A 28nm RRAM-Based 81.1 TOPS/mm^2/bit Compute-In-Memory Macro with Uniform and Linear 64 Read Channels under 512 4-bit Inputs

Peng Yao[1*], Qiumeng Wei[1], Dong Wu[1], Bin Gao[1], Siyao Yang[1], Ting-Ying Shen[2], Qingtian Zhang[1], Sining Pan[1], Jianshi Tang[1], He Qian[1], Lu Jie[1*], and Huaqiang Wu[1]
[1]School of Integrated Circuits, Tsinghua University, Beijing, China.
[2]Xiamen Industrial Technology Research Institute, Fujian, China.
[*]E-mail: yaopeng14@mail.tsinghua.edu.cn, jielu@mail.tsinghua.edu.cn

2024 IEEE European Solid-State Electronics Research Conference (ESSERC) | 979-8-3503-8813-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ESSERC62670.2024.10719511

***Abstract*—The rapid development of AI models imposes stringent demands on computing density at edge side. High parallelism computing-in-memory (CIM) macro with multi-level weights and analog inputs can meet the requirements. This work proposes a dual-loop clamping ADC to attain linear and uniform readouts, along with an incremental integration scheme to suppress thermal noise. A novel segmented array structure is employed to enhance computing flexibility with configurable parallelism. Based on these innovations, a 28nm high-parallelism resistive random-access memory (RRAM) CIM macro with 4-bit weight and 4-bit inputs is reported to attain a state-of-the-art normalized computing density of 81.1 TOPS/mm^2/bit with the largest 512 parallelism. In addition, this work reports high linearity (0.9985 of $\mu(R^2)$ and uniformity (0.0003 of $\sigma(R^2)$ across 64 parallel ADCs for the first time.**

***Index Terms*—compute-in-memory, edge computing, computing density, RRAM, ADC**

I. Introduction

As deep neural networks and large language models become increasingly prevalent on edge devices, the demand for high computing power and large storage for weight parameters has become essential. RRAM-based Computing-in-Memory (CIM) hardware possesses high storage density and non-volatility. Based on the crossbar array structure, it can realize fast in-situ matrix-vector multiplication (MVM) calculation with low power consumption. Consequently, edge devices equipped with CIM technology offer a promising solution for deploying complex models and networks within the constraints of power consumption and areas, providing the necessary computing power and memory storage capacity. However, RRAM-based CIM macros and processors reported in contemporary researches generally prioritize high energy efficiency. These works have limited computing density due to the single-bit input scheme and low input-parallelism [1], [2].

High-parallelism RRAM CIM with multi-level weights and inputs is a potential way to achieve high computing density and weight density. To support larger on-chip neural networks, each analog RRAM device can represent multi-bit weight to enhance the bit density. By applying analog voltages to the array simultaneously, multi-bit MVM operations are executed in a single calculation cycle. Nonetheless, this high-parallelism scheme poses severe challenges for CIM circuits: 1) The conductance load exhibits greater variation in high parallelism mode, leading to severe readout non-linearity and variance under non-ideal output impedance. 2) Thermal noise would be exacerbated by high-conductance load to deteriorate system accuracy significantly. 3) The full parallelism configuration lacks flexibility and fails to meet the algorithmic requirements of varying throughputs and accuracies.

In this work, we implement a RRAM-based analog CIM macro with high computing density and linearity using multi-bit encoded design. The work includes three major innovations. Firstly, we propose a novel ADC design featuring a dual-loop clamping structure. This design decouples the operating point of the clamping circuit from the array load, mitigating non-linearity in readouts and non-uniformity between channels. Secondly, an integrator with incremental integration-time configuration is proposed in the dual-loop clamping ADC (DLC-ADC) to execute I-V conversion of load current. This scheme with dynamic conversion speed suppress the thermal noise with negligible overhead of extra latency. Thirdly, a segmented array structure based on a 2-level WL driver (including global WLs, or WLGs, and local WLs, or WLLs) is adopted, which enables the adjustment of computing parallelism to align with the required computation scale, power and precision.

This paper is organized as follows: Section II describes the fabricated macro design and circuit structures; Section III illustrates measurement results; Section IV concludes the paper.

II. Macro Structure and Circuit Design

A. Macro Overview

The proposed macro includes a 512×512 2T2R array. It could perform MVM operations with 4-bit input, 4-bit weights, and 8-bit output precision within a single calculation cycle. A differential 4-bit weight (W_P–W_N) is encoded in a single 2T2R weight cell [3] (Fig. 2(a)), with each device programmed to

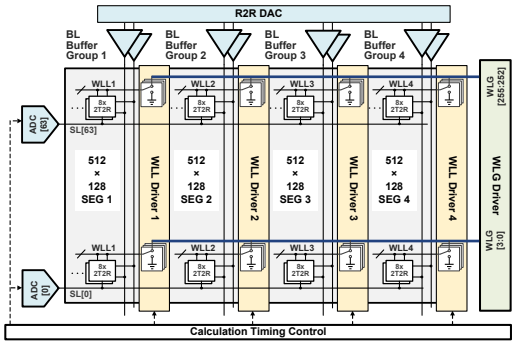


Fig. 1. Diagram of the proposed macro.

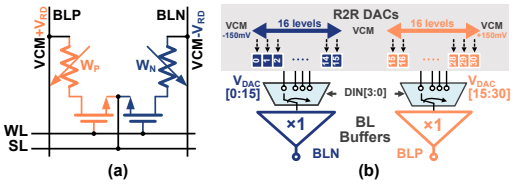


Fig. 2. (a) Illustration of 2T2R weight cell. (b) Structure of 4-bit bipolar input circuits.

one of eight discrete conductance-levels. A pair of differential voltages are applied to BLP and BLN representing the 4-bit inputs. The common-mode input voltage is VCM, and the differential range is from 0 to 150mV and 0 to -150mV, respectively. These voltages are generated by 31 groups of resistive ladder DACs (R2R DACs), which are routed to the BLP/BLN through switches and driving buffers (depicted in Fig. 2(b)). To reconcile the disparity in pitch between the peripheral circuits and the high-density array, each BLN/BLP is shared by two rows of weight cells, and each SL is shared by four columns of weight cells. The target cell is selected by activating the corresponding WL

To enhance the configurability of the system, the array and input buffers are partitioned into four segments, each with 128-input parallelism and capable of operating independently. The 64 channels of ADC could precisely clamp SL voltages and simultaneously readout calculation results. This macro can perform INT4 or INT8 computations to meet the algorithmic precision demands for running various models on edge devices. Shift and addition operations required by INT8 calculation are executed outside the macro in this design.

B. Dual-loop Clamping ADC with Incremental Integration

To realize the high-linearity calculation under high parallelisms, we proposes an innovative ADC design as depicted in Fig. 3. A dual-loop clamping structure is integrated in ADC to eliminate readout non-linearity and non-uniformity between different channels. An incremental integration-time method is adopted to suppress the noise under high input-parallelism.

Traditional readout circuits include a separate part of clamping circuit. After signal is established, it is first sampled and then quantized by the ADC. To ensure the linearity under high input-parallelism, clamping circuits require a sufficient loop gain to reduce the output impedance. In addition, it is necessary to expedite the settling of the SL clamping voltage to reduce the read latency. These generally lead to large power consumption and decreases the energy efficiency. Moreover, it could also cause large energy expense to lower clamping noise that would be amplified by the large total SL conductance.

The proposed DLC-ADC could mitigate these non-ideal effects in traditional readout circuits. The first feedback loop is employed to provide a high bandwidth. This loop is based on a gain boosting structure, in which a common source amplifier accelerates the voltage settling on the SL. Switching capacitors in the feedback of Loop1 is to provide a stable DC operating point to the amplifier, and they are refreshed during the RST_EN phase in each calculation cycle. The corresponding timing diagram is presented in Fig. 4. To improve its tolerance to mismatches and power supply fluctuations, offset cancellation is applied as follows. The amplifier output AOUT is connected to the gate of clamping transistor GC through a capacitor C_{OS}. Before the calculation, $SW_{S1\text{-}S3}$, and SW_{S5} are connected, while SW_{S4} is disconnected to sample the offset on C_{OS}. In the calculation phase, switching capacitors is refreshed periodically to sustain the gain boosting loop and the offset between AOUT and GC is cancelled.

The second loop provides a large effective DC gain to reduce clamping error. This loop involves an IDAC feedback loop embedded in the SAR ADC that feeds current back to the clamping node to compensate the load current. A G_m-C integrator is embedded in Loop2 for I-V conversion and filtering of high-frequency noise. The residue current is integrated on C_1 and C_2 during the calculation cycle, as depicted in Fig. 4. After that, the comparator compares the voltage output of the G_m-C integrator, i.e., the voltage on C_1 and C_2. The SAR logic then controls the IDAC to converge the residual current towards zero according to the comparison results. After the SAR conversion, the bias current on the NM_1 is maintained to be approximately constant. Consequently the operating point of the clamping circuit is decoupled form

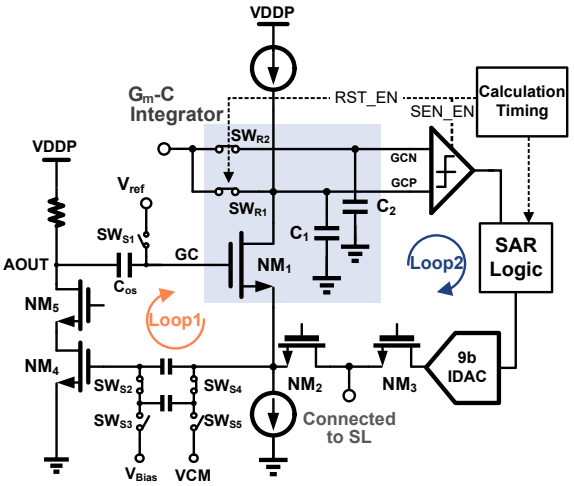


Fig. 3. Proposed ADC with a dual-loop clamping structure.

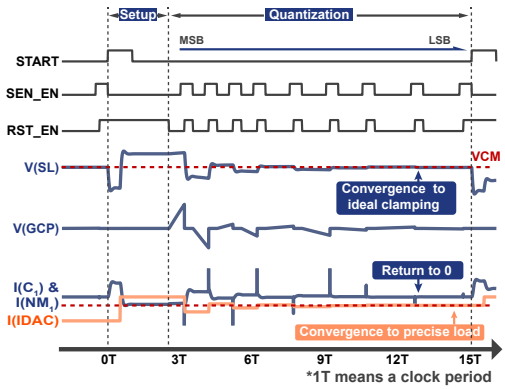


Fig. 4. ADC calculation timing.

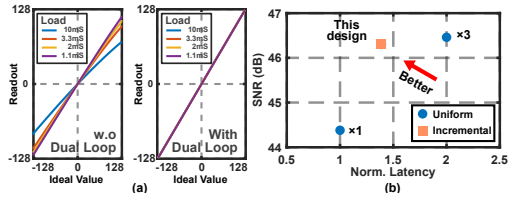


Fig. 5. Simulation results of dual-loop clamping and dynamic conversion speed.

the varying load. Furthermore, by directly comparing the residue load current rather than pre-sampled signals, this DLC-ADC can effectively reduce dynamic errors caused by incomplete signal settling. This approach consequently lowers the bandwidth demands for both the clamp circuits and input buffers.

Additionally, we adopt a incremental integration scheme with dynamic conversion speed to suppress the thermal noise and save the quantization time. Corresponding controlling signals are shown in Fig. 4. As the SAR conversion proceeds, higher resolution is required to correctly quantize LSB signals. To improve the SNR, the integration duration is also progressively increased from comparisons of MSB to LSB in this method.

C. Segmented Array with Configuration Parallelism

To increase the flexibility of high-parallelism analog CIM, the input and output parallelism, as well as the associated peripheral circuits, should be dynamically activated in line with the network scale and adapted to the varying computation precision requirements. This work introduces a 2-level WL driver to subdivide the array into four independent segments, where the input-parallelism of each segmentation is 128. Firstly, it reduces the capacitive load of the global WL, which accelerates the WL switching for shorter readout latency. Secondly, it facilitates the adjustment of input parallelism, thereby providing flexible adaptation to the throughput and accuracy needs of various networks. Due to the smaller load conductance under low input-parallelism, the calculation is more accurate. For applications with high accuracy requirements, the input can be split into several segments, calculated sequentially at low parallelism. The multiple partial sums are accumulated subsequently. Furthermore, by powering down unnecessary input buffers, a computing mode with fewer activated segments can achieve lower power consumption.

III. Characterization Results

A. Evaluation of Dual-Loop Clamping ADC

Fig. 5(a) shows the simulated linearity and slopes in readouts across various weight loads. A gain boost-based clamping circuit is compared to the proposed DLC-ADC. Load conductance in the simulation ranges from 10mS to 1.1mS, representing the typical load range under high input-parallelism. Without dual-loop clamping, the quantization slope varies with the changing load conductance. Moreover, under 10mS load, the readout shows prominent non-linearity. The simulation verifies that the dual-loop clamping structure can realize uniform and linear readout across varying loads.

Fig. 5(b) presents the speed and precision advantages regarding the proposed incremental integration scheme. We evaluate the precision loss and readout latency under ×1, ×3, and ×1.8 (this design) integration-time, respectively. It verifies that the proposed time configuration approaches the highest precision with only 69% latency.

B. Measured MVM Accuracy

To qualify the precision of MVM calculation of the proposed ADC under high parallelism, 10,000 groups of input vectors are used. Each vector contains 512 sets of 4-bit input data that is randomly distributed from 0x0 to 0xf. The weight array is programmed using a fixed pattern with a sparsity of 50%. Fig. 6(a) exhibits the measured readouts under 512 input-parallelism within a single channel. The ratio of the Root Mean Square Error (RMSE) to the Full-Scale Range (FSR) is 2.68%. Fig. 6(b) plots all the readout results of 64 output channels. Each data point represents an average derived from 8 read iterations to illustrate the denoised trend. Fig. 6(c) shows the correlation coefficients R^2 between the readout results and the ideal values of all channels, all of which are distributed above 0.997. RMSE ratios are also sprad within

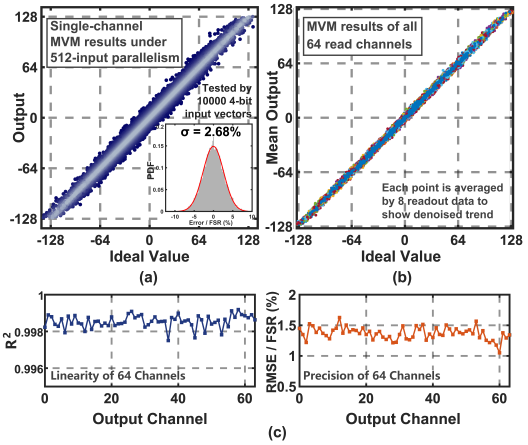


Fig. 6. a) Measured MVM calculation results under input parallelism of 512. b) Averaged MVM calculation results of all 64 output channels. c) Linearity (R^2) and precision (RMSE/FSR) of all output channels.

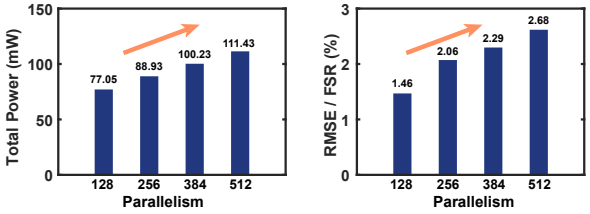


Fig. 7. Power and precision under different input-parallelisms.

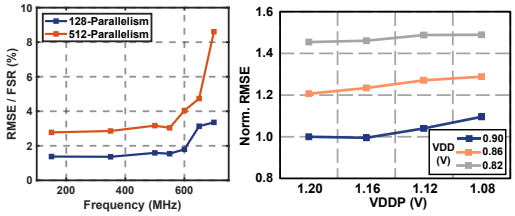


Fig. 8. Measured calculation precision under different clock frequency and power supply voltages.

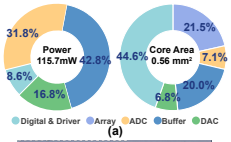
0.5%. These results verify that the proposed DLC-ADC has enhanced linearity and uniformity.

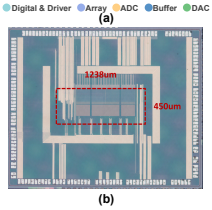
C. Measured System Performance

To verify the flexibility, we measured the performance under different segmentation configurations, as shown in Fig. 7. By decreasing the input-parallelism from 512 to 128, the precision improves by 1.85 times and and power decreases by 34.38mW. Fig. 8 shows the measured macro performance across clock frequency and power supply variations. Here, VDDP refers to the power supply voltage for the ADC clamping circuits, while VDD denotes the power supply voltage for the remaining analog circuits. The macro can operate at a maximum frequency of 650MHz accompanied by a quantization latency of 23.08ns when dealing with the precision of 4-bit input, 4-bit weight, and 8-bit output. Fig. 9 displays a breakdown of area and power consumption, as well as other fundamental details of the macro. Power is measured under 512-input, 64-output parallelism, with all calculation-related peripheral circuits activated. Algorithmic models (e.g. MLP and CNN) are demonstrated with software comparable accuracies based on the macro.

IV. Conclusion

In this work, we report a RRAM-based CIM macro integrating a novel ADC structure with dual-loop clamping and configurable segmented array. The proposed RRAM-based CIM macro maintains high MVM precision under parallelism of 512, clock frequency of 650M Hz. This macro realizes a high 1-bit normalized energy efficiency of 628.8 TOPS/W/bit and storage density of 2.68Mb/mm^2. It presents the MVM linearity of $\mu(R^2)$ = 0.9985 and uniformity of $\sigma(R^2)$ = 0.0003 for 64 parallel ADCs, and shows the state-of-the-art normalized computing density of 81.1 TOPS/mm^2/bit with the largest parallelism. It is verified that the fabricated macro can meet the demands for various algorithmic models.





Technology (nm)		28nm HPC+
Weight Cell		2T2R (RRAM)
Memory Capacity		512Kb
Core Area (mm^2) *1		0.56
Memory Array Area (mm^2)		0.12
Clock Frequency (MHz)		150-650
Core Supply Voltage (V)		0.78-1.2
RRAM Operating Voltage (V)		2.2
RRAM Program Pulse Width (ns)		100-150
Single Calculation Latency (ns)		100-23.08
Throughput (TOPS) @ 4bIN/4bW/8bO		2.84
RMSE / FSR (%)	4bIN/4bW/8bO	1.46-2.68 *3
	8bIN/4bW/8bO	1.41-2.56 *3
Inference Accuracy	MNIST @ MLP *2	98.55% ± 0.04
	CIFAR-10 @ VGG11 *2	90.71% ± 0.14
	ImageNet @ ResNet50 *2	75.68% ± 0.09

*1 The core area includes analog and digital circuits related to computation and programming.
*2 Inference accuracy is obtained by simulation based on test results under 512 4-bit parallelism.
*3 Measured with parallelism ranging from 128 to 512 for 4-bit inputs.

(c)

Fig. 9. (a) Power and area breakdown. (b) Prototype die photo. (c) Summary information

	This work	VLSI 2023 [4]	ISSCC 2023 [5]	ISSCC 2022 [6]	VLSI 2021 [7]
Technology (nm)	28	40	22	40	14
NVM Device	RRAM	RRAM	RRAM	PCM	PCM
Calculation Precision	4bIN/4bW/8bO	1bIN/1bW/6bO	INT8	INT8	8bIN/ana W /8bO
Macro-level Storage Density (Mb/mm^2)	2.68	2.38	3.16 *4	N/A	0.1 *4
Accumulation Length	**512** *1	64	128	8	256
EE (TOPS/W)	19.3-39.3 *2	75.2	67.2-76.5	20.5-65	10.5
1bit Normalized EE (TOPS/W/bit)	308.8-628.8	75.2	4300.8-4896.0	1312.0-4160.0	336.0
Computing Density (TOPS/mm^2)	5.07 *3	7.01	0.81	N/A	1.59
1bit Normalized Computing Density (TOPS/mm^2/bit)	**81.14**	7.01	51.84	N/A	50.88
Linearity & Uniformity $\sigma(R^2)/\mu(R^2)$	**0.0003/ 0.9985**	N/A	N/A	N/A	N/A

*1 A pair of differential inputs associated with a pair of 2T2R weight unit counts as one accumulation.
*2 Measured from VDDC=0.78~0.9V, VDDP=1.08~1.2V, 650MHz, with 512 4bIN and 64 8bO, weight sparsity of 50%.
*3 Measured at VDDC=0.9V, VDDP=1.2V, 650MHz, with 512 4bIN and 64 8bO.
*4 Estimated, not directly reported.

Fig. 10. Comparison with state-of-the-art non-volatile memory-based CIM

Acknowledgment

This work is supported by NSFC (92064001, 62025111, 92064004 and 92164302), and Beijing Advanced Innovation Center for Integrated Circuits.

References

[1] W. Ye et al., "A 28-nm RRAM Computing-in-Memory Macro Using Weighted Hybrid 2T1R Cell Array and Reference Subtracting Sense Amplifier for AI Edge Inference," in IEEE Journal of Solid-State Circuits, vol. 58, no. 10, pp. 2839-2850.

[2] S. D. Spetalnick et al., "A 40nm 64kb 26.56TOPS/W 2.37Mb/mm^2 RRAM Binary/Compute-in-Memory Macro with 4.23x Improvement in Density and >75% Use of Sensing Dynamic Range," 2022 IEEE International Solid-State Circuits Conference (ISSCC), 2022,

[3] Q. Liu et al., "33.2 A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," 2020 IEEE International Solid-State Circuits Conference - (ISSCC), 2020, pp. 500-502.

[4] S. D. Spetalnick et al., "A 2.38 MCells/mm^2 9.81 -350 TOPS/W RRAM Compute-in-Memory Macro in 40nm CMOS with Hybrid Offset/IOFF Cancellation and ICELL RBLSL Drop Mitigation," 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2023, pp. 1-2.

[5] W. -H. Huang et al., "A Nonvolatile Al-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W," 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, pp. 15-17.

[6] W. -S. Khwa et al., "A 40-nm, 2M-Cell, 8b-Precision, Hybrid SLC-MLC PCM Computing-in-Memory Macro with 20.5 - 65.0TOPS/W for Tiny-Al Edge Devices," 2022 IEEE International Solid-State Circuits Conference (ISSCC), 2022, pp. 1-3.

[7] R. Khaddam-Aljameh et al., "HERMES Core – A 14nm CMOS and PCM-based In-Memory Compute Core using an array of 300ps/LSB Linearized CCO-based ADCs and local digital processing," 2021 Symposium on VLSI Circuits, 2021, pp. 1-2. pp. 1-3.