

Re-RIS: A Reconfigurable 3D RRAM In-Sensor Architecture for Low-Latency Machine Vision

Shiyang Li^{1,2}, Lixia Han³, Siyuan Chen^{1,2}, Lifeng Liu^{1,2}, Peng Huang^{1,2,*}

¹School of Integrated Circuits, Peking University, Beijing 100871, China

²Beijing Advanced Innovation Center for Integrated Circuits, Beijing 100871, China

³College of Integrated Circuits, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Email: *phwang@pku.edu.cn

Abstract—Cutting-edge machine vision applications impose stringent latency and energy efficiency demands on edge devices. To address these demands, In-Sensor Computing (ISC) architectures aim to eliminate data movement overhead, while 3D RRAM technology provides the hardware foundation of high memory density and massive computing parallelism. However, existing ISC architectures rely on static resource allocation, failing to address the dynamic “shifting bottleneck” in CNNs—where early layers are compute-bound and later layers are readout-bound. To address this, we propose Re-RIS, a Reconfigurable 3D RRAM In-Sensor architecture. By dynamically switching hardware granularity between high-parallelism and high-throughput modes, Re-RIS optimizes resource utilization for varying layer characteristics. Experimental results on VGG-16 demonstrate an end-to-end latency of 0.93 ms, achieving a 75% reduction compared to static baselines, with an energy efficiency of 244.6 TOPS/W and an area efficiency of 1.85 TOPS/mm².

Keywords—3D-stacked integration, Computing-in-memory, 3D RRAM, Computational image sensor

I. INTRODUCTION

Real-time perception systems on edge devices impose stringent latency constraints (<10 ms). Traditional cloud and edge computing paradigms struggle to meet these demands due to significant data movement overheads. While In-Sensor Computing (ISC) mitigates sensor-to-processor transmission [1–4], current static implementations still fall short of optimal real-time performance, as illustrated in the latency comparison in Figure 1.

This limitation in traditional ISC stems from a fundamental efficiency challenge: the shifting bottleneck. Analysis of CNN layers reveals that the performance-limiting factor changes dynamically. Early layers are compute-bound, characterized by massive input feature maps but small weights, requiring high computational parallelism. Conversely, later layers are readout-bound, involving large weight matrices that create a data interface bottleneck. Static ISC architectures force a compromise: optimizing for parallelism fragments the large weights of later layers, causing excessive readout latency, while optimizing for readout utilization limits the parallelism of early layers.

To address this, we propose Re-RIS, a Reconfigurable 3D RRAM In-Sensor architecture. By dynamically adapting the hardware granularity, Re-RIS eliminates the trade-off between computation and readout efficiency, enabling real-time inference.

II. PROPOSED ARCHITECTURE

A. 3D Heterogeneous Integration

Re-RIS leverages a three-wafer vertical stack to maximize bandwidth and density, as illustrated in Figure 2. The Top

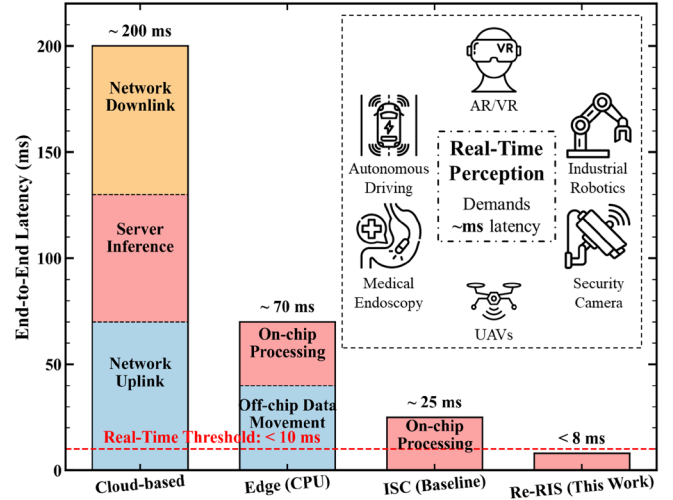


Figure 1. End-to-end latency comparison across computing paradigms.

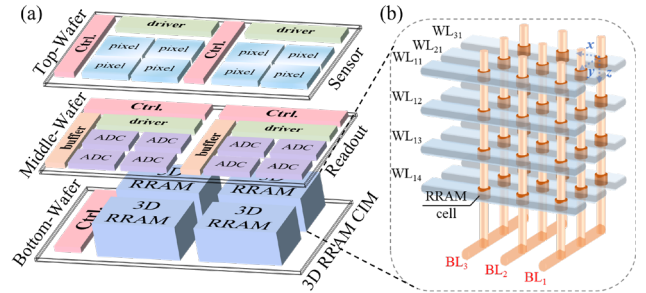


Figure 2. Re-RIS system architecture. (a) Heterogeneous 3-wafer stack. (b) 3D RRAM cross-point array.

Wafer contains the pixel array for image acquisition. The Middle Wafer serves as the control hub, housing ADCs, buffers, and the reconfiguration logic. The Bottom Wafer consists of high-density 3D Vertical RRAM (VRRAM) macros, which store the entire CNN model on-chip to avoid off-chip memory access.

B. Reconfigurable Granularity

The core innovation is the dynamic reconfiguration mechanism shown in Figure 3. The array is composed of fundamental units connected via programmable switches, with drivers replicated for each unit to enable independent operation. Although adding these switches and drivers incurs a minor area overhead, they provide essential architectural flexibility. By toggling the connection states of these switches, the architecture can alternate between two operating modes tailored to the specific workload of current layer:

- **High-Parallelism Mode:** For compute-bound layers, switches isolate units into independent Small Macros. Weights are duplicated, and inputs are broadcasted to maximize parallel MAC operations.

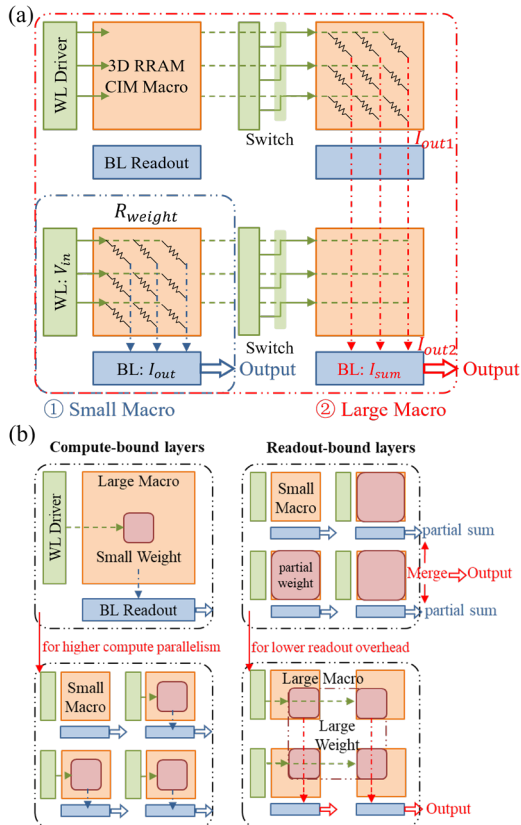


Figure 3: (a) Reconfiguration mechanism. (b) Dynamic dataflows.

- **High-Throughput Mode:** For readout-bound layers, switches merge adjacent bit-lines (BLs) to form Large Macros. This allows large weight matrices to be mapped without fragmentation, enabling a unified, high-bandwidth readout path.

Guided by a mapping algorithm that determines the specific connection topology and granularity, Re-RIS dynamically adapts hardware to match the shifting bottleneck. This ensures optimal resource utilization, achieving a configuration that satisfies our area and latency requirements.

III. EXPERIMENTAL RESULTS

A. Experimental Setup

We evaluated Re-RIS using a cross-layer simulation framework. Circuit-level parameters were extracted from rigorous HSPICE simulations, calibrated against reported 3D RRAM [5,6] and industrial PDKs. Crucially, the simulation explicitly models circuit non-idealities, specifically focusing on IR drop across the vertical arrays and crosstalk effects induced by leakage currents, ensuring realistic physical constraints. We also conduct architectural-level exploration to determine optimal parameter configurations and evaluate their impact on inference accuracy.

B. Design Space Exploration

The "macro width" determines the granularity of reconfiguration, presenting a fundamental trade-off between performance and area. As shown in Figure 4(a), our exploration on VGG-16 reveals that reducing the macro width from 512 to 32 decreases latency from 27.63 ms to 0.93 ms, albeit at the cost of increased area (from 2.6 mm² to 7.2 mm²) due to control overhead.

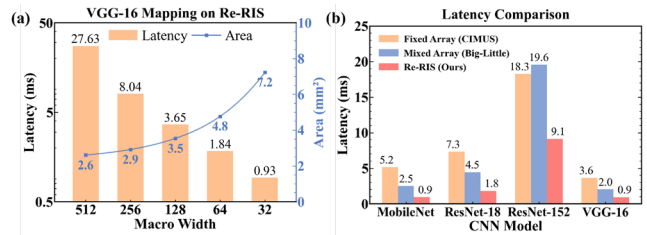


Figure 4: (a) Design space exploration on VGG-16. (b) Latency comparison across diverse CNN models.

TABLE I. BENCHMARK RESULTS OF RE-RIS AND OTHER DESIGNS

| VGG-16 for ImageNet task | Energy efficiency | Frames rates |
|--------------------------|--------------------|----------------|
| Sony with CNN DSP [7] | 2.05Tops/W | 120fps |
| CIMUS [2] | 57.7Tops/W | 800fps |
| Big-Little Chiplets [8] | 48.5Tops/W | 833fps |
| Ours: Re-RIS | 244.6Tops/W | 1075fps |

This Pareto frontier empowers users to identify and implement the optimal configuration that minimizes latency according to their specific acceptable area constraints.

C. Performance Comparison

Table I compares Re-RIS with state-of-the-art static (CIMUS [2]) and heterogeneous (Big-Little [8]) architectures. Key observations from the evaluation are detailed as follows:

- **Latency & Efficiency:** On VGG-16, Re-RIS achieves 0.93 ms latency and 244.6 TOPS/W energy efficiency. This represents a 75% reduction in latency compared to the static baseline under an identical hardware resource budget, primarily due to the elimination of readout stalls in later layers.
- **Versatility:** As illustrated in Figure 4(b), these gains are consistent across diverse workloads. For lightweight models like MobileNet and deep networks like ResNet-152, Re-RIS consistently outperforms static counterparts.
- **Thermal:** The distributed processing nature of Re-RIS inherently mitigates localized hotspots compared to centralized accelerators by spreading heat generation across the active array. Comprehensive thermal modeling and detailed heat dissipation simulations will be discussed in our future work.

IV. CONCLUSION

We presented Re-RIS, a Reconfigurable 3D RRAM In-Sensor architecture. By dynamically switching between parallelism-centric and readout-centric modes, Re-RIS effectively resolves the shifting bottleneck dilemma in CNNs. Comprehensive evaluations demonstrate that Re-RIS achieves superior latency and energy efficiency compared to static counterparts, paving the way for next-generation real-time edge intelligence.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research Plan of China under Grant 2023YFB4402400; in part by the National Natural Science Foundation of China Program under Grant 62034006 and Grant 62474005; in part by the STIC under Grant QYJS-2022-2200-B; in part by the 111 Project Program under Grant B18001.

REFERENCES

- [1] F. Zhou, and Y. Chai. "Near-sensor and in-sensor computing," *Nature Electronics*, vol. 3, no. 11, pp. 664-671, 2020.
- [2] L. Han *et al.*, "CIMUS: 3D-stacked computing-in-memory under image sensor architecture for efficient machine vision," *IEEE Transactions on Computers*, 2025.
- [3] Z. Liu *et al.*, "NS-CIM: A current-mode computation-in-memory architecture enabling near-sensor processing for intelligent IoT vision nodes," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 9, pp. 2909-2922, 2020.
- [4] N. Tang *et al.*, "First Demonstration of 1T FDSOI-Based > 1000fps Image Sensor with In-pixel Computing," in *Symposium on VLSI Technology and Circuits*, 2025.
- [5] Q. Huo *et al.*, "A computing-in-memory macro based on three-dimensional resistive random-access memory," *Nature Electronics*, vol. 5, no. 7, pp. 469-477, 2022.
- [6] C. Ma *et al.*, "First implementation of monolithic integrated CIM with 1Mb ultra-high-density 8-layer 3D VRRAM, achieving high computing density (204.8 GOPs/mm²) and FoM (2.13×10⁶ GOPS²/W/mm²) for efficient scientific computing," in *Symposium on VLSI Technology and Circuits*, 2025.
- [7] R. Eki, *et al.*, "9.6 A 1/2.3 inch 12.3 Mpixel with on-chip 4.97 TOPS/W CNN processor back-illuminated stacked CMOS image sensor," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, 2021.
- [8] G. Krishnan *et al.*, "Big-little chiplets for in-memory acceleration of dnns: A scalable heterogeneous architecture," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022.