A High-Throughput Energy–Area-Efficient Computing-in-Memory SRAM Using Unified Charge-Processing Network

Ying-Tuan Hsu, Chun-Yen Yao, Tsung-Yen Wu, Tzi-Dar Chiueh, *Fellow, IEEE,* and Tsung-Te Liu, *Member, IEEE*

***Abstract*—This letter presents a computing-in-memory (CIM) static random-access memory (SRAM) using efficient data processing and conversion circuits to enhance the throughput, energy, and area efficiency performance. The proposed unified charge-processing network simultaneously provides both signal processing and data conversion functions with maximum resource utilization, realizing significant performance improvements in energy and area efficiency by 37.5% and 15.4%, respectively. Measurement results from a prototype fabricated in a 28-nm CMOS technology show that the proposed CIM SRAM achieves a high throughput of 186.18 GOPS, with energy and area efficiencies of 41.87 TOPS/W and 3288.4 GOPS/mm^2, which demonstrates the performance improvements of 2.26×, 1.12×, and 2.89×, respectively, when compared with the state-of-the-art results. The proposed CIM SRAM can achieve 88.87% classification accuracy on the CIFAR-10 dataset.**

***Index Terms*—Charge processing, computing-in-memory (CIM), convolutional neural network (CNN), machine learning (ML), static random-access memory (SRAM).**

I. Introduction

To enable the emerging artificial intelligence (AI) applications, the computing-in-memory (CIM) architecture is proposed to enhance the computation efficiencies of machine-learning (ML) processing tasks [1]–[9]. The CIM architecture avoids the expensive data movement by performing the computation directly inside the memory. This results in substantially higher energy efficiency than the conventional Von Neumann computation architecture, whose energy consumption is dominated by the data movement [10]. A CIM static random-access memory (SRAM) has the advantage of processing information directly acquired from the bit lines, which minimizes data movement energy and performs efficient operations of multiplication and accumulation (MAC) in parallel. Fig. 1 compares the performances of recent ML accelerators and CIM SRAMs. While most of the CIM SRAMs could realize higher energy efficiency, their operating frequency and throughput are relatively lower than those of the existing ML accelerators. This seriously limits the application space of CIM SRAM to AI applications with low or moderate performance requirements. Therefore, it is essential to further enhance the throughput and efficiency of a CIM SRAM, so its speed and throughput performances could be comparable with the state-of-the-art ML accelerators for high-performance AI applications.

Compared to a conventional SRAM, mainly composed of a bit-cell array and peripheral circuits, a CIM SRAM requires additional MAC processing and data conversion circuits to support ML computation tasks. These extra circuits become the performance bottleneck and ultimately limit the throughput, energy, and area efficiency of a CIM SRAM [9]. Traditional charge summing [1], [2] and current summing [3] MAC architectures for CIM SRAM can perform efficient MAC computations. However, their classification performances are seriously limited by the nonideality of the SRAM access transistors, such as the transistor mismatch and the data-dependent current characteristics. Compared to the previous two MAC methods, the charge redistribution architecture [4] demonstrates better classification performance for high-performance AI applications. Still, it suffers from worse area and energy efficiency due to the additional capacitors required for the charge processing operation.

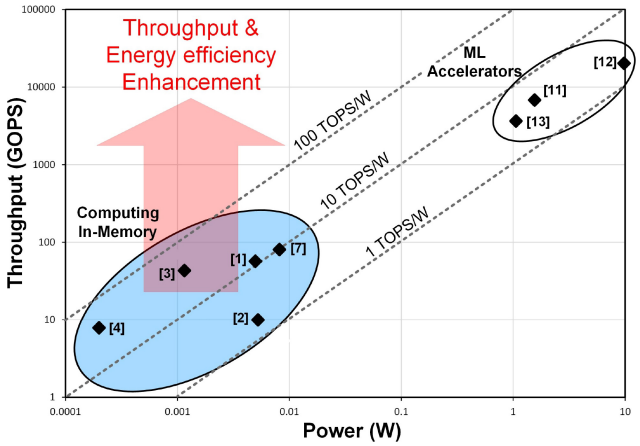

Fig. 1. Comparison of CIM SRAMs and the state-of-the-art ML accelerators.

The data conversion circuits, including digital-to-analog converter (DAC) and analog-to-digital converter (ADC) interfaces, typically consume a huge amount of energy in a CIM SRAM, more than 60% reported in [9]. The DAC with PWM activation inputs used in [4] realizes an efficient data conversion performance for CIM SRAM. However, the PWM scheme suffers from a substantial performance tradeoff between the output resolution and conversion time. This makes the PWM approach infeasible for high-performance AI applications. On the other hand, the binary-weighted current-steering DAC [1] can achieve high conversion speed, but it consumes extra static energy during the data conversion. As for the choice of ADC, compared to the integrating ADC [4] and the flash ADC [9], the SAR ADC [3] potentially offers an energy-efficient data conversion solution to high-performance CIM SRAMs that require both high throughput and resolution performances. However, similar to the charge redistribution MAC architecture, the corresponding area overhead from the capacitor array would seriously degrade the overall efficiency of CIM SRAM.

In this letter, we present a high-throughput, energy–area-efficient CIM SRAM for high-performance AI applications. This work overcomes the traditional performance bottleneck of CIM SRAM resulting from the data processing and conversion circuits by employing

Manuscript received June 6, 2021; revised July 9, 2021; accepted July 30, 2021. Date of publication August 10, 2021; date of current version August 26, 2021. This work was supported in part by the Ministry of Science and Technology, Taiwan; in part by the Intelligent and Sustainable Medical Electronics Research Fund in National Taiwan University; and in part by MediaTek Inc. under Contract MTKC-2021-0167. This article was approved by Associate Editor Mingoo Seok. *(Corresponding author: Tsung-Te Liu.)*

The authors are with the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: ttliu@ntu.edu.tw).

Digital Object Identifier 10.1109/LSSC.2021.3103759

2573-9603 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: National Taiwan University. Downloaded on March 09,2025 at 06:52:29 UTC from IEEE Xplore. Restrictions apply.

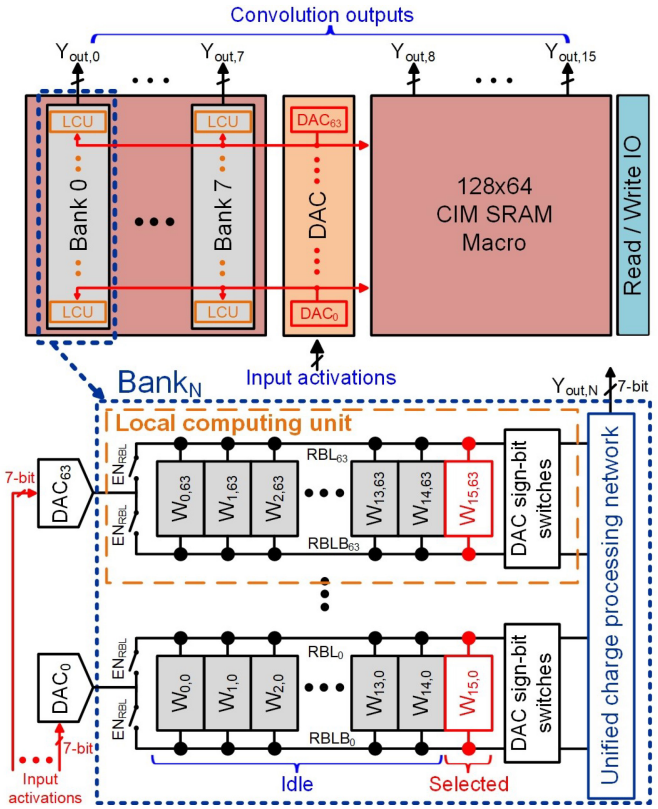

Fig. 2. Overall architecture of the CIM SRAM.

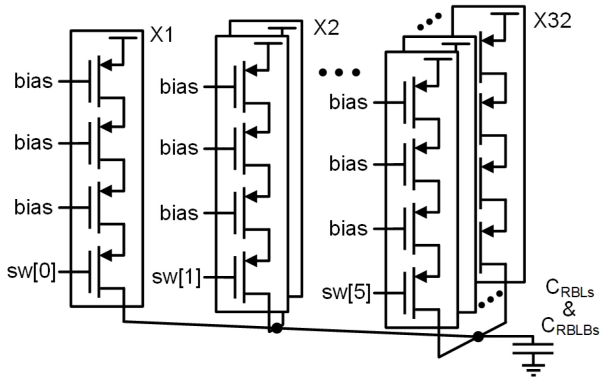

Fig. 3. Schematic of the dynamic binary-weighted current-steering DAC.

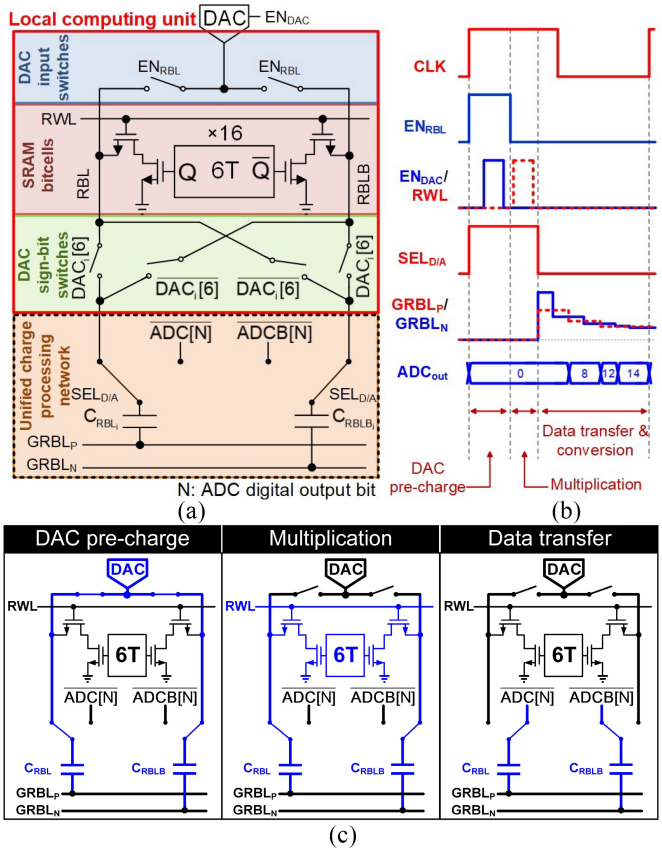

Fig. 4. (a) Structure of the LCU. (b) Timing diagram of CIM operation. (c) Different operation steps of the LCU.

1) a dynamic current-steering DAC that enables the high-speed operation with improved energy efficiency and linearity performance by 2.01× and 0.65 ENOB, respectively and 2) a unified charge-processing network (UCPN) that simultaneously provides highly efficient signal processing and data conversion functions, effectively boosting the area and energy efficiencies by 1.15× and 1.38×, respectively. The 28-nm CIM SRAM design achieves a high throughput of 186.18 GOPS, with energy and area efficiency of 41.87 TOPS/W and 3288.4 GOPS/mm^2 for 7-b signed magnitude activation, 1-b weight, and 5-b output. These represent 2.26×, 1.12×, and 2.89× performance improvements in throughput, energy efficiency, and area efficiency, respectively, when compared to the state-of-the-art high-performance CIM SRAM designs [3], [7].

The remainder of this letter is organized as follows. Section II introduces the proposed CIM architecture using UCPN and the corresponding circuit design in detail. The implementation and the measurement results of the CIM SRAM prototype are presented in Section III.

II. Proposed CIM Architecture and Circuit Design

A. System Overview

Fig. 2 shows the overall architecture of the proposed CIM SRAM. The proposed CIM SRAM consists of 64 7-bit DACs, 16 computation banks, and a UCPN that offers data conversion capability equivalent to 16 7-bit ADC circuits with a maximum filter depth of 64. In each computation, the DACs would first convert the digital activation inputs into the corresponding analog voltages for precharging the computation banks. Afterward, the multiplication operations would be performed in the computation banks. Finally, the UCPN would accumulate the multiplication results and generate the digitized outputs.

B. Digital-to-Analog Converter

To further improve the energy efficiency while maintaining the speed performance of CIM SRAM, a dynamic binary-weighted current-steering DAC is employed in this design, as shown in Fig. 3. It consists of a binary-weighted current source array and an output capacitor mainly composed of capacitors on RBL and RBLB, C_{RBLs} and C_{RBLBs}. Compared to the previous high-speed DAC design [1], the proposed structure eliminates both the quiescent current and voltage headroom problems caused by the diode-connected nMOS transistor while maintaining a similar speed advantage. This effectively increases the energy efficiency and linearity of DAC by 2.01× and 0.65 ENOB, respectively.

C. Computation Bank

The computation banks would conduct multiplication operations after the corresponding analog activation inputs have been precharged on RBLs. Each computation bank consists of 64 local computing units (LCU), as shown in Fig. 2. Fig. 4(a) shows the schematic of the LCU composed of two DAC input switches, 16 10T SRAM bitcells,

Authorized licensed use limited to: National Taiwan University. Downloaded on March 09,2025 at 06:52:29 UTC from IEEE Xplore. Restrictions apply.

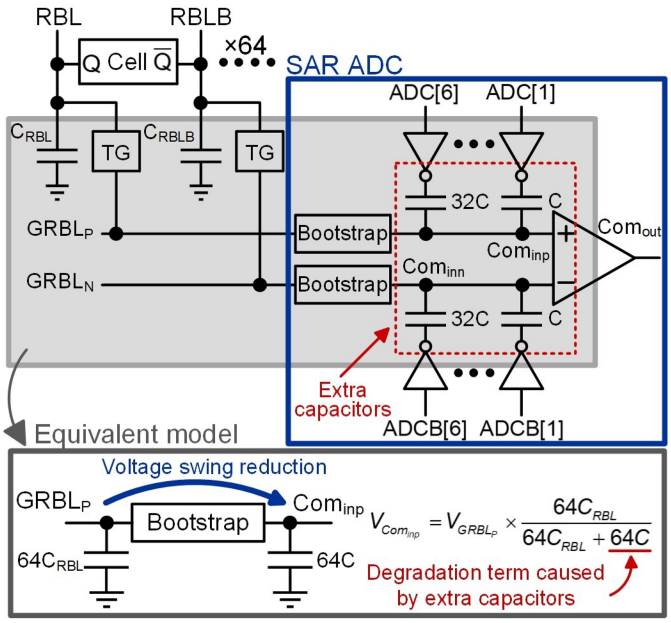

Fig. 5. Problems of the charge redistribution circuit connected to a SAR ADC.

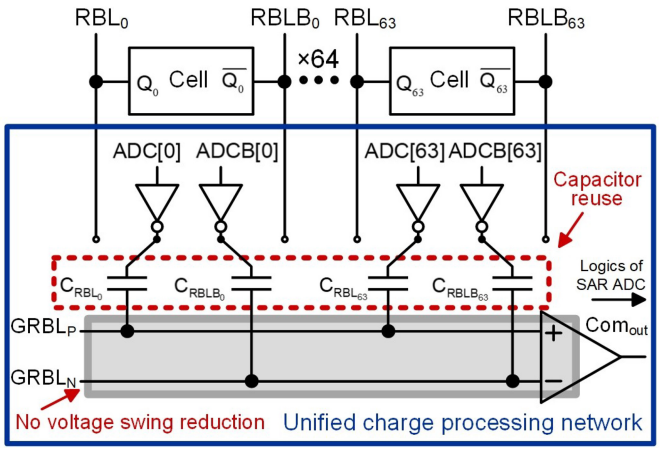

Fig. 6. Proposed UCPN.

and 4 DAC sign-bit switches. The multiplication operation carried out by the LCU involves three steps: 1) DAC precharge; 2) multiplication; and 3) data transfer, as shown in Fig. 4(c).

1) DAC Precharge: Before the DAC precharge, the DAC sign-bit switches would be first set according to the sign bit of the activations, and the bottom plates of capacitors C_{RBL} and C_{RBLB} would be connected to RBLs and RBLBs accordingly. During the DAC precharge, the signals of EN_{RBL} and EN_{DAC} are activated to precharge the C_{RBL} and C_{RBLB}. The two global read bit lines, GRBL_P and GRBL_N, are initially discharged to the ground to ensure that the voltage difference between the two capacitors is equal to the DAC precharged voltage.

2) Multiplication: During the multiplication, the precharge path would be first disconnected, and the C_{RBL} and C_{RBLB} would maintain the DAC precharged voltage in the previous step. The signal RWL is then activated to perform the multiplication. According to the data stored in SRAM bitcells, the voltage on either RBL or RBLB would be discharged to zero. After the discharge process finishes, the RWL would be turned off to complete the multiplication operation.

3) Data Transfer: During the data transfer, the signal $\mathrm{SEL}_{D/A}$ would switch the bottom plates of the capacitors C_{RBL} and C_{RBLB} to the ADC switch nodes in the proposed UCPN. The transferred data (charge) would be accumulated on GRBL_P and GRBL_N, and then further processed and converted into digital outputs in the UCPN. Fig. 4(b) shows the timing diagram and operations of the LCU and UCPN. All of the C_{RBLs} and C_{RBLBs} have the same capacitor value, which forms a unary capacitor array. During the data conversion, this capacitor array is reused in the proposed UCPN. Different numbers of the capacitors are switched in the power of two to realize the binary-weighted capacitor switching in each iteration of SAR ADC conversion.

D. Unified Charge-Processing Network

The SAR ADC architecture could potentially realize a high-throughput efficient data conversion for CIM SRAM operation. However, when used together with the high-performance charge redistribution MAC architecture, it suffers from several design issues compared to other MAC approaches, such as the current summing method [3]. Fig. 5 illustrates these design issues by using the CIM SRAM design in [8] as an example, which employed a charge redistribution MAC architecture to realize the product accumulation operation and a SAR ADC for final output digitization. Since the SAR ADC requires additional switching capacitors and bootstrap circuits, the corresponding area overhead would be significantly increased by 15.4%. Moreover, because the bootstrap circuit and switching capacitor array require an extra charge sharing process, the output swing of the final accumulation signal would be severely reduced by 1.6×, as illustrated by the equivalent circuit model in Fig. 5. As a result, this approach suffers from significant voltage swing reduction and capacitor overhead. Furthermore, to maintain the ADC performance with the reduced input swing, the size and bias current of the ADC comparator must be substantially increased. This causes excessively larger energy consumption of the ADC comparator and seriously deteriorates the overall energy efficiency of CIM SRAM.

To circumvent these design issues, we proposed a single UCPN to replace the charge redistribution block and the ADC in the traditional CIM SRAMs, as shown in Fig. 6. The proposed UCPN simultaneously provides both signal processing and data conversion functions with maximum resource utilization, significantly boosting overall energy and area efficiency. Instead of using separate capacitor arrays for data processing and conversion functions, one single capacitor array on RBL and RBLB is employed in the proposed UCPN to store the MAC results during data computation. At the same time, it also serves as the ADC switching capacitors for final data digitalization. This effectively enhances the area efficiency by 1.15×. Moreover, the proposed UCPN minimizes the signal propagation steps without the need for extra bootstrap circuits. This eliminates the charge sharing issue shown in Fig. 5 and maintains the maximum available signal swing. By using the bottom-plate sampling technique in the UCPN, the control switches required in the traditional CIM SRAMs and the associated charge injection and coupling issues can be avoided, enhancing the overall signal integrity. Altogether, these techniques improve the energy efficiency by 1.38×, when compared with the conventional CIM SRAM design using a charge redistribution MAC and a SAR ADC without UCPN such as [8].

III. Implementation and Measurement Results

A 16-kb SRAM test chip was implemented in a 28-nm CMOS technology to verify the proposed CIM SRAM design. The MOM capacitor technology was used to implement the capacitors in this test chip. Fig. 7 shows the die photograph of the test chip.

A. ML Accuracy Results

To evaluate the ML classification accuracy of the proposed CIM SRAM, we implemented a ResNet20 neural network on the proposed

Authorized licensed use limited to: National Taiwan University. Downloaded on March 09,2025 at 06:52:29 UTC from IEEE Xplore. Restrictions apply.

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

	JSSC'2020 [3]	ISSCC'2020 [5]	ISSCC'2020 [6]	JSSC'2020 [7]	**This Work**
Technology	55 nm	28 nm	28 nm	55 nm	**28 nm**
Supply voltage (V)	1–0.8	0.85–1	0.7–0.9	0.9	**0.9**
Input precision	2 / 4	4 / 8	4 / 8	2 / 8	**7**
Weight precision	5	4 / 8	8	2 / 8	**1 / 2**
Output precision	5 / 7	12 / 20	16 / 20	7 / 19	**5–7**
Accuracy (CIFAR-10)	90.2%–90.42%	91.5%–91.94%	91.9%–92.02%	91.2%–91.93%	**87.50%–88.87%**
Throughput (GOPS)	43.2–21.2	N/A	N/A	82.29–5.14	**186.18–45.51**
Energy efficiency (TOPS/W)	37.5–18.37	30.4–7	33.52–11.54	10.1–0.6	**41.87–12.37**
Area efficiency (GOPS/mm^2)	1136.7–557.8	N/A	N/A	13.8–0.9	**3288.4–803.83**

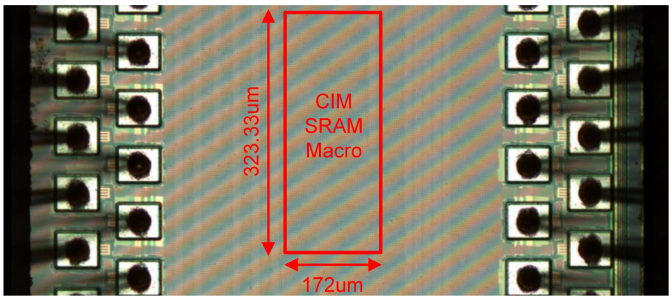

Fig. 7. Die photograph.

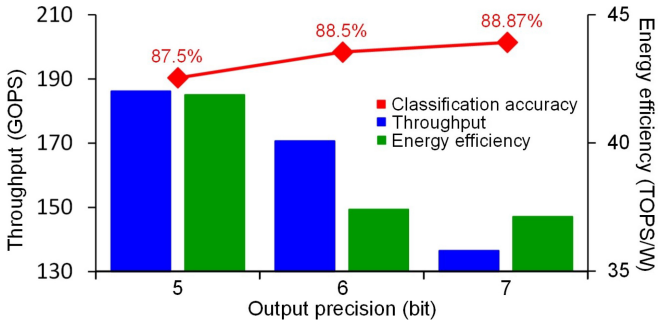

Fig. 8. Throughput and energy efficiency versus output precision.

design to perform the CIFAR-10 task. Compared to the baseline accuracy of 89.81% with pure software computation, the proposed CIM SRAM demonstrates only a slight accuracy degradation of 0.94% with 7-bit output precision. Fig. 8 shows the measured throughput and energy efficiency results with different output precisions, respectively, together with the corresponding classification accuracies. By lowering the output precision from 7 to 5 bits, the energy efficiency and the throughput of CIM SRAM can be enhanced accordingly, while the decline in classification accuracy is less than 1.37%.

B. Performance Comparison

Table I summarizes the performances of the proposed CIM SRAM design and compares them with the previous high-performance CIM SRAMs that have reported classification accuracy results for the CIFAR-10 task. The proposed CIM SRAM achieves a peak throughput of 186.18 GOPS, with energy and area efficiency of 41.87 TOPS/W and 3288.4 GOPS/mm^2, all of which significantly outperform the previous designs. Compared with the most efficient CIM SRAM [3], the proposed design realizes a 2.89× higher area efficiency while consuming 1.12× lower energy. In addition, it achieves a 2.26× higher throughput performance than the previous high-performance CIM SRAM design [7]. This clearly demonstrates the advantages of throughput, area, and energy efficiency of the proposed CIM SRAM, making it a promising solution to high-performance AI applications.

ACKNOWLEDGMENT

The authors would like to thank TSMC and the Taiwan Semiconductor Research Institute, Taiwan, for providing chip fabrication and technical support.

REFERENCES

[1] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.
[2] M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 642–655, Feb. 2018.
[3] X. Si *et al.*, "A twin-8T SRAM computation-in-memory unit-macro for multibit CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, Jan. 2020.
[4] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.
[5] J.-W. Su *et al.*, "15.2 A 28 nm 64 Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 240–242.
[6] X. Si *et al.*, "15.5 a 28 nm 64 Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 246–248.
[7] Y.-C. Chiu *et al.*, "A 4-Kb 1-to-8-bit configurable 6T SRAM-based computation-in-memory unit-macro for CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 10, pp. 2790–2801, Oct. 2020.
[8] 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.
[9] 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.
[10] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
[11] J. Song *et al.*, "7.1 An 11.5TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile SoC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 130–132.
[12] Y. Yamada *et al.*, "7.2 an 20.5TOPS and 217.3GOPS/mm^2 multicore SoC with DNN accelerator and image signal processor complying with ISO26262 for automotive applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 132–134.
[13] C.-H. Lin *et al.*, "7.1 A 3.4-to-13.3TOPS/W 3.6TOPS dual-core deep-learning accelerator for versatile AI applications in 7 nm 5G smartphone SoC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 134–136.

Authorized licensed use limited to: National Taiwan University. Downloaded on March 09,2025 at 06:52:29 UTC from IEEE Xplore. Restrictions apply.