A 28-nm Static-Power-Free Fully Parallel RRAM-Based TD CIM Macro With 1982 TOPS/W/Bit for Edge Applications

Songtao Wei, *Graduate Student Member, IEEE*, Peng Yao, *Member, IEEE*, Xinying Guo, Dong Wu, Lu Jie, *Member, IEEE*, Qi Qin, *Student Member, IEEE*, Bin Gao, *Senior Member, IEEE*, Jianshi Tang, *Senior Member, IEEE*, He Qian, Sining Pan, *Member, IEEE*, and Huaqiang Wu, *Senior Member, IEEE*

***Abstract*—Analog computing in memory (CIM) based on resistive nonvolatile memory (NVM) has encountered several issues, such as low parallelism, low computing accuracy, and considerable power consumption. In this letter, a temporal unit based on design technology co-optimization (DTCO) for resistive random access memory is proposed for the first time, with the advantage of eliminating dc current and reducing the deviation of mapped weight. A time-domain (TD) array based on the proposed temporal unit features performing fully parallel matrix–vector multiplication (MVM) in a static-power-free manner, without the consideration of IR drop and limited sensing margin (SM). Besides, a low-power time–digital converter (TDC) with local offset elimination further boosts energy efficiency (EF) and computing accuracy. The fabricated 28-nm TD CIM macro achieves a state-of-the-art normalized EF of 1982 and 1387 TOPS/W/bit under 1b-input, ternary-weight and 4b-input, signed 4b-weight, respectively.**

***Index Terms*—Computing in memory (CIM), design technology co-optimization (DTCO), matrix–vector-multiplication (MVM) acceleration, resistive random access memory (RRAM).**

I. Introduction

Tiny machine learning (ML) [1] has become a popular frontier in recent years, in which DNN is deployed on Internet of Things (IoT) devices with constrained hardware resources and stringent energy efficiency (EF) requirements. Computing in memory (CIM) has been validated as a promising architecture for this application. It performs calculation at the data storage location, eliminating high-cost data transfer between computation and storage units [2]. Analog CIM based on resistive random access memory (RRAM) shows great potential in EF due to lower wake-up power and promising bit density compared to its digital and SRAM-based counterparts [3], [4].

Prior resistive RRAM-based CIM macros either perform multiplication-and-accumulation (MAC) in the current domain or the voltage domain. As shown in Fig. 1(a), prior works in the current domain [5], [6], [7] suffer from considerable IR drop-induced accuracy degradation and large static power consumption. Voltage-domain implementation [8], [9], on the other hand, is constrained by low parallelism due to limited sensing margin (SM). Additionally, CIM works in both domains suffer from large conductance variation of RRAM devices, which degrades the computation accuracy.

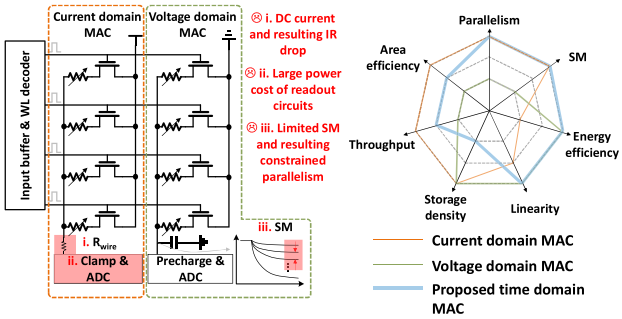

Fig. 1. (a) Drawbacks of previous current- and voltage-domain RRAM-based MAC. (b) Comparison between the proposed TD RRAM MAC and previous RRAM MAC.

In this letter, we propose an RRAM-based time-domain (TD) CIM macro for the first time to address the issues in previous works. The advantages of this letter are presented in Fig. 1(b), and the contributions are summarized as follows.

1) A novel RRAM-based TD cell based on a tunable-delay buffer (DB) is first presented to reduce weight mapping deviation by 26% from the perspective of device characteristics of RRAM, thus improving the reliability of the matrix–vector-multiplication (MVM) computation.
2) A TD-CIM macro based on the proposed TD cell is built to perform MVM in a fully dynamic manner, which eliminates dc currents, relieves IR drop, and maintains sufficient signal margin. Finally, the low RMS error (0.94%) in MVM computing is measured under full input parallelism (320 input data in parallel).
3) A low-power time-to-digital converter (TDC) with a novel local offset canceling (LPOSC-TDC) scheme to enhance the system's EF and eliminate offset at the cost of only an extra 3% area and $<1\%$ power consumption.

With the above contributions, the fabricated 28-nm TD-CIM macro achieves a state-of-the-art (SOTA) normalized EF of 1982 and 1387 TOPS/W under 1b input, ternary weight and 4b input, signed 4b weight configuration, respectively. Besides, only 0.94% RMS is exhibited under 4b-input configuration.

II. RRAM-Based TD CIM Implementation

A. TD CIM Macro Overview

The overall architecture of this CIM macro is presented in Fig. 2(a). This macro mainly consists of a TD array with a capacity of 320 $\times$ 128 TD cells, which is divided into four subarrays

Received 2 October 2024; revised 30 November 2024; accepted 13 December 2024. Date of publication 19 December 2024; date of current version 6 January 2025. This work was supported in part by NSFC under Grant 62422405, Grant 62025111, Grant 62495100, and Grant 92464302; in part by the Beijing Advanced Innovation Center for Integrated Circuits; and in part by the Shanghai Municipal Science and Technology Major Project. This article was approved by Associate Editor Bongjin Kim. *(Corresponding authors: Peng Yao; Sining Pan.)*

Songtao Wei, Peng Yao, Xinying Guo, Dong Wu, Lu Jie, Qi Qin, Jianshi Tang, He Qian, and Sining Pan are with the School of Integrated Circuits, Tsinghua University, Beijing 100084, China (e-mail: pyao@mail.tsinghua.edu.cn; psn@tsinghua.edu.cn).

Bin Gao and Huaqiang Wu are with the School of Integrated Circuits, Tsinghua University, Beijing 100084, China, and also with the International Innovation Center of Tsinghua University, Shanghai 200062, China.

Digital Object Identifier 10.1109/LSSC.2024.3520593

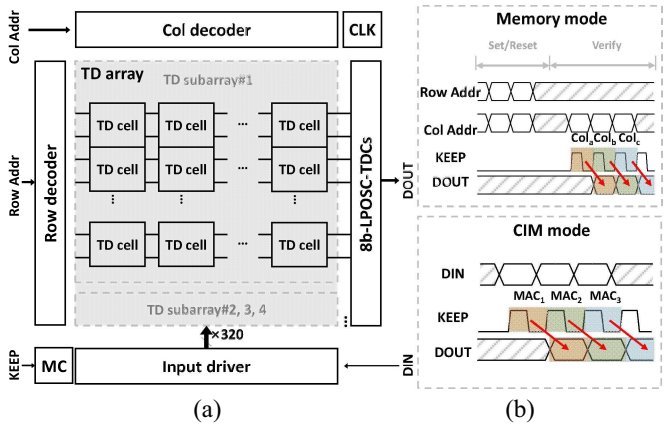

Fig. 2. TD CIM macro overview: (a) macro architecture and (b) timing diagram under different working modes.

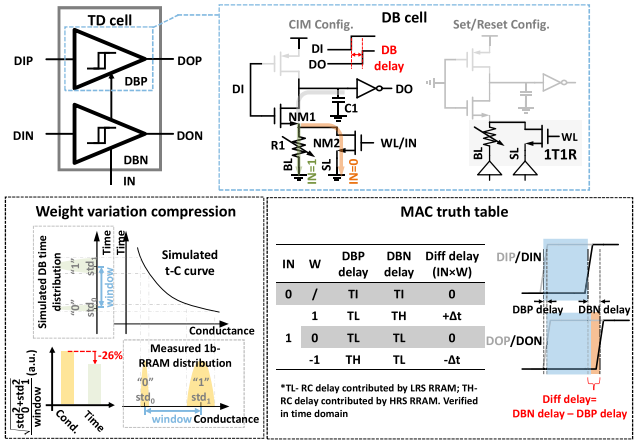

Fig. 3. TD cell working principle: basic component (DB cell) structure (top), simulated weight deviation compression effect (lower left), and TD cell MAC truth table (lower right).

for separated activation. The row decoder and column decoder are equipped to program and verify a single TD cell. Input driver and output LPOSC TDCs are used to apply 320 1-bit inputs concurrently and quantize the output differential delay signals to 8b digital outputs. Multibit input can be achieved by bit-serial input loading in consistent cycles and multibit signed weights are realized by mapping one signed m-bit weight to m consecutive TD cells in the same column. Besides, a local timing generator (CLK) and mode selection (MC) module are equipped to ensure proper working flow. In memory mode, a single TD cell can be selected to program under the guidance of Row/Col Addr in the SET/RESET phase. Then, one specific column of TD cells can be accessed in one Keep period according to Col Addr in the Verify phase. In CIM mode, 320 1-bit inputs are loaded during Keep = 0, then all rows of selected subarrays are triggered by the Keep signal and then DOUT is generated in the next Keep cycle. This procedure can be executed in consecutive Keep cycles as depicted in Fig. 2(b).

B. TD Cell Implementation and TD MVM Acceleration

Fig. 3 depicts the structure of the proposed TD cell, which comprises a pair of DB-based cells. The DB cell consists of a 2-stage inverter and one regular 1T1R cell as the degenerate device of the first-stage inverter. In CIM configuration, BL and SL are connected to the ground, and DI is given a rising edge to trigger this computing

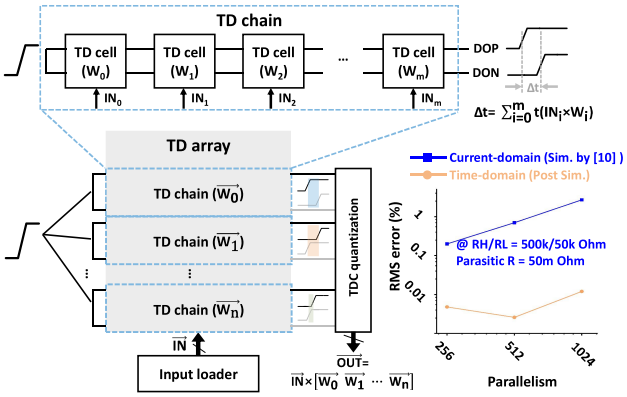

Fig. 4. TD chain for inner product (top), TD array for MVM acceleration (bottom left), and computing accuracy comparison (bottom right).

unit. When IN = 0, WL is high, and the charge stored on C1 is swiftly leaked through the on-state NM1/NM2 to the ground, where the driving capability of NM2 is larger than that of NM1. As a result, an intrinsic delay (TI) is generated. When IN = 1, WL is low, and the charge on C1 flows through NM1 and R1, resulting in a delay modulated by R1. In the SET/RESET configuration, DI = 0, and the DB cell works under the 1T1R configuration. Instead of programming conductance into specific levels, the DB cell is verified in the TD to ensure rigorous addition in CIM mode. A pair of DB cells represents a ternary weight and forms a TD cell. The multiplication of the 1b input and the ternary weight is locally performed in the TD as depicted in Fig. 3.

Besides, the TD design is proposed to suppress conductance variation. As shown in Fig. 3, the measured 1b-RRAM device exhibits a compact distribution in low conductance state while a large deviation is observed in high conductance region. Due to compression of the high conductance errors, the temporal distribution corresponding to the high conductance distribution becomes narrower. As a result, the programming error of the 1-bit weight in the TD is reduced by 26% compared to that in the conductance domain, thereby improving the reliability of RRAM-based CIM using this novel design technology co-optimization weight format.

A TD computing array is assembled based on the proposed TD cell. As shown in Fig. 4, multiple TD cells are connected serially to form a delay chain (TD chain). As a result, the delay differences contributed by each TD cell are accumulated in the TD and the inner product between the 1b-input vector and ternary-weight vector is realized. The inner production result is represented by the output relative delay of the last TD cell in the TD chain if the differential inputs of the initial TD cell are applied with the same rising edge. Multiple delay chains sharing the same inputs are triggered by the same rising edge to implement fully input/output parallel MVM in one TD array. Compared to traditional designs in the current or voltage domain, MVM in the proposed RRAM-based TD CIM is executed in a fully dynamic and energy-efficient manner without SM and parallelism limitations.

The RMS error between the simulated MAC results and the ideal MAC output under different parallelism in the current domain [10] and TD (post-simulation with RCC) are presented in Fig. 4. This indicates that the TD design could enable larger computing parallelism by suppressing IR drop effects. Additionally, the computation error remains about 0.01% even under the 1024 parallelism configuration, demonstrating its significant potential in applications characterized by high parallelism and minimal error loss.

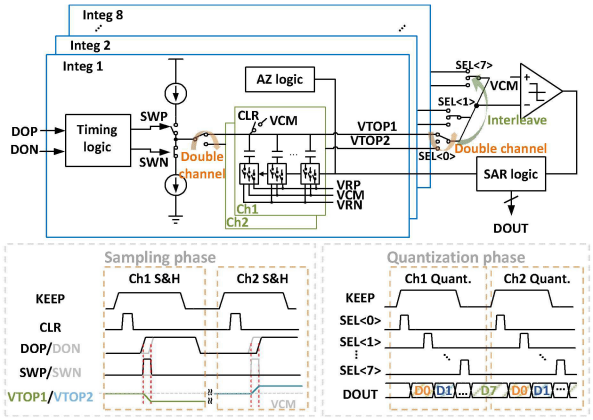

Fig. 5. Structure and working flow of LPOSC-TDC.

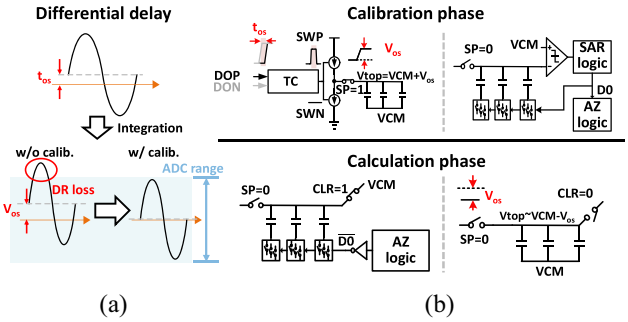

Fig. 6. (a) DR loss due to intrinsic delay mismatch from TD array. (b) Offset cancellation process of LPOSC-TDC.

C. LPOSC-TDC Structure and Implementation

The proposed LPOSC-TDC primarily consists of eight sets of passive integrators and a SAR-based ADC sharing capacitive digital-to-analog converters (CDACs), as illustrated in Fig. 5. The eight integrators first integrate the TD computation results of the eight output channels onto the CDAC during the first keep cycle, converting them into corresponding voltage values. During the second keep cycle, these eight voltages are then quantized through the shared SAR logic and comparator to produce eight sets of digital code outputs in an interleaved manner. Furthermore, each integrator contains two sets of CDAC to form a double-channel structure, ensuring MVM results from TD array sampling and quantization in every Keep period. Auto-zero (AZ) logic is applied to store the digital code for offset canceling, which will be detailed later.

However, intrinsic delay distinction exists due to mismatch between two differential DB chains in one TD chain when all inputs are equal to 0, which will cause dynamic range (DR) loss, as shown in Fig. 6(a). To tackle this problem, an offset canceling scheme in the analog domain is applied with the proposed LPOSC-TDC. As depicted in Fig. 6(b), during the calibration phase, SP = 1, the timing controller receives pulses DOP/DON from all-0-input DB array, and delay mismatch t_{OS} is sampled on CDAC as VCM + V_{OS}. This voltage is then quantized as D0 and stored in AZ logic. In the calculation phase, the top plate of CDAC is connected to VCM during CLR = 1, and the bottom plates of CDAC are connected to VRP/VRN according to the reverse code of D0. Subsequently, the bottom plates of CDACs are connected to VCM and CLR = 0, resulting in VTOP switching to a voltage level approximately equal to VCM-V_{OS}, according to the law of charge conservation. As a result, offset induced by intrinsic delay mismatch is compensated in the voltage domain by adding a voltage bias before the time difference between DOP/DON is integrated on CDAC. As a result, the offset in the TD is eliminated using this offset cancellation in the voltage domain, incurring only 3% extra area and <1% power overhead.

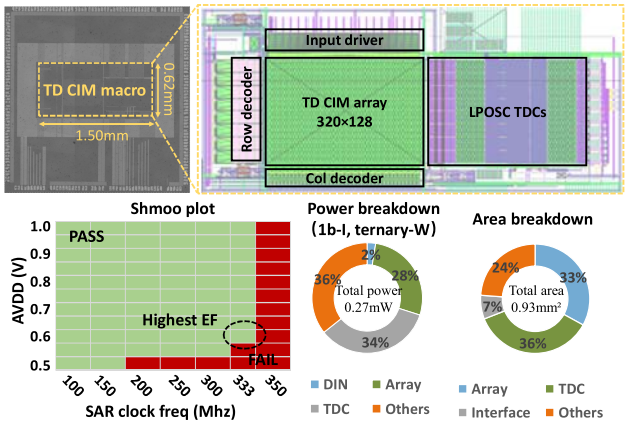

Fig. 7. Die micrograph with measured power breakdown, area breakdown, and shmoo.

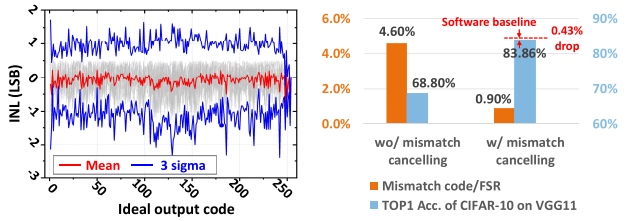

Fig. 8. Measured INL errors across LPOSC-TDCs (left), offset elimination and network inference accuracy improvement (right).

III. Chip Measurement and Results

A. Test-Chip and Overview

Fig. 7 shows the micrograph of the prototype chip fabricated in the UMC 28-nm technology along with the measured power breakdown, area breakdown, and shmoo characteristics. This 80K macro occupies an area of 0.96 mm^2. MAC between multibit input and weight are processed externally using shifting-and-adding (S&A), as the implementation of S&A on chip incurs negligible area and power overhead. Under 0.6/0.75/0.55 V analog/digital/array power supply and 333-MHz SAR clock frequency, this design achieves a peak EF of 1251 and 70 TOPS/W configuring with 1b-I/ternary-W and 4b-I/signed 4b-W under full parallelism.

B. TDC and MVM Performance

INL errors of LPOSC-TDCs across different output channels are presented in Fig. 8, with less than ±1.5 LSB INL linearity of LPOSC-TDCs across different output channels. Additionally, the measurement results for all LPOSC-TDCs under all-0 input configurations of the TD-CIM array are presented. The measured output code corresponding to intrinsic delay mismatch from the TD array indicates the offset is compressed from 4.6% to 0.9% of the full-scale range (FSR), resulting in an improvement of the TOP-1 accuracy of CIFAR10 on VGG-11 from 68.8% to 83.9%, which incurs only 0.43% accuracy degradation compared to software baseline. As shown in Fig. 9, under fully parallel with bit-serial 4-bit input, signed 4-bit weight, and 12-bit output configuration, 0.94% RMS error and 1.90% STD error can be achieved across 2k random input vectors, validating the feasibility of this TD RRAM CIM macro for SoC integration.

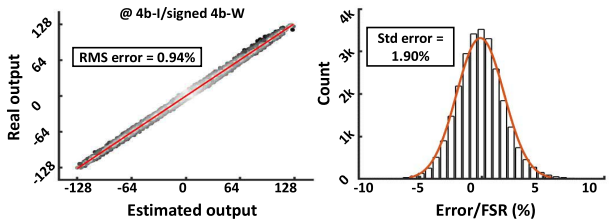

Fig. 9. MVM measured results—real output versus estimated output (left) and error distribution (right).

TABLE I
COMPARISON WITH PRIOR ART

	This work		ISSCC 2022[11]		ISSCC 2021[8]		VLSI 2021[6]	ISSCC 2022[9]		VLSI 2023[7]	ISSCC 2023[3]	
Technology/nm	28		28		22		14	22		40	22	
Computing domain	Time		Time		Voltage		Current	Voltage		Current	Current	
MVM devices	SLC-RRAM		SRAM		SLC-RRAM		4b-PCM	SLC-RRAM		SLC-RRAM	SLC/MLC-RRAM	
Parallelism(single array MAC)	320		64		4		256	8		256	Up to 128	
Input/W/output precision (bit)	1/ternary/8	4/signed-4/12	INT4	INT8	1/2/4	4/4/10	8/4/8	1/1/3	8/8/19	1/1/6	INT4	INT8
Throughput (TOPS)	0.341	0.021	4.965	1.241	0.418	0.099	1.008	5.120	0.142	up to 0.184	26.970	8.210
Area efficiency (TOPS/mm^2)	0.367	0.023	0.86-4.08		0.070	0.017	1.587	0.284	0.008	up to 7.008	2.660	0.810
Nomalized area efficiency (TOPS/mm^2/b)	0.582	0.455	-		0.139	0.264	50.784	0.284	0.506	up to 7.008	42.560	51.840
Energy efficiency (TOPS/W)	1250.5[c]	70.0[c]	148.1	37.0	197.5	47.3	10.5	1286.4	21.6	350.0	287.6	76.5
Normalized energy efficiency[a] (TOPS/W/b)	1982.0	1387.9	2369.6	2368.0	395.0	756.8	336.0	1286.4	1382.4	350.0	4601.6	4896.0
Output ratio	0.81	0.69	1.00		1.00		0.40	1.00		0.75	1.00	
FOM (=Normalized EF [b]output ratio)	1600.5	964.0	2369.6	2368.0	395.0	756.8	134.4	134.4	1286.4	1382.4	262.5	4601.6	4896.0

[a] Output ratio is defined as MAC output precision / full precision; [b] FOM equals to normalized EF × output ratio. [c] Exclude power consumption apart from TD array, input driver and TDCs.

C. Comparison With Prior Art

A comparison with CIM prior art featuring accelerating MVM is presented in Table I. This 28-nm fabricated TD CIM macro achieves a SOTA normalized EF of 1982.0 and 1387.2 TOPS/W/bit under 1b-I/ternary-W and 4b-I/signed 4b-W configuration, respectively. Since the accuracy loss exists in high-parallel analog CIM due to quantization error, a figure-of-merit (FOM) comparison should take the output ratio into account, which can be expressed as output precision divided by full precision. This letter provides a peak FOM of 1600.5 TOPS/W/bit, indicating its feasibility for edge applications.

IV. CONCLUSION

In this letter, we propose a 28-nm fabricated TD CIM macro based on RRAM for the first time. Using the proposed RRAM-based TD MVM scheme along with LPOSC TDC, this macro achieves a SOTA EF and computing accuracy compared to its voltage-domain and current-domain nonvolatile memory-based counterparts. Besides, reduced deviation of mapped weight using TD cell and minimal error under high-parallel MVM demonstrate the reliability of this RRAM-based TD CIM macro. The measurement results confirm the suitability of the proposed TD CIM macro for edge devices with extremely high EF requirements

REFERENCES

[1] J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, and S. Han, "Tiny machine learning: Progress and futures," *IEEE Circuits Syst. Mag.*, vol. 23, no. 3, pp. 8–34, Oct. 2023.

[2] H. Jia, H. Valavi, Y. Tang, J. Zhang, and N. Verma, "A programmable heterogeneous microprocessor based on bit-scalable in-memory computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 9, pp. 2609–2621, Sep. 2020.

[3] H.-H. Hsu et al., "A nonvolatile AI-Edge processor with SLC–MLC hybrid ReRAM compute-in-memory macro using current–voltage-hybrid readout scheme," *IEEE J. Solid-State Circuits*, vol. 59, no. 1, pp. 116–127, Jan. 2024.

[4] J.-W. Su et al., "A 28 nm 64 Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in *Proc. IEEE ISSCC*, 2020, pp. 240–242.

[5] W. Wan et al., "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, pp. 504–512, Aug. 2022.

[6] 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," in *Proc. IEEE Symp. VLSI Circuits*, 2021, pp. 1–2.

[7] 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," in *Proc. IEEE Symp. VLSI Circuits*, 2023, pp. 1–2.

[8] C.-X. Xue et al., "A 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7TOPS/W for tiny AI edge devices," in *Proc. IEEE ISSCC*, 2021, pp. 245–247.

[9] J.-M. Hung et al., "An 8-Mb DC-current-free binary-to-8b precision ReRAM nonvolatile computing-in-memory macro using time-space readout with 1286.4-21.6TOPS/W for edge-AI devices," in *Proc. IEEE ISSCC*, 2022, pp. 182–184.

[10] Q. Qi et al., "Hybrid precoding with a fully-parallel large-scale analog RRAM array for 5G/6G MIMO communication system," in *Proc. IEEE IEDM*, 2022, pp. 33.2.1–33.2.4.

[11] P.-C. Wu et al., "A 28nm 1Mb time-domain computing-in-memory 6T-SRAM macro with a 6.6ns latency, 1241GOPS and 37.01TOPS/W for 8b-MAC operations for edge-AI devices," in *Proc. IEEE ISSCC*, 2022, pp. 1–3.