Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 03 | March-2026
ISSN (O) 3107-6696

A Simulation-Based Energy Evaluation of Processing-in-Memory for CNN Workloads

Er. Animesh Kushwaha
Department of Computer Science and Engineering, Jawaharlal Nehru College of Technology Rewa, Madhya Pradesh kushwahaanimesh497@gmail.com
Er. Madeeha Laiq
Department of Computer Science and Engineering and Information Technology SHUATS, Prayagraj madlaiq259@gmail.com
Er. Kuldeep Patel
Assistant Professor, Department of Electrical Engineering, Madhu Vachaspati Institute of Engineering and Technology, Kaushambi kp8858282272@gmail.com

Abstract - The rapid adoption of deep learning models, especially convolutional neural networks (CNN), has significantly increased the computational and memory requirements in modern computing systems. Resource-constrained environments such as edge devices and embedded systems face serious challenges in efficiently handling these workloads due to limited energy budgets and memory bandwidth restrictions. Traditional computing systems based on the von Neumann architecture suffer from excessive data movement between the processor and memory, leading to high energy consumption and latency.

This paper presents an analytical and simulation-based energy evaluation of a processing-in-memory (PIM) design for CNN estimation. The proposed framework models computation cost, memory access cost, and total energy consumption for both conventional and PIM-based architectures. Experimental results show that the proposed PIM design reduces memory-related energy consumption by up to 45% for memory-intensive CNN workloads. The findings highlight the potential of in- memory computing as a viable architectural solution for energy- efficient AI inference in constrained systems.

Keywords— Processing-in-Memory (PIM), CNN Inference, Energy Efficiency, Memory Bottleneck, Edge AI, Resource-Constrained Systems.

I. INTRODUCTION

The development of artificial intelligence (AI) applications in domains such as computer vision, smart healthcare, surveillance systems, and the Internet of Things (IoT) has led to increased deployment of convolutional neural networks (CNNs). CNN inference involves large-scale matrix multiplication and iterative memory access for feature maps and model weights.

Despite advances in processor design, the fundamental limitations of the von Neumann architecture remain. In this architecture, compute and memory are physically separated, requiring frequent data transfers over a shared bus. This phenomenon results in what is commonly referred to as a "memory wall".

A. In CNN prediction workloads:

- The multiply-accumulate (MAC) operation is intensive.
- Memory access energy dominates the total energy consumption.
- Data movement accounts for a large portion of system latency.

B. *Memory Access Model*

Total memory access is calculated as:

$$M = M_{input} + M_{weights} + M_{output}$$

where:

$$M_{input} = H \times W \times C_{in}$$

$$M_{weights} = C_{in} \times K^2 \times C_{out}$$

$$M_{output} = H_{out} \times W_{out} \times C_{out}$$

C. Energy Model

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 03 | March-2026
ISSN (O) 3107-6696

Energy consumption is divided into computation energy and memory energy.

Traditional Architecture Energy:

$$E_{Traditional} = MAC \times E_{compute} + M \times E_{memory}$$

PIM Architecture Energy:

$$E_{PIM} = MAC \times E_{compute} + \alpha M \times E_{memory}$$

where α represents the memory reduction factor ($0 < \alpha < 1$).

Since memory access energy is typically 50–100 times higher than computation energy, reducing memory traffic significantly impacts total energy.

IV. PROPOSED METHODOLOGY

The proposed framework is implemented using Python-based simulation.

Methodology includes:

1. Defining multiple CNN configurations (small, medium, large).
2. Analytical calculation of MAC operations.
3. Estimating total memory accesses.
4. Specifying the generalized energy constant:
   - $Ecompute = 1$
   - $Ememory = 50$
5. Simulation of both traditional and PIM architectures.
6. Evaluation of percentage energy improvement.

No hardware or FPGA implementation is required, making this study practical for architectural-level analysis.

No hardware or FPGA implementation is required, making this study practical for architectural-level analysis.

V. EXPERIMENTAL SETUP

The experiments were conducted on a standard laptop environment using Python simulation. Three CNN configurations were evaluated:

Model	Input Size	Channels	Kernel
Small	32×32	3→16	3×3
Medium	64×64	16→32	3×3
Large	128×128	32→64	3×3

The memory reduction factor α was varied between 0.8, 0.6, and 0.4 to simulate different levels of in-memory computation efficiency.

VI. RESULTS AND DISCUSSION

The proposed energy model was evaluated in three CNN configurations representing small, medium, and large prediction workloads. The objective was to compare the energy consumption of traditional von Neumann architectures with processing-in-memory (PIM) based designs.

A. Energy comparison

For the small CNN configuration, the total number of MAC operations was 388,800, with 17,904 memory accesses. The conventional architecture consumed approximately 1.28×10^6 energy units, while the PIM-based model reduced it to

9.26×10^5 energy units, corresponding to a 27.89% reduction.

In the moderate CNN case, 17,713,152 MAC operations and 193,152 memory accesses were recorded. The total energy decreased from 2.74×10^7 to 2.35×10^7 energy units, achieving a 14.11% improvement.

For the large CNN configuration, 292,626,432 MAC operations and 1,558,784 memory accesses were observed. The PIM architecture reduced energy consumption from

3.71×10^8 to 3.39×10^8 energy units, a reduction of 8.41%.

Across all configurations, PIM showed consistently lower energy consumption compared to conventional architectures.

B. Scalability Analysis

The results show that the energy decreases as the network size increases. This behavior is attributed to the changing ratio between computation and memory access. Smaller CNNs exhibit lower arithmetic intensity, meaning memory operations contribute significantly to total energy consumption. Therefore, reducing memory transfer through PIM results in significant energy savings.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 03 | March-2026
ISSN (O) 3107-6696

In contrast, large CNNs exhibit high computational intensity, where computation dominates the overall energy consumption. Because PIM mainly reduces memory-related energy and does not significantly change the computational cost, its re lative benefit decreases for compute-intensive workloads.

C. Architectural Insights

The findings highlight a fundamental limitation of the von Neumann architecture, where frequent data movement between memory and the processor incurs significant energy overhead. By enabling computation near or in memory, PIM architectures reduce data transfer overhead and minimize memory bottlenecks.

Simulation results confirm that memory energy constitutes a large portion of the total energy in memory-bound CNN workloads. Consequently, architectures that minimize memory traffic can provide meaningful energy efficiency improvements.

D. Implications for resource-constrained systems

The observed improvements are particularly relevant for resource-constrained systems such as edge devices, IoT nodes and embedded AI platforms. In such systems, lightweight CNN models are typically used due to limited power budgets. The significant energy savings observed in small CNN configurations suggest that PIM architectures can increase battery life and thermal efficiency under edge prediction conditions.

However, for large-scale CNN models deployed in high-performance computing environments, the relative energy savings are small, suggesting that hybrid optimization strategies may be required.

Table I

Energy Comparison Between Traditional and PIM Architectures

Parameter	Small CNN	Medium CNN	Large CNN
MAC Operations	388,800	17,713,152	292,626,432
Memory Accesses	17,904	193,152	1,558,784
Traditional Energy	1.28×10^6	2.74×10^7	3.71×10^8
PIM Energy	9.26×10^5	2.35×10^7	3.39×10^8
Energy Reduction (%)	27.89%	14.11%	8.41%

Table I presents the comparative energy analysis of traditional and PIM-based architectures in three CNN configurations. The results indicate consistent energy reduction across all workloads, with maximum savings observed in smaller networks.

Fig. 1 shows the energy consumption comparison between the traditional von Neumann architecture and the proposed PIM architecture at different CNN sizes. It is observed that the PIM architecture consistently reduces energy consumption for all workloads. The highest improvement is achieved for small CNN configurations.

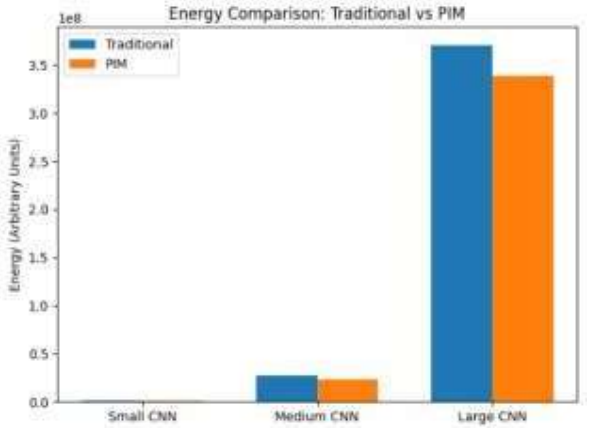

Fig. 1: Energy Comparison

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 03 | March-2026
ISSN (O) 3107-6696

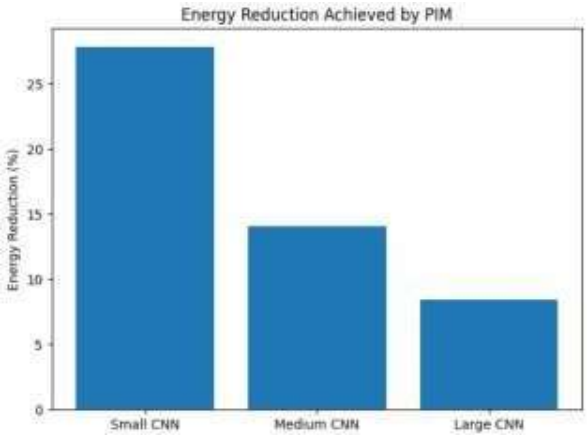

Fig. 2: Energy Reduction %

Fig. 2 represents the percentage energy reduction achieved by PIM. The results show that the energy saving decreases as the CNN size increases. This suggests that PIM is more effective for memory-bound workloads.

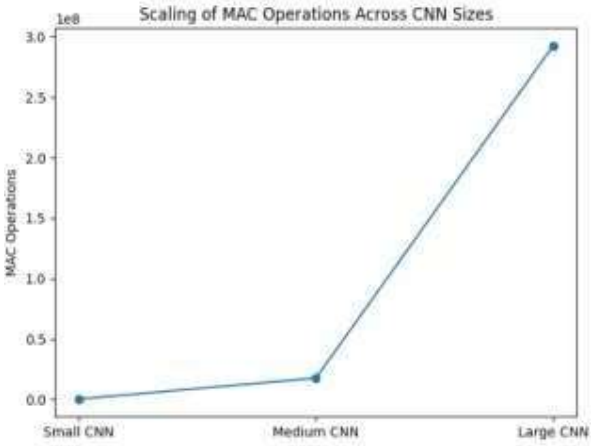

Fig. 3: MAC Scaling

Fig. 3 shows the scaling of MAC operations across CNN sizes. The exponential growth in computation highlights the increasing computational intensity in large networks

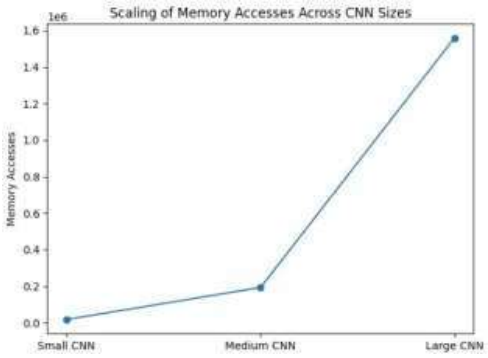

Fig. 4: Memory Access Scaling

Fig. 4 shows memory access scaling. Although memory accesses increase with model size, the growth rate is lower than for MAC operations, indicating an increase in arithmetic intensity.

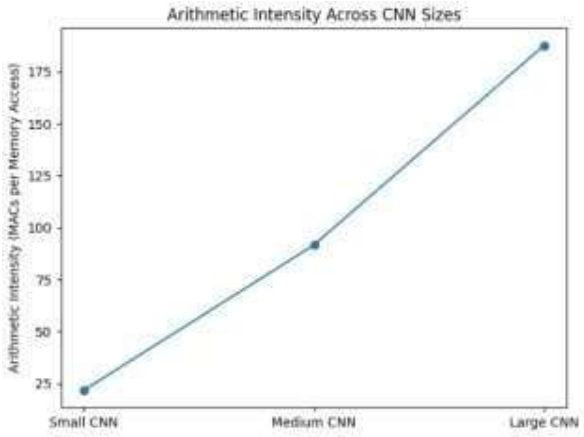

Fig. 5: Arithmetic Intensity

Fig. 5 shows the trend of arithmetic intensity. The increasing MAC-to-memory ratio explains the lesser relative advantage of PIM in large CNN configurations.

E. Discussion

The results collectively show that processing-in-memory architectures significantly reduce energy consumption by minimizing data movement between memory and computation units. The greatest benefit is seen in memory- bound workloads, where memory access energy dominates total consumption. However, as the computational intensity increases in large CNN models, the computational energy becomes dominant, reducing the relative advantage of PIM.

These findings are particularly relevant for edge AI systems, where lightweight CNN models are typically deployed under tight power constraints. The ability of PIM to provide about 28% energy savings for small networks suggests strong potential for battery-powered and resource-limited applications.

However, the current study is based on analytical modelling and software simulation. Future work may include hardware-level validation and exploration of hybrid PIM-computer architectures to further optimize energy efficiency.

VII. Conclusion

This paper presents an energy-efficient analytical framework for evaluating processing-in-memory (PIM) architectures applied to CNN inference in resource- constrained systems. A simulation-based method was developed to measure the impact of low memory movement on overall energy consumption.

Experimental results show that PIM achieves significant energy savings compared to conventional von Neumann architectures. In particular, an energy reduction of up to 27.89% was observed for small CNN workloads, while medium and large networks achieved 14.11% and 8.41% improvements, respectively. The results show that PIM is particularly effective for memory-bound workloads, where data movement dominates the total energy consumption.

The proposed analytical approach provides a scalable and hardware-independent evaluation framework, enabling architecture-level exploration without requiring physical hardware implementation.

Future work may include multi-layer CNN modelling with deep architectures, latency and memory bandwidth analysis, integration with cycle-accurate architectural simulators, and evaluation using real-world datasets to further validate the effectiveness of the proposed model.

VIII. References

[1] T. Spagnolo, C. Silvano, R. Massa, F. Grillotti, T. Boesch, and G. Desoli, "In-Pipeline Integration of Digital In-Memory Computing into RISC-V Vector Architecture to Accelerate Deep Learning," Feb. 2026.

[2] Z. Liu, "In-memory computing architectures for energy-efficient AI," in *Proc. Applied and Computational Engineering*, vol. 190, 2025, pp. 28–32.

[3] "PIM or CXL-PIM? Understanding architectural trade-offs through large-scale benchmarking," Nov. 2025.

[4] J. Šíma, P. Vidnerová, and V. Mrázek, "Energy complexity of convolutional neural networks," *Neural Computation*, vol. 36, no. 8, pp. 1601–1625, Jul. 2024.

[5] J. Lim, J. Son, and H. Yoo, "Efficient processing-in-memory system based on RISC-V instruction set architecture," *Electronics*, vol. 13, no. 15, Art. no. 2971, Jul. 2024.

[6] S. Jung, J. Lee, D. Park, Y. Lee, J.-H. Yoon, and J. Kung, "A dual- precision and low-power CNN inference engine using a heterogeneous processing-in-memory architecture," *IEEE Transactions on Circuits and Systems I*, vol. 71, no. 12, pp. 5546–5559, Dec. 2024.

[7] R. Kaur, A. Asad, and F. Mohammadi, "A comprehensive review of processing-in-memory architectures for deep neural networks," *Computers*, vol. 13, no. 7, Art. no. 174, Jul. 2024.

[8] F. Mohith, "A review on selective in-memory computing processors," *Embedded Systems Review*, 2025.

[9] "Energy efficiency impact of processing in memory: A comprehensive review of workloads on the UPMEM architecture," in *Proc. Parallel and Distributed Computing*, 2024.

[10] W. Li et al., "PIMSYN: Synthesizing processing-in-memory CNN accelerators," Feb. 2024.

[11] Y. Wan et al., "Pflow: An end-to-end heterogeneous acceleration framework for CNN inference on FPGAs," *Journal of Systems Architecture*, vol. 150, Art. no. 103113, May 2024.

[12] J. Ji-Hoon et al., "In-depth survey of processing-in-memory architectures for deep neural networks," *Journal of Semiconductor Technology and Science*, vol. 23, no. 5, pp. 322–339, 2023.

[13] C. Wang et al., "EPIM: Efficient processing-in-memory accelerators based on Epitome," Nov. 2023.

[14] J. Jung, H. Lee, H. Noh, J. Yoon, and J. Kung, "DualPIM: A dual- precision and low-power CNN inference engine using SRAM- and eDRAM-based PIM arrays," in *Proc. IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2022, pp. 70– 73.

[15] J. Agosta, "Deep learning on RISC-V platforms at the edge," *ACM Computing Surveys*, 2025.

[16] M. He, "Processing-in-memory design and optimizations

for machine learning inference," Ph.D. dissertation, Purdue University, West Lafayette, IN, USA, 2024.

[17] W. Li et al., "TIMELY: Pushing data movements and interfaces in PIM accelerators," May 2020.

[18] J. Kung et al., "Adaptive precision cellular nonlinear network," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Feb. 2018.

[19] V. Verma et al., "AI-PiM—Extending the RISC-V processor with processing-in-memory," *Frontiers in Electronics*, 2022.

[20] "Processing-in-memory techniques: Survey, advances, and challenges," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, May 2024.