A 4541 TOPS/W Saliency-Aware Analog Computing In-Memory Macro with Charge-Domain Saliency Detector

Shimpei Ando, Satomi Miyagi, Wenlun Zhang, Yung-Chin Chen, and Kentaro Yoshioka
Keio University, Yokohama, Japan
Email : shimpeiando@keio.jp

***Abstract*—This paper presents the first silicon implementation of a Saliency-Aware Analog Computing In-Memory (SACIM) macro that dynamically optimizes computational precision based on data importance. While conventional approaches to reduce ADC power consumption in ACIMs rely on sparsity-based methods that dynamically adjust ADC resolution, their zero-counting technique remains incompatible with bit-parallel computing—a key strength of ACIMs. Our SACIM introduces an online charge-domain saliency detector that efficiently identifies and prioritizes critical computations while maintaining bit-parallel compatibility. The design features a novel Super-Skip mode that reduces ADC power consumption by over 70%, and we propose a current-integration based multi-bit input driver that offers accurate signal generation with low implementation cost. Together, these deliver an overall 5.8× boost in energy efficiency—a higher improvement rate than has been demonstrated by previous sparsity-based approaches.**

Fabricated in 65nm CMOS, the SACIM prototype achieves peak 4541 TOPS/W energy efficiency while maintaining high inference accuracies of 92.0%/69.0% on CIFAR-10/100 datasets, with minimal accuracy degradation (0.7% and 0.3% respectively) compared to full-precision computation. This work demonstrates a promising direction using saliency for energy-efficient, high-performance ACIM-based deep learning accelerators.

I. Introduction

Analog Computing In-Memory (ACIM) technology has emerged as a promising architecture for minimizing both data movement and computational power in DNN computing [1-9]. However, a significant challenge in ACIM is the high power consumption of ADCs required for reading computation results. To address this issue, previous works have proposed sparsity-based techniques that dynamically adjust ADC resolution according to data characteristics [1-3]. These approaches estimate the required ADC output range by counting zeros in the input/weight vectors, allowing them to reduce unnecessary ADC cycles and conserve power by adapting the precision to the actual data content. However, this zero counting sparsity method faces two significant limitations:

1) **Architectural Limitation:** Incompatibility with bit-parallel CIMs [4], which are crucial for ACIM power efficiency through multi-valued input computation;
2) **Sparsity Limitation:** Only sparsity in either input or weight can be utilized, significantly limiting energy efficiency gains.

To overcome power efficiency limitations of sparsity-based ACIMs, we propose a Saliency-Aware ACIM (SACIM) which operates based on "Saliency", a measure of computation result importance. SACIM supports bit-parallel inputs and introduces several key innovations:

1) **Super-Skip mode** aggressively exploits saliency to reduce ADC power consumption by up to 70%
2) **Charge-domain saliency detector** achieves efficient analog-domain saliency detection
3) **Current integration-based multi-bit input driver** reduces input signal generation overheads

Our SACIM prototype in 65nm CMOS achieves a peak energy efficiency of 4541 TOPS/W (1b/1b) and inference accuracies of 92% and 69% on CIFAR-10 and CIFAR-100, respectively. Compared to full precision ADC results, saliency-based computation introduces minimal accuracy loss (0.7% CIFAR-10, 0.3% CIFAR-100) while reducing ADC power by over 70%.

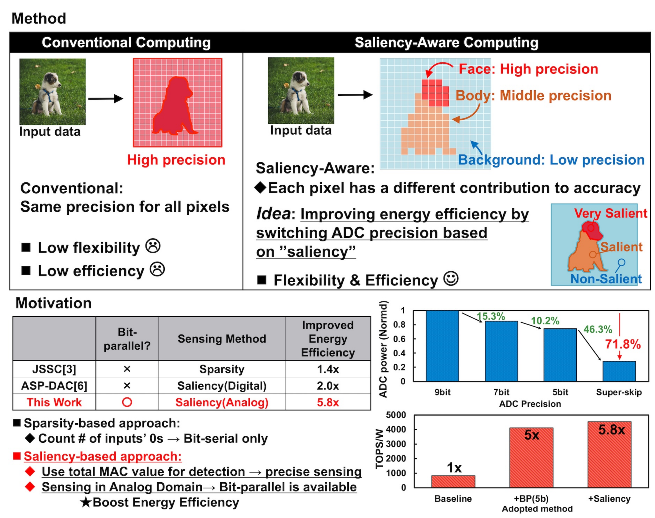


Fig. 1. Top: Concept of saliency-aware computing. Bottom: Block diagram of the proposed SACIM implementation with dynamic precision control based on data importance.

II. Saliency-Aware Analog CIM (SACIM)

A. Key SACIM Concepts

Fig. 1 shows the concept of the saliency-aware computing based on [6] and the proposed SACIM, which dynamically controls computation precision based on computing data importance. As illustrated in Fig.1, in image recognition, pixels constituting the "dog" significantly influence classification results (Very-Salient), whereas background pixels have minimal impact (Non-Salient). The fundamental principle of saliency-aware computation is to perform high-precision operations

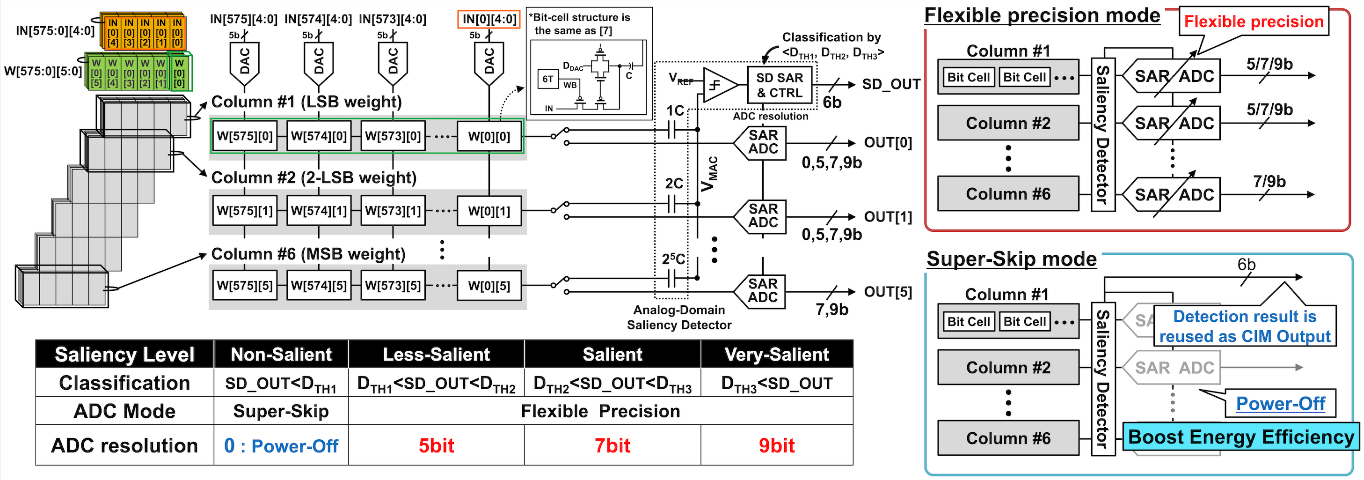


Fig. 2. Detailed operational modes of SACIM architecture: Flexible precision mode sets independent ADC resolution per column, and Super-Skip mode utilizes saliency detection results as CIM output with most ADCs powered off. Bottom shows saliency level classification.

only on important data while executing lower-precision computations on non-salient data, thereby achieving enhanced power efficiency without sacrificing accuracy. Our analysis shows that in CNNs, this importance (saliency) can be derived simply from convolution operation output magnitudes. Building on this finding, SACIM determines saliency by generating an analog voltage corresponding to the entire MAC output through charge sharing and assessing its amplitude.

Our SACIM offers two key innovations: **1) Bit-parallel Compatibility:** enables handling of sparsity in both inputs and weights for bit-parallel CIMs; **2) Super-Skip Mode:** for non-salient data, the saliency detection result can directly serve as the final MAC output without accuracy loss, significantly improving energy efficiency. While previous sparsity-based research [1-3] achieved efficiency improvements of up to 1.4×, SACIM achieves up to 5.8× improvement through aggressive ADC precision optimization and bit-parallel computation.

B. SACIM Operation

Fig. 2 shows the operational concept of the SACIM. Building on the bit-cell design from [7], our macro comprises six columns (#1-#6) and processes 5-bit parallel inputs with 6-bit serial weights, achieving a comprehensive 576 × 5b × 6b MAC operation. The SACIM operates through a three-phase process: 1) The 576-row ACIM computations in the column CIM generate analog MAC results corresponding to each weight bit; 2) These individual results are binary-weighted within the Analog-Domain Saliency Detector (ADSD) to synthesize VMAC, a single analog approximation of the entire MAC output; 3) the dedicated SAR ADC within the ADSD (SD-SAR) converts VMAC to determine computation saliency by comparing against predefined thresholds (DTH1, DTH2, DTH3).

Based on this comparison, we categorize computations into four distinct saliency levels, enabling us to independently and dynamically assign different precision levels to each of the column ADCs (#1 - #6) online. This column-specific precision control is a key feature of our design, allowing fine-grained

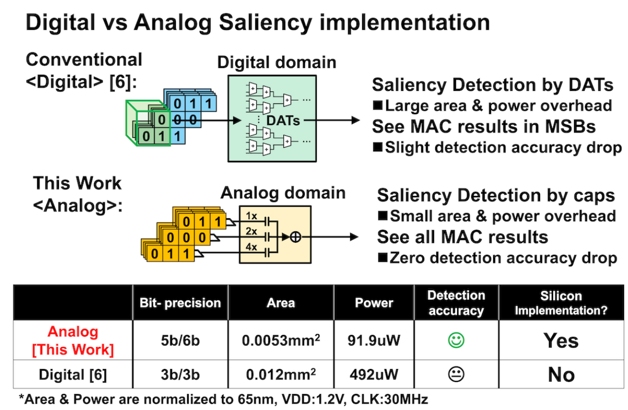


Fig. 3. Comparison between conventional digital saliency detection [6] and our proposed analog approach. Our analog implementation achieves more accurate saliency detection with 2× lower area and 5× lower power consumption through direct charge-domain sensing of the full MAC output.

optimization based on each column's contribution to the final result. The threshold values are decided through software simulations, offering flexibility to optimize for energy efficiency, accuracy, or balanced performance based on application requirements. For Very-Salient data, full 9-bit ADC precision ensures maximum accuracy, while Less-Salient and Salient data utilize reduced 5-bit and 7-bit precision respectively to conserve power. Most significantly, Non-Salient data triggers our proposed Super-Skip Mode, which bypasses ADCs #1–#4 entirely while utilizing only ADCs #5–#6 and combining their results with the SD-SAR output to form the complete MAC result. This aggressive precision optimization strategy achieves up to 70% reduction in ADC power consumption without compromising computational accuracy.

C. Digital vs. Analog Saliency Detection Approaches

Fig. 3 shows a detailed comparison with the digital-based saliency computing approach [6]. Their method only uses the top 3 MSBs of digital CIM MAC results for saliency detection, requiring a power and area hungry Digital Adder Trees for accumulation, and yet captures just a fraction of the computation result.

In contrast, our proposed *fully analog approach* determines saliency from the *entire* 576×5b×6b of analog MAC output, providing three fundamental advantages: (1) complete visibility of the full computation result for more accurate saliency estimation, (2) elimination of digital conversion overhead before saliency detection, and (3) dramatically lower implementation cost. Based on layout-level comparison, our fully-analog saliency detection method reduces area and power consumption by over 2× and 5× respectively compared to [6]. Furthermore, while [6] remains at the simulation level, we demonstrate the world's first silicon prototype that successfully integrates and validates saliency-aware computing in real hardware.

III. Analog Domain Saliency Detector(ADSD)

Fig. 4 shows the ADSD circuit implementation. For saliency detection, we capture the six column analog MAC outputs (#1-#6) using binary weighted capacitors to generate VMAC, an analog voltage that represents the entire MAC operation result. While prior work [9] demonstrated similar analog-domain accumulation, it relied on power-intensive buffers to achieve the high-precision analog computation necessary for ACIM applications.

Our approach offers a significant advantage: since VMAC serves only for saliency detection and processing Non-Salient MAC results, the required computation precision is substantially lower. This enables us to implement the entire process through passive charge-sharing, eliminating power-hungry amplifiers and achieving remarkable power and area efficiency. The dedicated SD-SAR converts VMAC to determine saliency level, which then dynamically configures the precision of column ADCs by switching CDAC and SAR Logic configuration, enabling variable resolution based on the detected importance. The balancing capacitor ensures uniform charge-sharing ratios across all column capacitors, preventing inconsistencies that would arise from variations in weighted capacitor values. The unit capacitance used in the ADSD is set to 2.1 fF, which is sufficiently large to ensure robustness against mismatches and parasitic capacitance. The ADSD capacitor array occupies only 0.002 mm², which is approximately 1.9% of the total CIM macro area, making it extremely compact.

IV. Current-Integration multi-bit row driver.

Fig. 5 illustrates the proposed current-integration multi-bit row driver. While bit-parallel processing significantly improves the efficiency of ACIM, using multiple power supplies has resulted in large system overhead [8], necessitating an area-efficient multi-bit row driver. In ACIMs, a critical challenge is that the driver load varies according to the total weight of the row (ΣW). Rather than compensating for this issue using a C-DAC which requires a large area [7], we propose compensating for ΣW variation by controlling the integration time (t_{int}). The t_{int} generation circuit is implemented by a compact, fully-digital variable delay circuit, making it suitable for scaling.

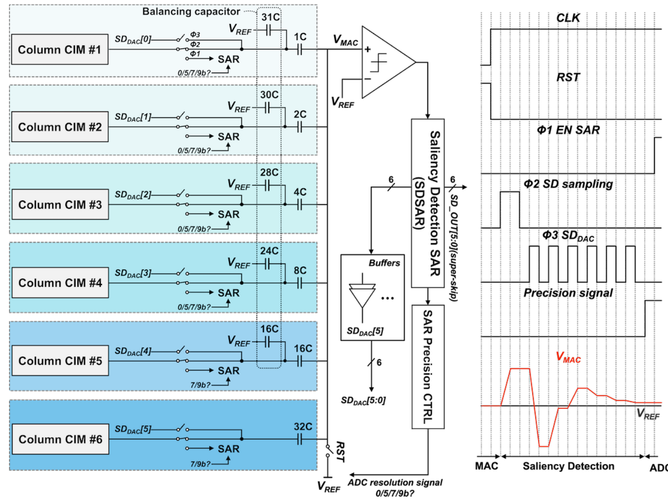


Fig. 4. Schematic of Analog Domain Saliency Detector(ADSD) and operation waveforms.

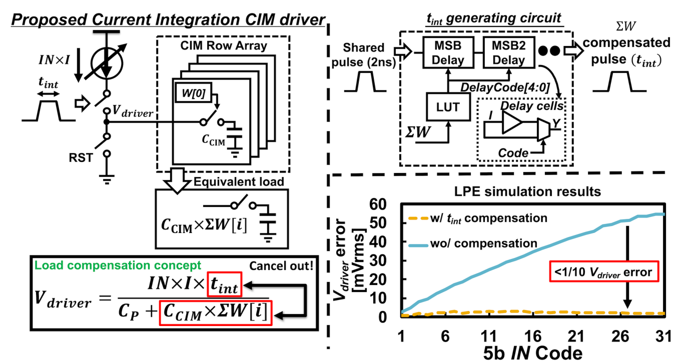
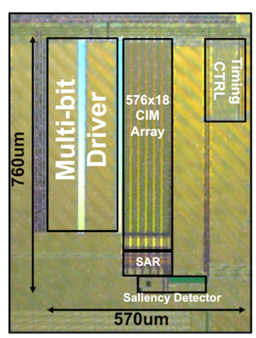


Fig. 5. Schematic and concept of proposed Current-Integration multi-bit row CIM driver.

LPE simulations validate our approach, showing that our t_{int} based load variation compensation reduces the driver voltage generation error to 1/10 compared to the uncompensated case.

V. Experimental results and Comparison

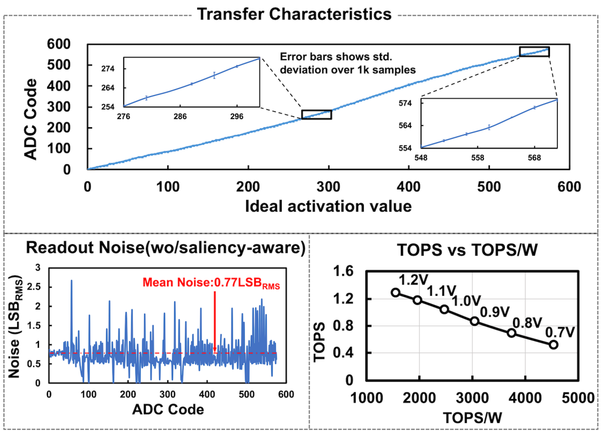
The prototype SACIM was fabricated in 65nm CMOS as shown in Fig. 6.



Chip Summary		
Technology	TSMC 65nm	
Supply Voltage	0.7-1.2	
Input/weight precision	1-6b	
Core Area	0.433mm^2	
Array size	10Kb	
Peak Energy Efficiency (TOPS/W)	4541	
CIFAR-10 / CIFAR-100 accuracy(%)	Fixed precision	92.7 / 69.3
	Saliency (Dynamic precision)	92.0 / 69.0

Fig. 6. Die photo and chip summary.

Fig. 7 presents the measured CIM column characteristics, showing an average noise of 0.77 LSBRMS sampled across 1k data. Proposed SACIM achieves a peak energy efficiency of 4541 TOPS/W at 0.7V. To optimize the fundamental accuracy-power tradeoff, ADC precisions for saliency levels #1-#6 were



Saliency Configuration

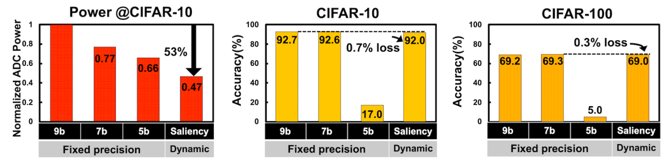
	Non-Salient	Less-Salient	Salient	Very-Salient
ADC #1-#4	Skip	5b	7b	9b
ADC #5-#6	7b	7b	7b	9b
ADC Power Ratio	0.28	0.78	0.85	1
CSNR[dB] [7]	13	16	20	26
Usage Ratio @CIFAR-10	50%	25%	15%	10%

Fig. 7. Saliency-Aware CIM macro measured performance and saliency configuration table.

configured based on the Computing SNR (CSNR) metric [7], ranging from 13 dB to 26 dB. Our implementation follows a hierarchical precision strategy, where Non-Salient level operates with ADCs #1-#4 powered off and #5-#6 at 7-bit precision, while Very-Salient level utilizes all ADCs at full 9-bit precision to maximize accuracy. This allows us to flexibly control power consumption across different importance levels of the data without incurring large accuracy penalties.

To systematically validate our saliency-aware methodology, we performed threshold tuning using CIFAR-10 training data and analyzed MAC output distributions to identify optimal operating points that balance inference accuracy and ADC power consumption. This comprehensive tuning process builds upon our CSNR-based framework by determining precise allocation boundaries for Non-Salient, Less-Salient, Salient, and Very-Salient categories. Our simulation results indicate that around half of the inputs can be safely handled in Non-Salient mode (7-bit or Super-Skip) without significantly harming classification accuracy, leading to less than 1% degradation relative to a fixed 9-bit baseline. By selectively applying higher precision only to the most critical (Very-Salient) computations, the overall accuracy loss remains within 0.7% on CIFAR-10, matching our threshold-tuning estimates. These results confirm that the saliency-based mechanism effectively targets power reduction at less important data while preserving key computations at full precision.

Fig. 8 shows the measurement results and comparison table between the proposed SACIM and the state-of-the-art ACIMs. The SACIM macro achieved inference accuracies of 92.0% and 69.0% on CIFAR-10 and CIFAR-100, respectively, with accuracy losses of only 0.7% and 0.3%. When compared to conventional fixed-precision approaches using 9-bit, 7-bit, and 5-bit ADCs, our saliency-based design with Super-Skip mode reduces power consumption by up to 53% while maintaining high inference accuracy on CIFAR-10. This is particularly significant as traditional fixed 5-bit precision schemes suffer from substantial accuracy degradation. By leveraging saliency for bit-parallel processing and extensively optimizing ADC precision and power consumption, the proposed SACIM achieves a peak energy efficiency of 4541 TOPS/W, outperforming previous works that utilize sparsity.



Comparison Table

<table>
<tr><th></th><th colspan="4">This Work</th><th>JSSC'23[3]</th><th>JSSC'20 [4]</th><th>JSSC'24[7]</th><th>ASP-DAC'24[6]</th></tr>
<tr><td>Technology</td><td colspan="4">65nm</td><td>28nm</td><td>65nm</td><td>65nm</td><td>65nm</td></tr>
<tr><td>CIM Type</td><td colspan="4">Charge</td><td>Charge</td><td>Charge</td><td>Charge</td><td>Hybrid</td></tr>
<tr><td>Array Size</td><td colspan="4">10Kb</td><td>16Kb</td><td>572Kb</td><td>84Kb</td><td>9Kb</td></tr>
<tr><td>Bit Precision</td><td colspan="4">1-6b</td><td>1-16b</td><td>1-8b</td><td>4-8b</td><td>4-8b</td></tr>
<tr><td>ADC Precision</td><td colspan="4">0/5/7/9b</td><td>8b</td><td>8b</td><td>8/10b</td><td>3b</td></tr>
<tr><td>Eff. enhancement method</td><td colspan="4">Saliency</td><td>Sparsity</td><td>-</td><td>-</td><td>Saliency</td></tr>
<tr><td>Bit-parallel</td><td>Baseline</td><td>+BP(5b)</td><td colspan="2">+Saliency</td><td>×</td><td>×</td><td>5b</td><td>4b</td></tr>
<tr><td>Peak Energy Eff. (TOPS/W)*2</td><td>824</td><td>4119</td><td colspan="2">4541</td><td>165</td><td>400</td><td>4094</td><td>370</td></tr>
<tr><td rowspan="2">Inference accuracy CIFAR-10/CIFAR-100(%)</td><td>9b ADC</td><td>7b ADC</td><td>5b ADC</td><td>Saliency</td><td rowspan="2">91.0/-</td><td rowspan="2">92.4/-</td><td rowspan="2">91.7/-</td><td rowspan="2">- /72.1</td></tr>
<tr><td>92.7/69.2</td><td>92.6/69.3</td><td>17.0/5.0</td><td>92.0/69.0</td></tr>
</table>

[2]Normalized to 1b/1b & 65nm process

Fig. 8. Measurement results and Comparison Table.

VI. Conclusion

This paper presents the world's first silicon implementation of a Saliency-Aware Analog Computing In-Memory (SACIM) macro that dynamically adjusts computation precision based on data importance. By leveraging Super-Skip mode and a charge-domain saliency detector, SACIM achieves 4541 TOPS/W, significantly improving energy efficiency up to 5.8× while maintaining high accuracy (0.7% loss on CIFAR-10, 0.3% on CIFAR-100). The proposed approach reduces ADC power consumption by over 70%, demonstrating the effectiveness of saliency-driven precision control.

References

[1] B. Zhang et al., "A 177 TOPS/W, Capacitor-based In-Memory Computing SRAM Macro with Stepwise-Charging/Discharging DACs and Sparsity-Optimized Bitcells for 4-Bit Deep Convolutional Neural Networks," *IEEE CICC*, 1-2, 2022.

[2] P. Chen et al., "7.8 A 22nm Delta-Sigma Computing-In-Memory (ΔΣCIM) SRAM Macro with Near-Zero-Mean Outputs and LSB-First ADCs Achieving 21.38TOPS/W for 8b-MAC Edge AI Processing," *IEEE ISSCC*, 140-142, 2023.

[3] C. -Y. Yao et al., "A Fully Bit-Flexible Computation in Memory Macro Using Multi-Functional Computing Bit Cell and Embedded Input Sparsity Sensing," *IEEE JSSC*, vol. 58, no. 5, 1487-1495, 2023.

[4] H. Jia et al., "A Programmable Heterogeneous Microprocessor Based on Bit-Scalable In-Memory Computing," *IEEE JSSC*, vol. 55, no. 9, 2609-2621, 2020.

[5] S. -E. Hsieh et al., "7.6 A 70.85-86.27TOPS/W PVT-Insensitive 8b Word-Wise ACIM with Post-Processing Relaxation," *IEEE ISSCC*, 136-138, 2023.

[6] Y. -C. Chen et al., "OSA-HCIM: On-The-Fly Saliency-Aware Hybrid SRAM CIM with Dynamic Precision Configuration," *IEEE ASP-DAC*, 539-544, 2024.

[7] K. Yoshioka, "A 818–4094 TOPS/W Capacitor-Reconfigured Analog CIM for Unified Acceleration of CNNs and Transformers," *IEEE JSSC*, Early Access, 2024.

[8] J. Lee et al., "Fully Row/Column-Parallel In-memory Computing SRAM Macro employing Capacitor-based Mixed-signal Computation with 5-b Inputs," *IEEE Symp. VLSI Circuits*, 1-2, 2021.

[9] X. Yang et al., "A 4-Bit Mixed-Signal MAC Macro With One-Shot ADC Conversion," *IEEE JSSC*, vol. 58, no. 9, 2648-2658, 2023.