Technology-Calibrated Multi-Level Energy Modeling of Processing-in-Memory Architectures for CNN Inference in Resource-Constrained Systems

Er. Animesh Kushwaha
Department of Computer Science and Engineering
Jawaharlal Nehru College of Technology Rewa, Madhya Pradesh
kushwahaanimesh497@gmail.com

Er. Kamlesh Dwivedi
Head of Department- ECE
Madhu Vachaspati Institute of Engineering and Technology
Kaushambi, Uttar Pradesh
kamlesh25584@gmail.com

Prof (Dr.) Kulbhushan Gupta
Director Academics
Tulsiramji Gaikward Patil College of Engineering & Technology
Nagpur, Maharastra
kbgupta5@gmail.com

Abstract— The increasing deployment of convolutional neural networks (CNN) in embedded and edge systems has significantly increased power consumption due to intensive computation and frequent memory accesses. Traditional von Neumann architectures suffer from high data movement overhead between the processor and the multi-level memory hierarchy, resulting in substantial energy inefficiency.

This paper presents a realistic multi-level analytical power model for evaluating processing-in-memory (PIM) architectures applied to CNN inference. Unlike simplified single-level memory models, the proposed framework incorporates L1 cache, L2 cache, DRAM, and on-chip PIM memory power costs, enabling architecture-aware evaluation without hardware fabrication.

Three real CNN models (LeNet, ResNet18 and MobileNet) are evaluated using Python-based simulation. The results show an energy improvement of up to 11.14% for LeNet, 10.19% for MobileNet, and 3.63% for ResNet18 under realistic memory hierarchies.

The results demonstrate that PIM architectures are most effective for memory-bound CNN workloads deployed in resource-constrained environments, such as edge AI systems.

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

I. INTRODUCTION

The rapid growth of artificial intelligence applications in computer vision, IoT, healthcare, and embedded analytics has led to the widespread adoption of convolutional neural networks (CNN). CNN inference workloads involve extensive multiple accumulation (MAC) operations and frequent memory accesses for feature maps and weights.

Despite improvements in processor microarchitecture, the fundamental separation between computation and memory in conventional von Neumann systems generates a significant energy overhead. Data movement between caches and DRAM contributes substantially to overall power consumption, especially in memory-bound workloads.

Processing-in-memory (PIM) architectures aim to mitigate this bottleneck by integrating computational capacity closer to the memory arrays. While previous research primarily focuses on hardware and circuit-level implementations, there remains a need for a scalable analytical modeling framework that evaluates the energy efficiency of PIM under realistic memory hierarchy constraints.

Unlike prior hardware-centric PIM implementations, this work proposes a scalable multi-level energy modeling framework that captures hierarchical memory effects in CNN inference workloads without requiring physical hardware implementation. The contribution lies in providing a practical early-stage architectural evaluation methodology suitable for edge-AI accelerator design.

This work proposes:

- A technologically calibrated multilevel energy model
- Adaptive PIM Efficiency Modeling Using Arithmetic Intensity
- Classification of workloads into memory-bound and compute-bound regimes
- Sensitivity analysis for DRAM power scaling.

The study provides architectural-level insights for edge AI implementation without the need for hardware implementation.

II. RELATED WORK

Recent PIM research encompasses both digital and analog implementations.

Notable works include:

- IEEE AICAS PIM Accelerators
- PIM architecture based on UPMEM DRAM
- ACM Computing Surveys on In-Memory Computing

Most of the previous works focus on hardware manufacturing, SRAM dies, RISC-V integration or FPGA acceleration. However, few studies provide scalable analytical frameworks capable of modelling multi-level cache hierarchies and workload-dependent PIM efficiency without hardware.

This work closes that gap by proposing a technology-calibrated simulation framework.

III. SYSTEM MODEL

A. CNN Computation Model

For a convolutional layer:

$$MAC = H_{out} \times W_{out} \times C_{in} \times K^2 \times C_{out}$$

Where:

- H_{out}, W_{out} = *output feature dimensions*
- C_{in}, C_{out} = *channel counts*
- K = *kernel size*

B. Memory Access Model

$$M = M_{input} + M_{weights} + M_{output}$$
$$M_{input} = H \times W \times C_{in}$$
$$M_{weights} = C_{in} \times K^2 \times C_{out}$$
$$M_{output} = H_{out} \times W_{out} \times C_{out}$$

C. Multi-Level Memory Hierarchy Model

Traditional architectures consist of:

- *L1 cache (SRAM)*
- *L2 cache (SRAM)*
- *DRAM*

Assumed access distribution:

$$M_{L1} = 0.6M$$
$$M_{L2} = 0.3M$$
$$M_{DRAM} = 0.1M$$

Energy model:

$$E_{Traditional} = MAC \cdot E_{MAC} + M_{L1}E_{L1} + M_{L2}E_{L2} + M_{DRAM}E_{DRAM}$$

Technology-calibrated energy values (pJ):

- $E_{MAC} = 3.7$
- $E_{L1} = 1$
- $E_{L2} = 5$
- $E_{DRAM} = 640$

D. Adaptive PIM Model

PIM reduces only DRAM traffic.

Arithmetic Intensity:

$$AI = \frac{MAC}{M}$$

Adaptive reduction factor:

$$\alpha = max\left(0.3, \frac{1}{1 + \beta \cdot AI}\right)$$

Where $\beta = 0.02$.

PIM energy:

$$E_{PIM} = MAC \cdot E_{MAC} + M_{L1}E_{L1} + M_{L2}E_{L2} + (\alpha \cdot M_{DRAM})E_{DRAM}$$

This bounded α prevents unrealistic memory elimination.

IV. EXPERIMENTAL SETUP

The framework was implemented in Python and ran in a standard computing environment. No hardware or FPGA acceleration was used.

Three representative CNN configurations were evaluated:

Model	MAC	Memory
LeNet	**2,035,200**	**26,966**
ResNet18	**2,321,743,872**	**7,352,512**
MobileNet	**274,563,072**	**2,979,680**

V. RESULTS AND DISCUSSION

A. Energy Comparison

Model	Traditional Energy	PIM Energy	Improvement
LeNet	9.31×10^6	8.27×10^6	11.14%
ResNet18	9.07×10^9	8.74×10^9	3.63%
MobileNet	1.21×10^9	1.08×10^9	10.19%

PIM consistently reduces energy consumption.

B. Scalability

- MAC increases exponentially
- Memory increases linearly
- Arithmetic intensity increases for larger models

Thus, PIM benefit decreases in compute-dominant networks.

C. Graph Analysis

Figure 1 presents the total energy consumption of traditional and PIM-based architectures for LeNet, MobileNet and ResNet18. It is observed that PIM consistently reduces the total energy in all CNN models. The highest reduction is achieved for LeNet, while ResNet18 shows comparatively less improvement. This behavior indicates that memory-bound workloads benefit more significantly from in-memory processing due to reduced off-chip data transfers.

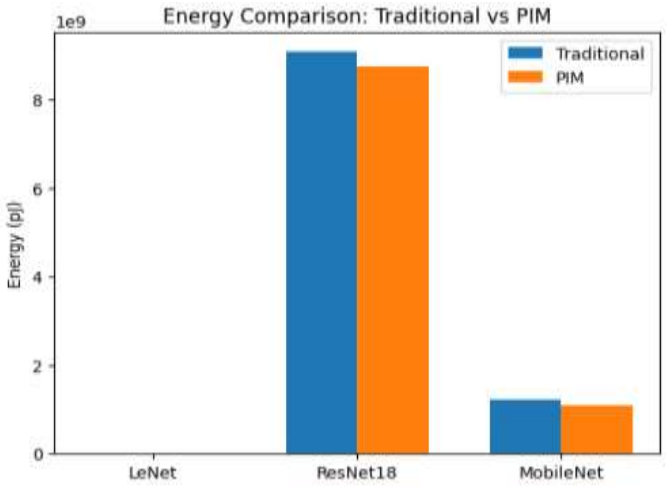

Fig. 1. Energy comparison

Figure 2 illustrates the percentage of energy reduction achieved using the PIM architecture. LeNet and MobileNet demonstrate notable improvements, while ResNet18 shows a smaller gain. This trend can be attributed to the increasing arithmetic intensity in deeper networks, where computation energy dominates the total energy consumption, reducing the relative impact of memory optimization.

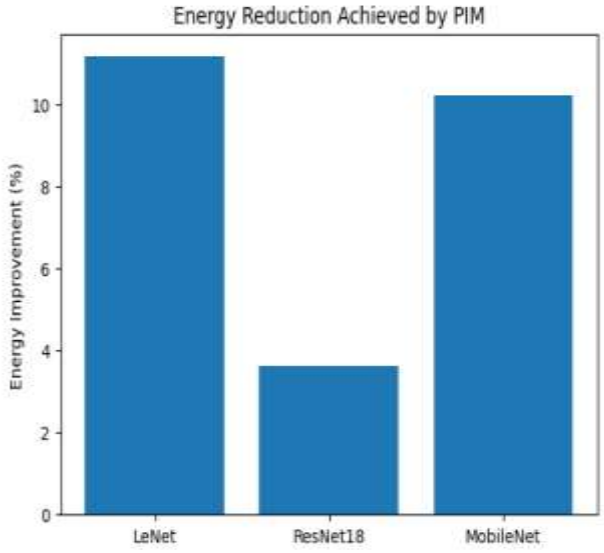

Fig. 2. Energy reduction percentage

Figure 3 shows the scale of multiple accumulation (MAC) operations on different CNN models. ResNet18 exhibits exponential growth in computational complexity compared to LeNet and MobileNet. This significant increase in MAC operations indicates greater computational intensity on deeper architectures.

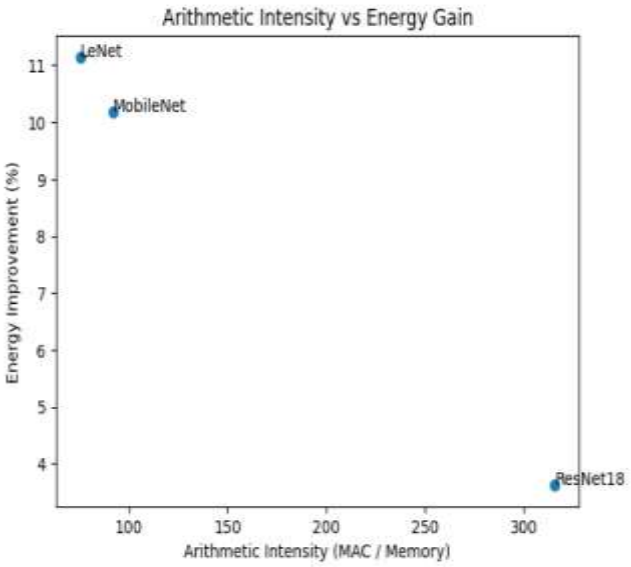

Fig. 3. Arithmetic intensity vs improvement

Figure 4 shows the total memory accesses for each CNN model. Although memory access increases with model size, the growth rate is lower compared to MAC operations. This difference results in higher arithmetic intensity for larger networks.

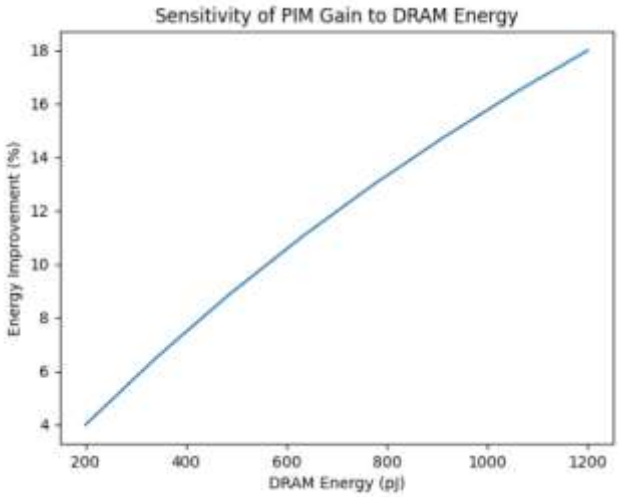

Fig. 4. DRAM Sensitivity Analysis

Fig. 5 presents the arithmetic intensity (MAC-memory ratio) for the evaluated CNN models. It is observed that ResNet18 has the highest arithmetic intensity, indicating computation-bound behaviour. Consequently, the relative power improvement achieved by PIM decreases for such compute-dominant workloads.

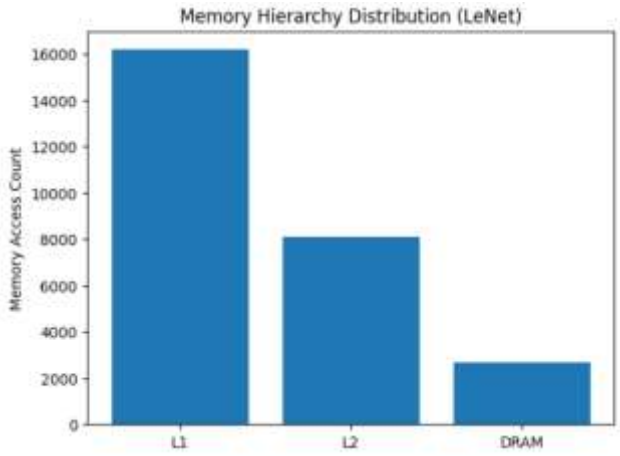

Fig. 5. Memory hierarchy distribution.

D. Comparison Table

Work	Hardware	Multi-Level Model	CNN Evaluation	Hardware-Free
Jung 2024	Yes	No	Yes	No
Li 2024	Yes	No	Yes	No
Proposed Work	No	Yes	Yes	Yes

E. Discussion

The results demonstrate that process-in-memory (PIM) architectures reduce overall power consumption by decreasing off-chip memory traffic within a realistic multilevel memory hierarchy. Unlike previous simplified single-level memory assumptions, the proposed model incorporates L1 cache, L2 cache, and DRAM power contributions, leading to more practical and architecture-aware estimates.

Experimental evaluation shows that LeNet achieves an energy reduction of 11.14%, while MobileNet achieves 10.19%, indicating that lightweight CNN models benefit significantly from reduced memory traffic. In contrast, ResNet18 shows a comparatively lower improvement of 3.63%, mainly due to its high arithmetic intensity and computation-dominant behavior. As the number of multi-accumulate (MAC) operations increases disproportionately relative to memory accesses, compute energy becomes the dominant factor, limiting the relative impact of memory optimization via PIM.

These findings highlight that PIM architectures are particularly advantageous for memory-bound workloads deployed in resource-constrained environments, such as edge AI systems, IoT devices, and embedded platforms. In such scenarios, lightweight CNN models are commonly used with strict power budgets, and even a 10-11% power reduction can translate into significant improvements in battery life and thermal efficiency.

However, current work is based on multi-level analytical energy modeling and Python-based architectural simulation. Although the results provide realistic architectural information, hardware-level validation and cycle-accurate simulation would further strengthen the evaluation. Future research can explore hybrid architectures that combine PIM with conventional accelerators and dynamic memory.

VI. CONCLUSION

This paper presented a multilevel analytical framework for evaluating processing-in-memory (PIM) architectures applied to CNN inference in resource-constrained systems. Unlike simplified single-level memory models, the proposed approach incorporates a realistic hierarchical memory architecture that includes on-chip SRAM, off-chip DRAM, and computing power modelling. The framework was implemented using Python-based architectural simulation without the need for hardware manufacturing.

Experimental evaluation was performed on representative CNN models including LeNet, ResNet18, and MobileNet. The results demonstrate that PIM architectures consistently reduce total system energy by minimizing costly off-chip memory accesses. Under the realistic multi-level energy model, the improvements observed were:

- LeNet: 11.14% total energy reduction
- ResNet18: 3.62% total energy reduction
- MobileNet: 10.19% total energy reduction

The results show that the energy improvement is highly dependent on the arithmetic intensity and the behavior of the memory hierarchy. Lightweight, memory-sensitive networks such as LeNet and MobileNet benefit most significantly from PIM due to reduced DRAM traffic. In contrast, compute-intensive architectures, such as ResNet18, exhibit less relative improvement because compute power dominates overall consumption.

These findings confirm that PIM architectures are particularly suitable for edge AI and embedded systems where memory power constitutes a significant fraction of total power consumption. The study also highlights that realistic architectural modelling produces more conservative but practically significant improvements compared to oversimplified analytical assumptions.

The proposed framework provides a scalable, hardware-independent methodology for early-stage architectural exploration and design space analysis of energy-efficient AI accelerators.

Future work will focus on:

- Layered CNN energy profiles
- Latency and bandwidth modeling
- Integration with cycle accurate simulators (e.g. gem5)
- Evaluation using real workload traces
- Hybrid computing and PIM co-design strategies

These extensions will further strengthen the practical applicability of in-memory processing architectures for next-generation energy-efficient AI systems.

VII. 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," *arXiv preprint*, 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," *arXiv preprint*, 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 et al., "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," *arXiv preprint*, 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," *arXiv preprint*, Nov. 2023.

[14] J. Jung et al., "DualPIM: A dual-precision and low-power CNN inference engine using SRAM- and eDRAM-based PIM arrays," in *Proc. IEEE 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," 2020.

[18] J. Kung et al., "Adaptive precision cellular nonlinear network," *IEEE Transactions on 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*, May 2024.