Each of these factors contributes to the amount of power required during the conversion process, and finding the right balance is crucial for optimizing energy efficiency. One of the primary sources of energy consumption in ADCs is dynamic power, which is proportional to the frequency of operation and the number of bits in the resolution. Higher-resolution ADCs require more bits to represent the analog signal accurately, leading to increased power consumption. In addition, higher-speed ADCs, which operate at faster clock rates, consume more energy due to the need for quicker conversion times. As a result, designing low-power ADCs requires careful consideration of these trade-offs between resolution, speed, and energy consumption. To reduce energy consumption, various techniques have been proposed in ADC design. One such approach is voltage scaling, where the supply voltage is reduced to decrease power consumption. Lowering the supply voltage reduces both dynamic and static power consumption, but it may also reduce the speed and resolution of the ADC. Therefore, voltage scaling must be carefully balanced to maintain the desired performance levels while minimizing energy usage. Another strategy for reducing energy consumption in ADCs is clock gating, where the clock signal is selectively disabled during periods when the ADC is not active. This technique reduces the unnecessary switching activity within the ADC, leading to lower dynamic power consumption. However, clock gating must be implemented in a way that does not interfere with the ADC's timing requirements, which can be challenging in high-speed applications.

Power-efficient ADC designs also take advantage of approximate computing methods, where some level of error is tolerated to reduce energy consumption. These approximate ADCs sacrifice a small degree of accuracy to achieve significant reductions in power usage. This approach is particularly useful in applications where perfect accuracy is not required, such as certain types of machine learning or image processing tasks, where a small amount of error can be tolerated in exchange for significant power savings [7]. In addition to reducing energy consumption during the conversion process, optimizing the ADC's interface with memory

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

elements can also help minimize energy waste. For example, reducing the number of times data must be read from memory or the frequency of ADC sampling can contribute to lower overall energy usage. Efficient integration of the ADC with memory systems is therefore essential for achieving energy-efficient CIM accelerators.

Furthermore, advanced technologies such as memristive devices, which offer lower power consumption than traditional transistors, are being explored for use in ADC designs. Memristive ADCs could provide a promising path toward energy-efficient and high-performance ADCs for CIM accelerators, allowing for further reductions in energy consumption. Finally, the use of machine learning algorithms to dynamically adjust ADC parameters in response to workload demands offers a novel approach to reducing energy consumption. By optimizing the resolution and sampling rate based on real-time data, adaptive ADCs can ensure that the system uses minimal energy while maintaining performance, further contributing to the energy efficiency of CIM accelerators [8].

IV. Area Efficiency of ADCs

Area efficiency is another critical design consideration when it comes to ADCs in CIM accelerators. The area consumed by the ADC must be carefully balanced with its functionality to ensure that the overall system remains compact and efficient. CIM accelerators aim to minimize the physical footprint of the system while maintaining high performance and the ADC plays a significant role in determining how much space is required for the system's core functions. The area consumed by an ADC depends on several factors, including the resolution, speed, and architecture of the ADC. Flash ADCs, for example, require a large number of comparators and logic gates to achieve high-speed conversions, leading to significant area overhead. While Flash ADCs are well-suited for applications requiring high-speed conversion, they come at the cost of a larger physical footprint. In contrast, Successive Approximation Register (SAR) ADCs are more compact and consume less area due to their simpler architecture, making them more suitable for applications where space is a critical concern. The resolution of the ADC is directly related to the area it occupies. Higher-resolution ADCs require more bits to represent the analog signal accurately, which in turn increases the number of components required in the conversion process. This results in a larger area footprint for high-resolution ADCs. Conversely, lower-

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

resolution ADCs are more compact but may sacrifice accuracy, which is undesirable in some applications. Therefore, the challenge lies in designing ADCs that offer a reasonable resolution while minimizing area consumption.

To reduce the area of ADCs, various design techniques have been proposed. One such approach is the use of multi-bit or pipelined ADC architectures, where the conversion process is broken down into multiple stages to reduce the complexity of each stage. These multi-stage designs allow for a more efficient use of space while maintaining the desired performance. Additionally, techniques such as layout optimization and the integration of ADC components into smaller fabrication technologies can help reduce the overall area consumed by the ADC. Another strategy for improving area efficiency is the use of hybrid ADC designs, which combine the strengths of different ADC architectures to balance speed, resolution, and area. For example, a hybrid ADC might use a SAR ADC for low-resolution, low-power operation and switch to a Flash ADC for high-speed conversions when needed [9]. This adaptive approach can help optimize the area consumption of the ADC based on the requirements of the application.

Moreover, reducing the complexity of the ADC's supporting circuitry, such as reference voltage generators and sample-and-hold circuits, can also contribute to smaller area requirements. By minimizing the number of components needed to support the ADC's operation, designers can reduce the total area consumed by the ADC without sacrificing its functionality. The integration of advanced materials, such as memristors or nanoscale transistors, also offers promising opportunities for reducing the area of ADCs. These materials have the potential to enable more compact and efficient designs, leading to smaller and more area-efficient ADCs. In high-performance CIM systems, it is essential to ensure that the ADC does not become a bottleneck in terms of area. Therefore, careful consideration of both the ADC's functionality and its area requirements is necessary to maintain the overall efficiency of the system.

V. Trade-Offs Between Energy and Area

The relationship between energy consumption and area efficiency is central to the design of ADCs in CIM accelerators. Typically, these two factors are inversely related—optimizing for one often results in a trade-off that increases the other. For example, high-resolution ADCs tend

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

to consume more energy due to the increased number of bits required for accurate conversion, but they also require more area to accommodate the additional components needed for the conversion process. Conversely, low-resolution ADCs are more area-efficient but may result in higher energy consumption, as additional processing stages may be required to compensate for the reduced resolution [10]. This trade-off between energy and area is especially significant in CIM accelerators, where both energy efficiency and area optimization are critical for achieving high-performance computing with minimal power consumption. Designers must carefully balance these two factors to ensure that the ADC meets the performance requirements of the application while adhering to the physical and energy constraints of the system. One approach to addressing the energy-area trade-off is through adaptive ADC architectures, which dynamically adjust their resolution and sampling rate based on workload conditions. By adjusting these parameters in real-time, the ADC can maintain optimal performance while minimizing both energy consumption and area usage.

Adaptive systems allow for more flexible and efficient hardware utilization, enabling the CIM accelerator to respond to varying computational demands. Another strategy for optimizing both energy and area is the use of hybrid ADC designs. Hybrid ADCs combine the strengths of different ADC architectures, such as SAR ADCs and Flash ADCs, to achieve the desired balance between speed, resolution, energy, and area. These designs enable the system to switch between different ADC modes depending on the specific requirements of the task at hand, resulting in both energy and area savings. Furthermore, the use of advanced fabrication technologies, such as 3D integration or nanometer-scale transistors, can help mitigate the trade-off between energy and area. These technologies enable the integration of smaller, more power-efficient components, leading to reduced energy consumption without sacrificing area efficiency. As a result, it becomes possible to design ADCs that offer both low energy consumption and small area footprints, even at high resolution and speed. Lastly, machine learning techniques can be employed to dynamically optimize the ADC’s performance.

By using machine learning algorithms to predict workload conditions and adjust ADC parameters, the system can minimize both energy and area usage. This approach allows the CIM accelerator to adapt to changing computational demands, ensuring that the system operates efficiently at all times. The trade-offs between energy and areas are complex, and there is no

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

one-size-fits-all solution. However, with the right combination of design strategies and technologies, it is possible to achieve an optimal balance that meets the needs of the application.

VI. ADCs for High-Performance CIM Systems

High-performance CIM accelerators, particularly those designed for demanding tasks such as AI and ML require ADCs that can handle large volumes of data at high speeds without compromising energy efficiency or area efficiency. These systems often require ADCs that offer both high resolution and fast conversion times to meet the performance demands of these complex applications. The design of ADCs for high-performance CIM systems must therefore focus on achieving optimal trade-offs between speed, resolution, power consumption, and area [11]. High-resolution ADCs are essential in high-performance CIM systems because they enable accurate conversion of analog signals, which is crucial for tasks such as image processing, signal processing, and data analytics. However, high-resolution ADCs tend to consume more power and occupy more area, making them less ideal for energy-constrained systems. Therefore, balancing the resolution requirements with the power and area constraints is a significant challenge in the design of ADCs for high-performance CIM systems. In addition to high resolution, high-performance CIM systems require ADCs that can operate at high speeds to handle the rapid influx of data. Fast ADCs are required to ensure that the system can process data in real-time, particularly in applications such as AI, where large datasets need to be processed quickly. However, high-speed ADCs typically require more complex circuitry and higher power consumption. The challenge lies in designing ADCs that can achieve the necessary speed while maintaining low power consumption and small area. One approach to achieving high-speed ADCs is the use of parallel processing techniques, where multiple ADCs operate concurrently to increase the overall conversion rate.

This method can help improve the speed of data conversion without significantly increasing power consumption or area. Additionally, pipelined ADC architectures, where the conversion process is divided into multiple stages, can also help achieve higher speeds with reduced power and area consumption. Moreover, the integration of machine learning techniques in high-performance CIM systems offers a promising approach for optimizing ADC performance. Machine learning algorithms can be used to adjust ADC parameters dynamically based on the

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

workload, ensuring that the ADC operates at the optimal resolution and speed for the task at hand. This dynamic adjustment allows for better utilization of system resources, resulting in lower power consumption and area usage while maintaining high performance.

In high-performance CIM systems, the ADCs must also be designed to handle large-scale data processing without becoming a bottleneck in the system. This requires careful optimization of the ADC's interface with memory elements and processors, ensuring that the conversion process does not delay data processing. Fast, efficient communication between the ADC and other components is essential to maintaining high system performance. Additionally, the design of high-performance ADCs for CIM accelerators requires advanced fabrication technologies to ensure that the ADCs can operate at the required resolution and speed while occupying minimal area. These technologies enable the development of smaller, more power-efficient ADCs that can handle the demanding requirements of high-performance CIM systems. Finally, the application of low-power design techniques, such as voltage scaling and adaptive techniques, is critical in ensuring that the high-performance CIM systems remain energy-efficient. By using these techniques, high-speed and high-resolution ADCs can operate efficiently without consuming excessive power, helping to maximize the performance and sustainability of the system.

VII. Future Directions and Challenges

The field of ADC design for CIM accelerators is rapidly evolving, with numerous opportunities for improvement and innovation. As computing workloads continue to grow in complexity and scale, the demand for more efficient and powerful ADCs will increase. One of the main challenges will be the ability to scale ADC designs to meet the performance needs of next-generation computing applications, such as deep learning and large-scale data analytics. One promising direction is the use of advanced materials and fabrication technologies to develop more compact and energy-efficient ADCs. For example, the use of memristive devices and nanotechnology could enable the creation of ADCs that consume less power and occupy smaller areas, while maintaining high resolution and speed. These technologies have the potential to revolutionize ADC design, leading to more efficient CIM accelerators. Another important area of research is the integration of ADCs with other hardware components in a more seamless manner.

Submitted: 05/11/2024 Accepted: 20/12/2024 Published 23/12/2024f

As CIM accelerators continue to evolve, ADCs must work efficiently in conjunction with memory units, processors, and other specialized hardware to optimize performance [12]. This requires innovative interconnect and integration strategies to minimize latency and maximize throughput. Furthermore, the use of machine learning techniques in ADC design will likely play a crucial role in future developments. Machine learning algorithms can be employed to optimize ADC parameters, such as resolution, sampling rate, and power consumption, in real-time based on workload conditions. This adaptive approach will allow CIM accelerators to dynamically adjust to varying computational demands, ensuring that they remain energy-efficient and high-performing under all conditions. Additionally, as the demand for low-power computing increases, especially in edge computing and mobile applications, energy-efficient ADCs will become even more critical.

The ability to design ADCs that balance power, area, and performance in an energy-constrained environment will be a key challenge moving forward. Finally, there is a growing need for standardization in ADC design for CIM accelerators. As the field progresses, developing standardized frameworks and design guidelines will help ensure interoperability and allow for the more widespread adoption of CIM-based architectures. Standardization will also facilitate more efficient comparisons between different ADC designs, leading to faster innovation and adoption of best practices.

VIII. Conclusion

The design of ADCs for Compute-in-Memory (CIM) accelerators presents a unique set of challenges due to the need to balance energy efficiency and area efficiency with high performance. As computing workloads continue to grow, the demand for high-speed, high-resolution ADCs will increase. However, achieving the necessary performance while minimizing power consumption and area requires careful consideration of various trade-offs. This paper has explored the energy and area trade-offs in ADC design for CIM accelerators, highlighting the importance of optimizing ADC performance for energy efficiency, area efficiency, and system integration. By utilizing techniques such as adaptive architectures, hybrid designs, and advanced materials, it is possible to achieve significant improvements in the efficiency of ADCs used in CIM accelerators. Looking ahead, the integration of machine learning techniques, advanced

fabrication technologies, and low-power design strategies will play a crucial role in advancing ADCs for CIM accelerators. These innovations will enable the development of more efficient, scalable, and sustainable computing systems capable of meeting the growing demands of modern computational workloads.

REFERENCES:

[1] R. Chen, "Analog-to-Digital Converters for Secure and Emerging AIoT Applications," Massachusetts Institute of Technology, 2023.

[2] M. Caselli, P. Debacker, and A. Boni, "Memory Devices and A/D Interfaces: Design Tradeoffs in Mixed-Signal Accelerators for Machine Learning Applications," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 69, no. 7, pp. 3084-3089, 2022.

[3] T. Andrulis, R. Chen, H.-S. Lee, J. S. Emer, and V. Sze, "Modeling analog-digital-converter energy and area for compute-in-memory accelerator design," *arXiv preprint arXiv:2404.06553,* 2024.

[4] Z. Chen, "Energy Efficient and Accurate In-Memory Computing Accelerators for Machine Learning," Rice University, 2022.

[5] R. Chen, H. Kung, A. Chandrakasan, and H.-S. Lee, "A bit-level sparsity-aware SAR ADC with direct hybrid encoding for signed expressions for AIoT applications," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 2022, pp. 1-6.

[6] K. He, I. Chakraborty, C. Wang, and K. Roy, "Design space and memory technology co-exploration for in-memory computing based machine learning accelerators," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1-9.

[7] S. Huang, H. Jiang, and S. Yu, "Hardware-aware Quantization/Mapping Strategies for Compute-in-Memory Accelerators," *ACM Transactions on Design Automation of Electronic Systems,* vol. 28, no. 3, pp. 1-23, 2023.

[8] S. Ji, "ANALYSIS OF THE ENERGY VS. ACCURACY TRADEOFF IN THE 10T1C BITCELL-BASED IN-MEMORY ARCHITECTURE," 2023.

[9] A. Kaul *et al.*, "Co-Optimization for Robust Power Delivery Design in 3D-Heterogeneous Integration of Compute In-Memory Accelerators," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2024: IEEE, pp. 1-2.

[10] S. Seyedfaraji, S. Jager, S. Shakibhamedan, A. Aftab, and S. Rehman, "OPTIMA: Design-Space Exploration of Discharge-Based In-SRAM Computing: Quantifying Energy-Accuracy Trade-offs," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. 1-6.

[11] C. Wolters, X. Yang, U. Schlichtmann, and T. Suzumura, "Memory Is All You Need: An Overview of Compute-in-Memory Architectures for Accelerating Large Language Model Inference," *arXiv preprint arXiv:2406.08413,* 2024.

[12] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE circuits and systems magazine,* vol. 21, no. 3, pp. 31-56, 2021.