SRAM In-Memory Computing Macro With Delta-Sigma Modulator-Based Variable-Resolution Activation

Vasundhara Damodaran, Ziyu Liu, Jian Meng, Jae-Sun Seo, *Senior Member, IEEE*, and Arindam Sanyal, *Member, IEEE*

***Abstract*—This letter presents an SRAM-based compute-in-memory (CIM) macro that uses 1-bit ΔΣ modulators to convert input and output activations to binary pulse waveform. The SRAM macro uses switched-capacitors for vector matrix multiplications and together with binary input activation improves linearity compared to current-domain SRAM CIM macros and allows reconfigurable activation resolution. The proposed macro is fabricated in 65 nm and benchmarked on MNIST and CIFAR-10 datasets with accuracies of 98.67% and 89.85%, respectively, with energy-efficiency in the range of 15.4–138.6 TOPS/W.**

***Index Terms*—Compute-in-memory (CIM), convolutional neural network (CNN), delta-sigma, static random access memory.**

I. Introduction

The most common operation in deep-learning models is multiply-and-accumulate (MAC) whose regularity and parallelism makes it suitable for hardware acceleration. However, the amount of memory access needed to perform MAC operations severely limits energy efficiency in conventional digital accelerators. Hence, compute-in-memory (CIM) has emerged as an appealing alternative to digital accelerators. By reducing communication costs of bringing together many input activations, neuron weights and distributing output activations, CIM breaks the von-Neumann bottleneck and improves energy-efficiency. Out of different CIM techniques, SRAM-CIM is widely used due to its high-energy efficiency and easy integration with CMOS ICs [1].

The growing complexity of datasets has prompted the need for MAC operations with multibit activations and weights. Multibit activations require either—a) accumulation of partial outputs using digital peripheral [2] at the cost of higher energy or b) multibit digital-to-analog converter (DAC) [3] or c) pulse-width modulation (PWM) that converts multilevel activation to binary pulse with proportionally varying width but is limited by linearity of PWM converter [4] or quantization error [5]. In addition, current-domain SRAM-CIM architectures suffer from nonlinearity in each bitcell due to the requirement of maintaining a constant current sink in each activated SRAM bitcell [5]. For large value of MAC output, the bitline (BL/RBL) voltage sees a large swing which makes the current in each activated SRAM bitcell nonlinear. This nonlinearity issue is addressed through charge-domain accumulation in capacitive SRAM-CIM architectures [1], [6]. However, linearity of MAC is still limited since analog input is sampled on the capacitor in through an nMOS switch. The analog input activation modulates threshold voltage V_t of the nMOS switch making the voltage sampled on the capacitor nonlinear.

This letter presents a switched-capacitor-based ΔΣ SRAM CIM technique that addresses the above challenges. The key contribution of this letter is the introduction of ΔΣ modulators for converting activations to binary pulse trains. The use of ΔΣ modulator reduces quantization error compared to [5] and allows variable resolution for input and output activations without requiring changes in hardware. While a recent work [7] has used ΔΣ for SRAM CIM, they use it to reduce input swing for correlated input activations and require a separate analog-to-digital converter (ADC) for digitizing output activations. Additionally, Chen et al. [7] lacked the ability to change resolution of input and output activations. Fabricated in 65-nm CMOS, the proposed ΔΣ SRAM CIM macro is benchmarked on MNIST and CIFAR-10 datasets achieving mean accuracies of 98.67% and 89.85%, respectively, with maximum energy efficiency of 138.6 TOPS/W.

II. Proposed Architecture

Fig. 1 shows architecture of the proposed ΔΣ SRAM array and the 9T1C bitcell schematic. 1-b ΔΣ modulators convert analog inputs into binary pulse train activation and apply to the SRAM bitcells through the RWL lines. The SRAM cells use switched-capacitor circuits to perform charge-domain multiplication of input activation and model weights stored in the SRAM cells similar to that in [6]. Accumulation is performed in charge-domain on the RBL bitlines and sent to output 1-b ΔΣ modulators and digitally decimated. Quantization noise-shaping in the ΔΣ modulators ensure lower-in-band quantization error than the analog-to-binary pulse conversion technique using counters in [5]. Accumulation in charge-domain through charge redistribution in the proposed architecture results in higher linearity than current-domain accumulation in conventional SRAM-CIM macros. Use of binary input activation and CMOS switch to sample input activation on the capacitor as shown in Fig. 1(b) improves linearity compared to [1] by removing dependency of sampled voltage on input activation. The CMOS switch also allows the SRAM bitcell to operate from low-supply voltage and improves energy efficiency. The layout of a single SRAM bitcell is shown in Fig. 1(b). A 2.4fF MOM capacitor is used to realize SRAM capacitor C_{sram} which is placed on top of the SRAM cell.

Fig. 2 shows the schematic of a single slice of the proposed macro using 4-bit signed weights where T_s is sampling period of ΔΣ modulators. Compensation capacitors are used to ensure all RBL lines see the same capacitive load during computations. The capacitors and RBL lines are discharged during ϕ_1. During ϕ_2, the SRAM cells compute the MAC products on each RBL line which are charge-shared with the binary weighted compute capacitor bank. The compute capacitors are disconnected from the SRAM cells during ϕ_3 and charge-shared with additional balancing capacitors to ensure correct binary weighted MAC result with sign bit operation. At the

Manuscript received 14 July 2023; revised 12 September 2023; accepted 11 October 2023. Date of publication 24 October 2023; date of current version 4 December 2023. This work was supported in part by NSF under Grant CCF-1948331, and in part by the SRC/DARPA JUMP 2.0 CoCoSys Center. This article was approved by Associate Editor Sudipto Chakraborty. *(Corresponding author: Vasundhara Damodaran.)*

Vasundhara Damodaran, Ziyu Liu, and Arindam Sanyal are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: vdamoda2@asu.edu).

Jian Meng and Jae-Sun Seo are with the Department of Electrical and Computing Engineering, Cornell Tech, New York City, NY 10044 USA.

Digital Object Identifier 10.1109/LSSC.2023.3327213

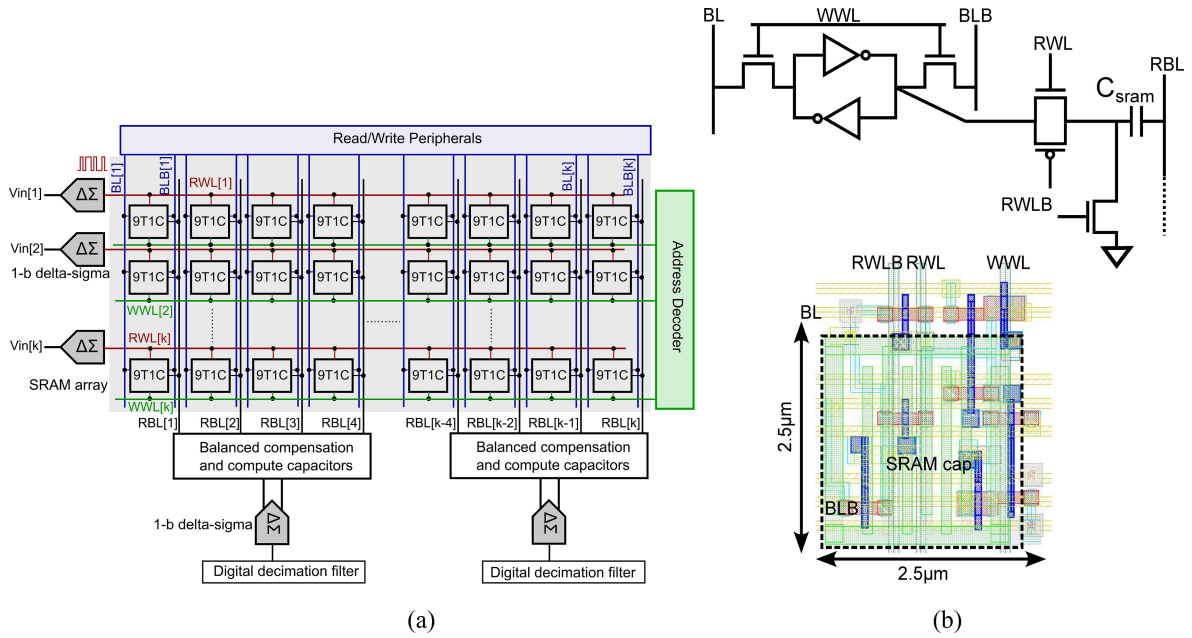
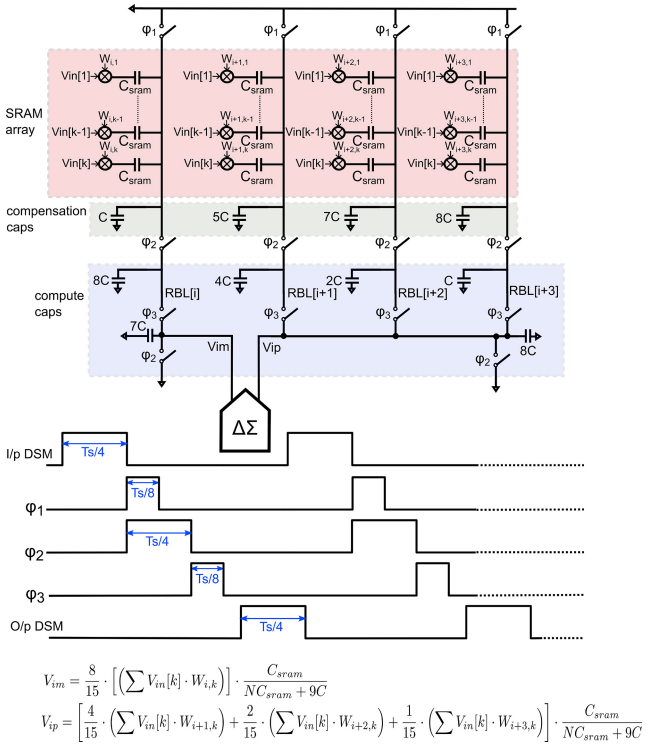


Fig. 1. (a) Proposed ΔΣ-based pulsed activation CIM macro. (b) Schematic of 9T1C SRAM bitcell and layout.



$$V_{im} = \frac{8}{15} \cdot \left[\left(\sum V_{in}[k] \cdot W_{i,k}\right)\right] \cdot \frac{C_{sram}}{NC_{sram} + 9C}$$

$$V_{ip} = \left[\frac{4}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+1,k}\right) + \frac{2}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+2,k}\right) + \frac{1}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+3,k}\right)\right] \cdot \frac{C_{sram}}{NC_{sram} + 9C}$$

Fig. 2. Circuit schematic of a single slice of the CIM macro and associated timing diagram.

end of ϕ_3, the sign-bit output V_{im}, and the remaining 3-bit output V_{ip}, are applied differentially to the output ΔΣ modulator as shown in Fig. 2 that provides a binary pulse train as output.

Fig. 3(a) shows schematic of the input 1-bit ΔΣ modulator. The output ΔΣ modulator has the same architecture with the exception of differential inputs. The ΔΣ modulator is fully dynamic to reduce power consumption. A first-order loop filter is used in this letter since the SRAM macro does not need a very high precision. The amplifier needs only a moderate gain since the 1-bit quantizer only uses the sign information of its inputs rather than amplitude to make decisions. The tradeoff with moderate gain of the amplifier is weak in-band high-pass quantization noise shaping performance. This is reflected in the transfer function of the ΔΣ modulator given by

$$D_{\text{out}} = \frac{z^{-1}V_x + \left(1 + 2/G - (1 + 1/G)z^{-1}\right)Q}{1 + 2/G - z^{-1}/G} \tag{1}$$

where V_x is input to the ΔΣ, Q is quantization error, and G is gain of the amplifier. The in-band quantization error is attenuated by $1/(G+1)$ which corresponds to -40 dB for $G = 100$. A single-stage dynamic amplifier shown in Fig. 3(a) is used due to the moderate gain requirements. The amplifier uses a capacitor as tail current source [8]. The capacitor is discharged during the ΔΣ sampling phase (ϕ_s) and provides bias current to the amplifier during the amplification phase ($\overline{\phi_s}$). During the amplification phase, voltage across the capacitor rises which reduces gate-to-source voltage of the input transistors and increases the open-loop voltage gain until the transistors enter subthreshold where the amplifier gain is maximum. Fig. 3(b) shows the frequency response of the ΔΣ modulator. The ΔΣ modulator has an SNDR of 28.1 dB at over-sampling ratio (OSR) of 8. Fig. 3(c) shows the classification accuracy on MNIST dataset as a function of amplifier gain G at OSR of 8. The classification accuracy increases by less than 0.4% as G is varied from 10 to 10 000 thus demonstrating low dependency of accuracy on amplifier gain thanks to the 1-bit ΔΣ architecture. G is set to 100 in this letter.

III. Measurement Results

The test-chip is fabricated in 65-nm process and the die photograph is shown in Fig. 4. The core circuits occupy an area of 0.1 mm^2 with the 64×64 SRAM array occupying an area of 0.03 mm^2. The test-chip operates from a supply voltage of 0.5–1.2 V for the SRAM array and 1.2 V for the other circuits. The operating speed of the entire macro is $1/T_s = 325$ kHz which is limited by buffers driving the sampling capacitor in the output ΔΣ. Charge sharing between capacitors in the SRAM array and the compute and compensation capacitors reduced the input common-mode voltage of the buffers to 105–250 mV which pushed the buffer input transistors

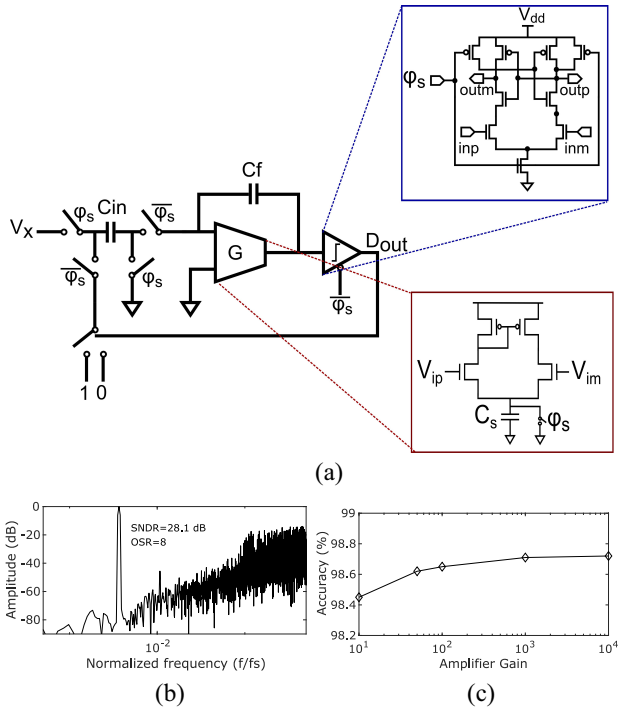


Fig. 3. (a) Schematic of 1-bit ΔΣ modulator. (b) FFT plot of ΔΣ modulator with sinusoidal input. (c) Accuracy versus amplifier gain at OSR of 8.

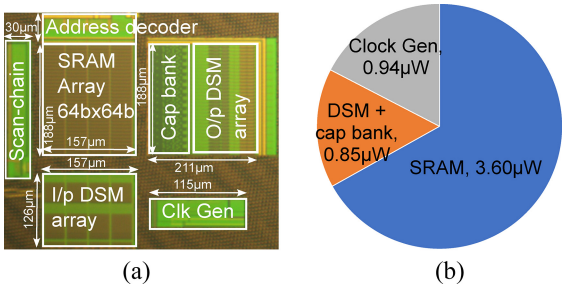


Fig. 4. (a) Die micro-photograph and (b) power breakdown at 1.2 V.

to subthreshold and limited the operating speed. This can be remedied by adopting floating inverter amplifiers [9] for buffer that is insensitive to input common-mode voltage. The ΔΣ modulators and clock generator consume 1.8 μW from 1.2-V supply while the SRAM array consumes 0.6–3.6 μW from 0.5–1.2-V supply. Offset in the input and output ΔΣ modulators are calibrated once in the foreground before characterization of the complete macro. Performance of the macro is evaluated on MNIST dataset and CIFAR-10 dataset. Input sizes of the MNIST and CIFAR-10 data are $1 \times 28 \times 28$ and $3 \times 32 \times 32$, respectively. For MNIST, we used convolutional neural network (CNN) with 2 convolutional layers and three fully connected (FC) layers: 6C3-MP2-16C3-MP2-120FC-84FC-10FC. For CIFAR-10, we evaluated VGG-like CNN [10] that has six convolutional layers and two FC layers: 128C3-128C3-MP2-256C3-256C3-MP2-512C3-512C3-MP2-1024FC-10FC, where nC3 represents a convolutional layer with n 3×3 filters, mFC is an FC layer with m neurons, and MP2 is a max-pooling layer with 2×2 pooling size.

Fig. 5(a) shows the measured average differential RBL voltage ($V_{ip} - V_{im}$) versus MAC value for a SRAM bitcell, with the RBL voltage obtained by spatial averaging of bitcells in a column and the output DSM bitstreams 1M times. The maximum differential RBL swing is limited by charge sharing with balancing and compensation capacitors and linearity of the buffer driving the output DSM.

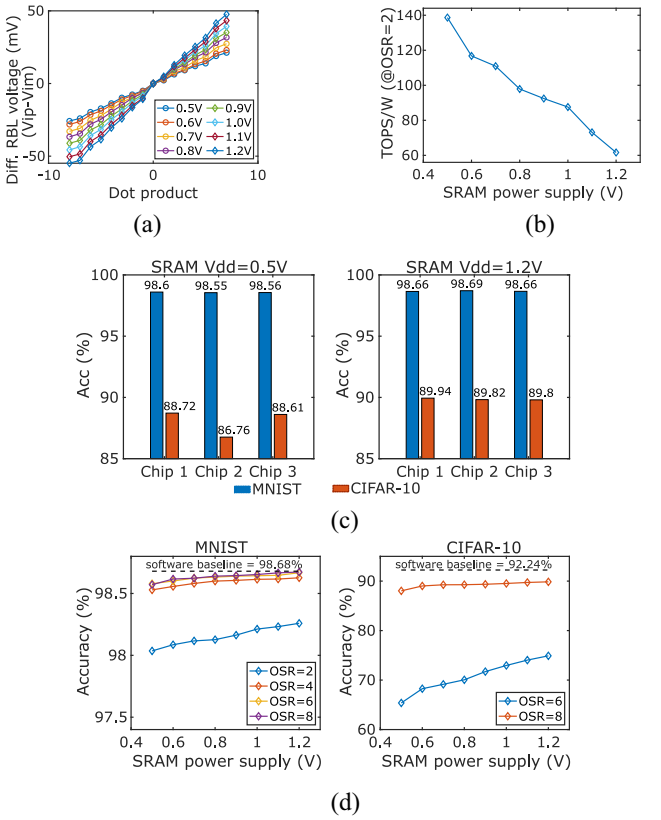


Fig. 5. Measured (a) mean differential RBL voltage and (b) energy efficiency as a function of SRAM supply (c) accuracy of three test-chips on MNIST and CIFAR-10 at OSR = 8 (d) accuracy of software baseline and mean accuracy of three test-chips versus OSR and SRAM supply.

Nonlinearity in the RBL transfer curve is due to static random mismatch between individual bitcells and capacitors, as well as charge-injection error in the switched-capacitor circuits. Linearity of the dot product is measured by the ratio of maximum RBL swing (Δ RBL) to worst-case root-mean-squared-error (RMSE) and varies from 40.5 at 0.5 V to 68.8 at 1.2-V SRAM power supply [11]. In comparison, Δ RBL/RMSE varies from 29.1 [12] to 63 [6] for state-of-the-art SRAM-CIM macros without ΔΣ modulation which demonstrate better linearity of the proposed architecture. Fig. 5(b) plots the energy-efficiency of the proposed macro as a function of SRAM supply voltage at OSR = 2. The MNIST and CIFAR-10 datasets are randomly partitioned into 90% split for training and 10% split for testing. Fig. 5(c) shows the measured accuracy for 3 test-chips for both datasets at SRAM supply voltages of 0.5 and 1.2 V, while Fig. 5(d) shows the measured accuracies averaged for three chips for both datasets as a function of OSR and SRAM supply voltages. The baseline accuracy of the floating point software neural network for MNIST is 99.06% and the accuracy drops to 98.68% when converted to fixed-point analog implementation. The baseline accuracy of the floating point software neural network for CIFAR-10 is 92.3% and the accuracy drops to 92.24% for fixed-point analog implementation. Table I compares this letter with state-of-the-art SRAM-CIM macros benchmarked on MNIST and CIFAR-10. The proposed architecture achieves high linearity and competitive classification accuracy and energy efficiency as state-of-the-art. The ΔΣ architecture allows change in resolution of input and output activations by trading off speed for accuracy and without requiring changes in the macro hardware. Limitation of the proposed macro are—1) area overhead of ΔΣ modulators which occupy 65% of the core area and 2) latency as the MAC results are computed over

TABLE I
COMPARISON WITH STATE-OF-THE-ART SRAM-CIM

	This Work	ISSCC'23 [7]	JSSC'20 [1]	JSSC'21 [12]	CICC'22 [6]
Process	**65nm**	22nm	65nm	40nm	28nm
Supply	**0.5-1.2V(SRAM) 1.2V(DSM)**	0.7–0.8V	0.8V(SRAM) 0.6V(ADC)	1.1V	0.8-1V
SRAM type	**9T1C**	6T+3T1C	8T1C	8T	9T1C
Macro size	**64× 64**	16kB	256× 64	512× 64	128× 128
Bitcell area	**6.2μm^2**	3.2μm^2	3.3μm^2	1.9μm^2	1.3μm^2
MAC compute	**charge-domain**	charge-domain	charge-domain	current-domain	charge-domain
Mac linearity[2]	**40.5@0.5V 68.8@1.2V**	24.7/82[1]	44	29.1	63
Readout	$\Delta\Sigma$	SAR+$\Delta\Sigma$	Flash	SAR	SAR
Activation precision	**1.6b@OSR=2 4.6b@OSR=8**	8b	1b	2b	4b
Weight precision	**4b**	8b	1b	2b	4b
MNIST acc. (%)	**98.57[3]@0.5V 98.67[3]@1.2V @OSR=8(MLP)**	–	98.30 (MLP)	98.20 (LeNet)	–
CIFAR-10 acc. (%)	**88.03[3]@0.5V 89.85[3]@1.2V @OSR=8(VGG)**	–	85.50 (VGG)	85.50 (VGG)	91.86 (VGG)
Efficiency (TOPS/W)[4]	**15.4–138.6**	16–21.4	671.5	82	139–177
Throughput (GOPS)[4]	**0.33 @OSR=2**	490–600	1638	122	56–255
Power (μW)	**5.4@1.2V 2.4@0.5V**	22897–37500	2439.3	2975.6	316.4–1834.5

[1] without/with ADC calibration; [2]maximum RBL swing/RMSE; [3]average of 3 chips; [4]1 MAC = 2 OP

multiple cycles. Performance of the proposed architecture can be further improved through technology scaling due to the highly digital nature of the macro.

REFERENCES

[1] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, "C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism," *IEEE J. Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, Jul. 2020.

[2] S. Yin, Z. Jiang, M. Kim, T. Gupta, M. Seok, and J.-S. Seo, "Vesti: Energy-efficient in-memory computing accelerator for deep neural networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 1, pp. 48–61, Jan. 2020.

[3] H. Wang, R. Liu, R. Dorrance, D. Dasalukunte, D. Lake, and B. Carlton, "A charge domain SRAM compute-in-memory macro with C-2C ladder-based 8-bit MAC unit in 22-nm FinFET process for edge inference," *IEEE J. Solid-State Circuits*, vol. 58, no. 4, pp. 1037–1050, Apr. 2023.

[4] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2018, pp. 488–490.

[5] Q. Dong et al., "A 351TOPS/W and 372.4GOPS compute-in-memory SRAM macro in 7 nm FinFET CMOS for machine-learning applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 242–244.

[6] 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," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, 2022, pp. 1–2.

[7] P. Chen et al., "A 22 nm delta-sigma computing-in-memory ($\delta\Sigma$ CIM) SRAM macro with near-zero-mean outputs and LSB-first ADCs achieving 21.38 TOPS/W for 8b-MAC edge AI processing," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 140–142.

[8] B. J. Hosticka, "Dynamic CMOS amplifiers," *IEEE J. Solid-State Circuits*, vol. 15, no. 5, pp. 881–886, Oct. 1980.

[9] X. Tang et al., "An energy-efficient comparator with dynamic floating inverter amplifier," *IEEE J. Solid-State Circuits*, vol. 55, no. 4, pp. 1011–1022, Apr. 2020.

[10] I. Hubara, M. Courbariaux, D. Soudry, R. EL-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[11] V. Damodaran, Z. Liu, J.-S. Seo, and A. Sanyal, "A 138-TOPS/W delta-sigma modulator-based variable- resolution activation in-memory computing macro," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, 2023, pp. 1–2.

[12] S. Jain, L. Lin, and M. Alioto, "±CIM SRAM for signed in-memory broad-purpose computing from DSP to neural processing," *IEEE J. Solid-State Circuits*, vol. 56, no. 10, pp. 2981–2992, Oct. 2021.