IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 January 2024, accepted 19 February 2024, date of publication 23 February 2024, date of current version 11 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369057

RESEARCH ARTICLE

A Dual-Wordline 6T SRAM Computing-In-Memory Macro Featuring Full Signed Multi-Bit Computation for Lightweight Networks

ZUPEI GU[1,2], SHUKAO DOU[1,2], HENG YOU[3], (Member, IEEE), YI ZHAN[1], (Member, IEEE), SHUSHAN QIAO[1,2], (Member, IEEE), AND YUMEI ZHOU[1,2]

[1]Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Nanjing Institute of Intelligent Technology, Nanjing, Jiangsu 211135, China

Corresponding author: Shushan Qiao (qiaoshushan@ime.ac.cn)

ABSTRACT In this paper, we present an analog-mixed-signal 6T SRAM computing-in-memory (CIM) macro. The macro uses dual-wordline 6T bitcells to reduce power consumption and write-disturb issues. The macro also proposes an analog computation logic circuit for high precision, energy efficient charge-domain computation. The bitcell structure combined with the analog computation logic circuit allows direct input of signed activations and weights to the chip for full signed computation. The proposed macro consists of four CIM blocks, each with four 32 × 8 compute blocks, a pulse generator, an analog computation logic circuit and a SAR-ADC. Fabricated in a 55 nm process, our CIM macro test chip achieves an energy efficiency of 7.3 TOPS/W. A comprehensive computing test that encompasses the entire range of inputs and weights has been conducted. The results show that the CIM macro test chip can achieve a precision of 79.51% in a 1-FE error range of 71.88%. The target application of the proposed CIM macro is lightweight neural networks, this is demonstrated by mapping a pre-trained network into the macro and achieving a recognition accuracy of 92.28% on the CIFAR-10 dataset. The design surpasses existing designs in comprehensive consideration of energy efficiency, technology and bit width.

INDEX TERMS Computing-in-memory, SRAM, dual-wordline, neural network, charge-domain, multiply-and-accumulate.

I. INTRODUCTION

Over the last few years, SRAM-based CIM designs have demonstrated their advantage in significantly improving energy efficiency through reducing data movement, showing potential application in deep learning (DL) based edge computing devices [1], [2], [3], [4], [5], [6], [7], [8]. Recent works show CIMs being implemented across a wide variety of technologies and covering more novel application scenarios [9]. This indicates a transformation in the research direction of CIM, from solely pursuing energy efficiency to pursuing a balance between performance and functionality. Recent works [10], [11], [12], strengthen the relationship between algorithm and CIM structure, thereby making CIM designs more domain specific to enhance performance for practical scenarios.

Analog-mixed-signal SRAM-based CIM designs have unique advantages compared to other CIM categories, such as high computation parallelism, significant shorter access time, industrial maturity and relatively low energy cost when performing signed bit and floating-point computing operations [13]. However, analog-mixed-signal CIMs face

The associate editor coordinating the review of this manuscript and approving it for publication was Teerachot Siriburanon.

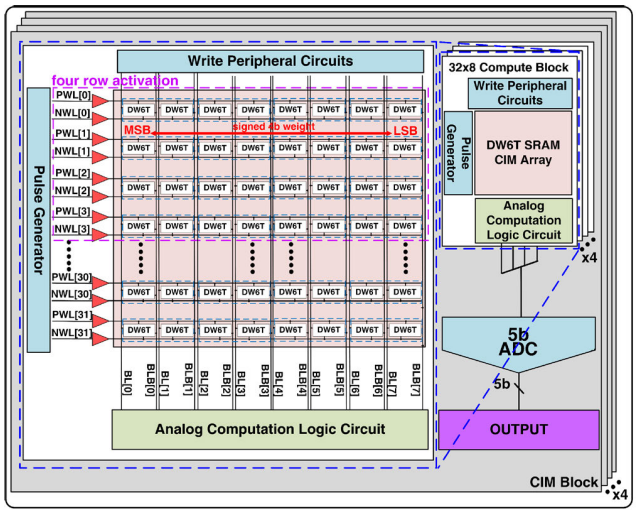


FIGURE 1. Structure of proposed DW6T SRAM based CIM.

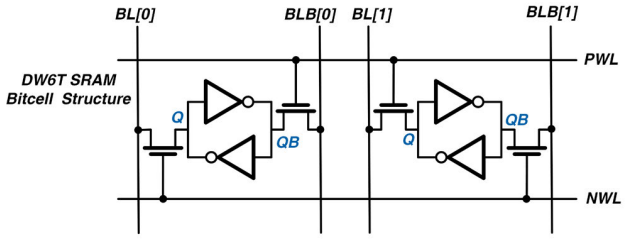
challenges including high ADC overhead and decreased precision caused by non-idealities. Apart from using design techniques to overcome these problems, a meaningful topic is to broaden the functions of analog-mixed-signal CIMs and find suitable fields of application to leverage the advantages of high computation speed without being affected by the disadvantages of certain accuracy lost. Analog-mixed-signal SRAM-based CIMs have become the preferred design for simple image and speech recognition tasks when moderate precision is required. The value of analog-mixed-signal integrates sensing, storage and computing functions on a single chip. In this situation, signals are sampled in the analog domain [14] and computation speed is required to keep up with the sensor sampling speed. Another field which shows their advantage is sparse lightweight networks [15], [16], a main category under neural networks. In contrast to digital SRAM-based CIMs, which trade more area and computing time for higher precision [17], [18], [19], [20], analog-mixed-signal SRAM-based CIMs perform small kernel convolution computations in a single cycle, significantly reducing computation time at the cost of minor recognition rate lost [21], [22], [23]. The lightweight network's high parallelism and energy efficiency computing requirements when using small convolution kernels, as well as its high tolerance for computational accuracy, perfectly match the characteristics of analog in-memory computing. However, application for analog-mixed-signal SRAM-based CIM designs on lightweight networks still need more research. The computing method and circuit structure that can perform the most efficient computation in the context of lightweight networks requires further discussion. Additionally, how to meet the needs of specific applications, such as full signed multi-bit computation, also deserves further research.

In this work, we introduce an analog-mixed-signal multi-bit CIM based on a 4kb Dual-Wordline 6T SRAM (DW6T SRAM) macro fabricated in 55 nm process. It supports full signed multi-bit computation for better recognition accuracy. The overall structure is shown in Fig. 1. The CIM macro is composed of four 1kb CIM blocks, with each CIM block containing four 32 × 8 compute blocks to perform accumulation of 16 multiplication results. The DW6T SRAM cells, which are modified from compact 6T foundry bitcells with no area increase, store the signed 4b weights for calculation. Utilization of the dual-wordline structure reduces power consumption and write-disturb issue [24] during the computation process. The inputs for calculation are given in the timing domain, produced by an on-chip simple and robust pulse generator. The 3b input is indicated by the length of the pulse and the symbol determines which wordline should be activated. An analog computation logic circuit accumulates the multiplication results on capacitors in the charge-domain, which are subsequently sampled by a SAR-ADC to generate signed 4b outputs. The CIM macro computes 128 multiply-and-accumulate (MAC) operations in parallel, achieving a total energy efficiency of 7.3 TOPS/W, using 16 signed 3b inputs and 64 signed 4b weights. The CIM macro is tested using computation operations under various scenarios for a direct result of the computation accuracy. Power consumption and energy efficiency results are also obtained in this process. Then a lightweight network is mapped onto the macro to perform recognition tests using the CIFAR-10 dataset. The analog-mixed-signal SRAM-based CIM macro, featuring a full signed computation method, exceeding energy efficiency and high precision, acts as a design worthy of reference in similar sparse lightweight network scenarios.

II. COMPUTING-IN-MEMORY ARCHITECTURE AND WORKING MECHANISM

A. COMPUTING-IN-MEMORY ARCHITECTURE

Fig. 2 presents the schematic and layout of the DW6T SRAM cells. As shown, 2 bitcells store one signed bit, while a row of 8 bitcells store a signed 4b weight. The SRAM bitcells are modified from compact foundry bitcells, with a modification that enables the transmission transistor to QB of the left bitcell to share the same wordline with the transmission transistor to Q of the right bitcell, while the other two transmission transistors share another wordline. Signed weight storage can be achieved by utilizing convolutional SRAM write-in method to store the values displayed in the first table in Fig. 2. During the computation phase, a charge accumulation capacitor is shared by the two bitlines of a single SRAM bitcell for the purpose of positive or negative result accumulation. The activation of positive wordline (PWL) or negative wordline (NWL) is determined by the input's symbol, whereas the charge current applied to the charge accumulation capacitor will be determined by the value of Q or QB. Due to this mechanism, two distinct storage methods can be used to store the value '0'. Depending on the activated wordline and stored weight, either the same amount of charge or no charge will be accumulated on both the positive and negative capacitors



Left Bitcell		Right Bitcell		Represent Value
Q	QB	Q	QB	
1	0	1	0	1
0	1	0	1	-1
0	1	1	0	0
1	0	0	1	0

	Left Bitcell		Right Bitcell	
	Q	QB	Q	QB
Charge Accumulation	Positive Cap		Negative Cap	

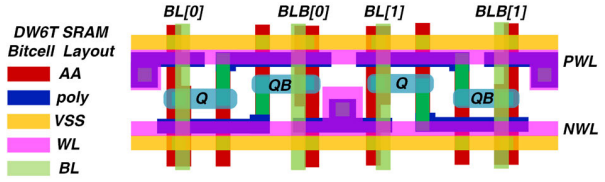


FIGURE 2. DW6T SRAM schematic and layout. Table shows method for storing signed bit in bitcells and the correspondence between storage node and capacitor during computation process.

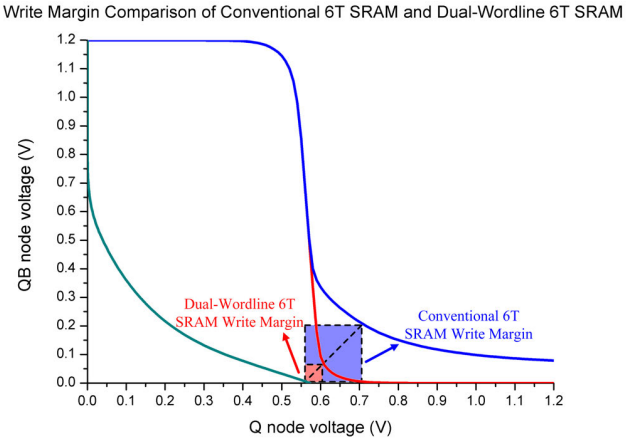


FIGURE 3. Write margin comparison of conventional 6T SRAM and dual-wordline 6T SRAM.

under both methods. Either way, they will neutralize when positive and negative results accumulate.

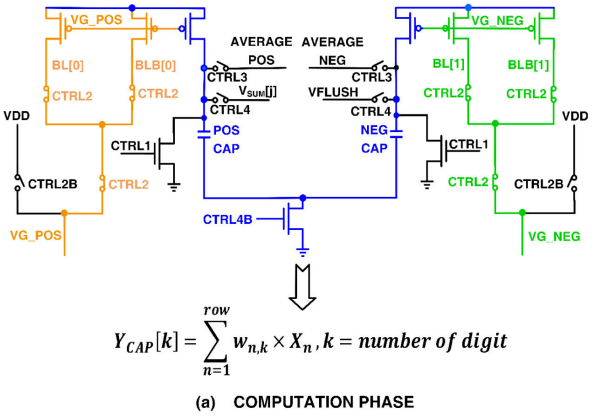
Utilizing the dual-wordline 6T SRAM structure reduces write-disturb issue during the computation process. We plot the ''butterfly curve'' of the conventional 6T SRAM and proposed dual-wordline 6T SRAM in Fig. 3 to compare their write margins. As shown, the write margin of the dual-wordline 6T SRAM is smaller, reducing the likelihood of cells getting overwritten during computation [25]. In dual-wordline 6T SRAM, the flip voltage for '1' to '0' when using 0.9 V and 1.2 V wordline voltage are 20 mV and 250 mV, respectively. In the analog computation logic circuit, the discharge current generated by the bitcells is converted into a charge current for the charge accumulation capacitors through current mirrors. As a result, when activating 1-4 rows simultaneously, the bitline voltages maintain at a stable level far above the flip voltage from '1' to '0', solving the write-disturb issue. Similarly, the state of the cell doesn't flip from '0' to '1'even when bitline voltage is 1.2 V.

B. FULL SIGNED MULTI-BIT COMPUTATION SCHEME

To accomplish the design objective of performing precise, energy efficient full-signed multi-bit computation, our proposed design is innovative in activation input method and result accumulation method. Dual wordlines are used to distinguish the signed digit, while the pulse width varies to signify the magnitude of activation. This activation input method harnesses the benefits of both pulse width modulation and multiple signal control to enhance accuracy. This makes our method stand out from existing single measure methodologies such as: wordline pulse count or pulse width modulation [4], [6], [18], analog input voltage on wordline [26], and multiple input wordlines or signals [1], [27]. Our approach for result accumulation also corresponds with our objective. Charge sharing between charge accumulation capacitors realizes multi-bit computation and ''capacitor stacking'' offers fast speed power saving accumulation between positive and negative results. This is different from [2] and [4] in that they achieve multi-bit computation through charge averaging, and from [28] and [29] in that they achieve multi-bit computation through charge redistribution. The entire signed multi-bit computation scheme is executed using the analog computation logic circuit, and its workflow is divided into four phases, controlled by signals CTRL1-CTRL4. The four phases are introduced in detail below.

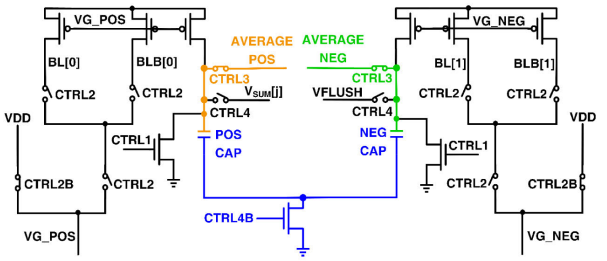
The first phase is the clear phase, and CTRL1 is activated. The charge accumulation capacitors, namely the positive capacitor (POS CAP) and the negative capacitor (NEG CAP) are cleared to prepare for the new cycle of computing.

The computation phase, which is controlled by CTRL2, is the second phase and is depicted in Fig. 4(a). As described in the previous section, activation inputs are given row-wise, in the form of a pulse. The length of the pulses corresponds to the data, while the selection of PWL and NWL varies depending on the symbol. The POS CAP and NEG CAP are charged based on the amount drained by the weights using a current mirror. The current mirror enhances the linearity of analog compute results by stabilizing the voltage on the bitlines throughout the computation phase. Modulating the transistor size of the current mirrors allows for the retention of small capacitor sizes, consequently reducing power consumption. In practical application, four rows are activated



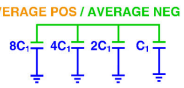
$$Y_{CAP}[k] = \sum_{n=1}^{row} w_{n,k} \times X_n, k = number\ of\ digit$$

(a) COMPUTATION PHASE

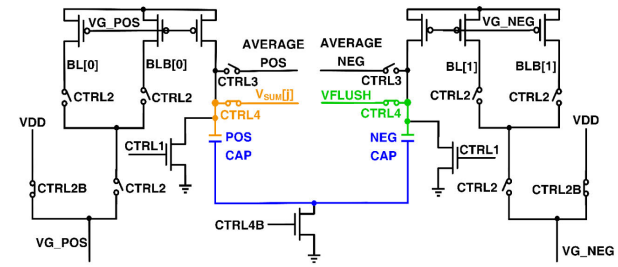


$$Y_{CAP\,TOTAL} = \frac{1}{2^{digit}-1} \sum_{i=1}^{digit} 2^{digit-i} Y_{CAP}[k]$$

$$= \frac{1}{15}(8Y_{CAP}[3] + 4Y_{CAP}[2] + 2Y_{CAP}[1] + Y_{CAP}[0])$$



(b) AVERAGING PHASE



$$V_{SUM}[j] = V_{POS} + (VFLUSH - V_{NEG})$$

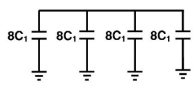
$$= VFLUSH + (V_{POS} - V_{NEG})$$

(c) POS-NEG ACCUMULATION

$$\frac{1}{N}\sum_{j=1}^{N} V_{SUM}[j]$$

$$= \frac{1}{4}(V_{SUM}[0] + V_{SUM}[1] + V_{SUM}[2]$$

$$+ V_{SUM}[4]) = \sum_{i=1}^{16} w_i \times X_i$$



(d) ADC SAMPLING

FIGURE 4. Signed multi-bit computation scheme, showing the workflow of the analog computation logic circuit. (a) Computation phase. (b) Averaging process. (c) Mechanism of how the positive and negative results accumulate. (d) ADC sampling process, the values on the positive capacitors are averaged and sampled.

simultaneously to accumulate 4 multiplication results in a computation unit. Charging of the capacitors finishes within

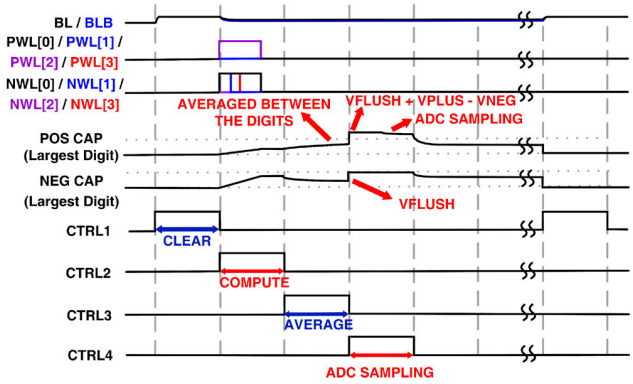


FIGURE 5. Waveform of the signed multi-bit computation scheme. Signals CTRL1-CTRL4 control the different phases of the scheme.

computation phase with CTRL2 having the same duration as the maximum input pulse.

The third phase is the averaging phase, where the positive and negative results of different digits are summed up separately, shown in Fig. 4(b). Signal CTRL3 controls the transistors for charge accumulation capacitor connection. The charge accumulation capacitors of 8 adjacent columns which form a 4b signed weight are geometric, with the biggest capacitance being 8 times the size of the smallest. During the averaging phase, bigger capacitance has a greater impact on the final averaged voltage, reflecting digit difference.

The fourth phase is when the positive and negative results accumulate, followed by ADC sampling, as illustrated in Fig. 4(c) and Fig. 4(d). This is controlled by CTRL4. During POS-NEG accumulation, the connection between ground and the bottom plates of POS CAP and NEG CAP is disconnected, whereas the connection between the two bottom plates remains. In this way, the POS CAP is "stacked" on top of NEG CAP. The voltage of NEG CAP top plate is then pulled up by VFLUSH, and since the electrons of the node connecting the two bottom plates are fixed, the voltage of this node increases by an equal amount as the voltage of NEG CAP top plate has been pulled up. Furthermore, due to the equal capacitance of POS CAP and NEG CAP, the voltage of POS CAP top plate (V_{SUM}[j]) increases the same amount as VFLUSH has pulled up minus the difference of the 2 capacitors, achieving the accumulation of positive and negative results, as shown in (1).

$$V_{SUM}[j] = VFLUSH + (V_{POS} - V_{NEG}) \quad (1)$$

This POS-NEG accumulate process is only performed by the analog computation logic circuit of the largest digit (with $8C_1$ POS CAP / NEG CAP) for faster accumulation and reduced error caused by switches. Four of these results (V_{SUM}[j]) are then averaged for accumulation, and sent to a SAR-ADC for ADC sampling which outputs 5b digital results. Due to this distinct calculation method, bigger ADC output corresponds to bigger positive calculation results, smaller ADC output corresponds to bigger negative calculation results and

TABLE 1. Impact of activating multiple wordlines on analog result precision.

Wordlines Activated	Stored Value	Bitline Voltage (mV)	Discharge Current (μA)
1	0	690	9.62
1	1	1200	0
2	0,0	638	18.6
2	0,1	693	9.58
2	1,1	1200	0
3	0,0,0	596	27.22
3	0,0,1	638	18.72
3	0,1,1	693	9.67
3	1,1,1	1200	0
4	0,0,0,0	565	35.26
4	0,0,0,1	598	27.09
4	0,0,1,1	638	18.61
4	0,1,1,1	694	9.67
4	1,1,1,1	1200	0

"2b'10000" represents the result of 0. The reference voltage of the ADC is adjustable according to the maximum and minimum accumulated calculation results. For instance, when the CIM macro is mapped as the last fully connected layer and calculation results are relatively low, reference voltage can be lowered for better distinction and enhance recognition rate. Our CIM comprises four CIM blocks of the circuits above, sharing the same inputs and different weights, a typical scenario in CNN layers of deep learning networks. The waveform timing diagram for the full signed multi-bit computation process is shown in Fig. 5 for overall examination. Changes of voltage on the charge accumulation capacitors during different phases are given as an example. During circuit design, the size and position of the transistors and capacitors are adjusted according to post-simulation results to reduce non-ideality issues caused by parasitics.

Activation of multiple wordline has an impact on bitcell storage stability and analog result precision. As analyzed in Section II-A, the dual-wordline bitcell structure solves write-disturb problems during computation. Table 1 lists the change in bitline voltage and discharge current corresponding to the difference in activated wordline number and stored value. Results are acquired through simulation using a column of 32 bitcells under tt corner. Through calculation, the average integral nonlinearity (INL) of the discharge current is 0.05 least significant bit (LSB). Considering the small value, the impact on precision is limited when only 4 rows are activated simultaneously, making it acceptable for applications in targeted lightweight network scenarios.

III. STATISTIC ANALYSIS

Fig. 6 illustrates the setup of our test environment, using our CIM macro for tests on accuracy, performance and recognition. The PC uses software tools Vivado and Vitis to program the FPGA board, and it analyzes the calculation results of the CIM macro. The FPGA is for data transfer between the PC and the CIM macro and mode control of the CIM macro. The CIM macro chip is packaged in the COB matter on the test PCB board. A power management circuit is also on the test PCB board to satisfy the different power supply needs of the CIM macro, which includes digital power, analog power, test circuit power, etc. The power analyzer is used to acquire power efficiency results throughout the performance test.

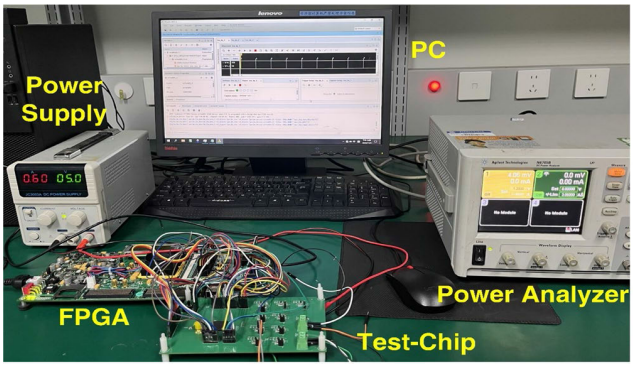
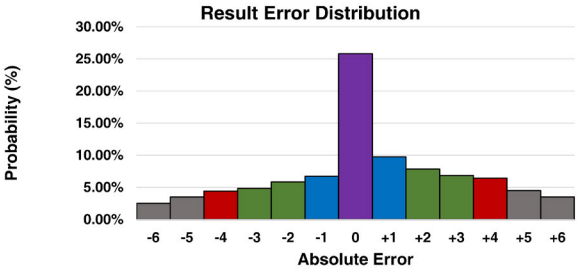


FIGURE 6. Test environment setup for the DW6T CIM macro.



Accuracy Analysis

Error Range	0	-1~+1	-3~+3	-4~+4
$1-FE$	100%	90.65%	78.13%	71.88%
Probability	25.79%	43.25%	68.65%	79.51%

$$Fiducial\ Error\ (FE) = \frac{\Delta_{Error\ Range}}{ADC\ Range} \times 100\% = \frac{\Delta_{Error\ Range}}{32} \times 100\%$$

FIGURE 7. Measured accuracy analysis and result error distribution.

A. COMPUTATION ACCURACY AND PERFORMANCE TEST

The computation accuracy test analyzes the results of 500 computation operations under various computing scenarios, covering the entire computation range. Fiducial Error (FE) is commonly used to express maximum errors coming from limited instrument resolution or quantization during analog-to-digital conversion [30].

$$Fiducial\ Error\ (FE) = \frac{\Delta_{Error\ Range}}{ADC\ Range} \times 100\% \quad (2)$$

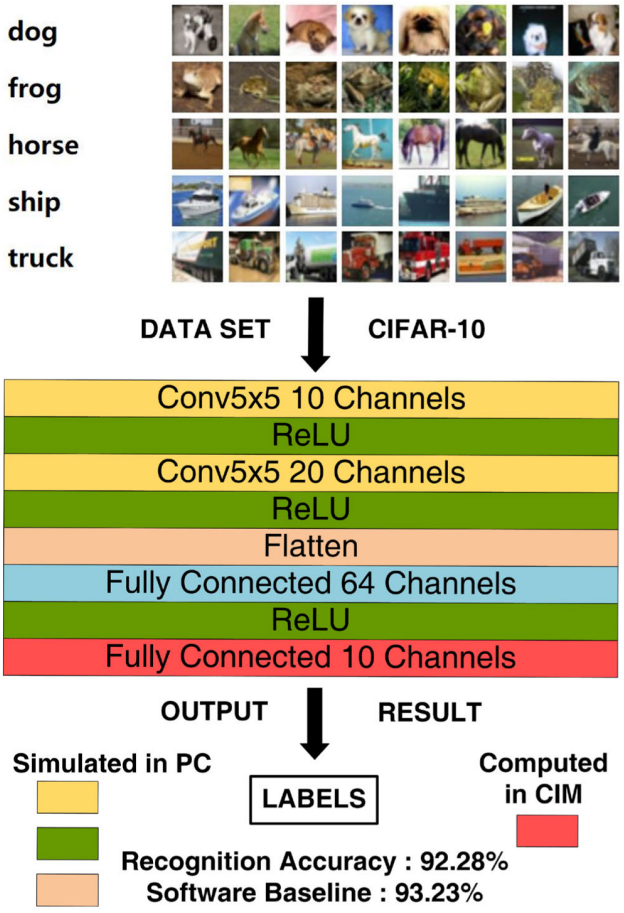


FIGURE 8. CIFAR-10 recognition network schematic.

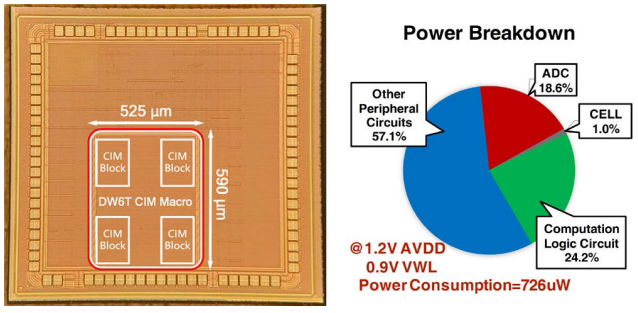


FIGURE 9. Chip micrograph and power breakdown.

By analyzing the value and probability of 1-FE, the influence of the computation result error distribution on the final recognition rate can be intuitively expressed. Using this method, accuracy analyzation reveals that more than 25% of the results have a 1-FE value of 100% (exactly correct), while 79.51% of the results have a 1-FE value of 71.88% (within a ±4 range of the theoretic result). Fig. 7 presents the result error distribution and FE calculation method. Absolute error is the difference between the output result by the CIM macro chip and the ideal result of the weights and inputs imported. Although the same ideal result can be produced by different sets of inputs and weights, when setting digital '1' as LSB, the

TABLE 2. Area statistic comparison.

	[3]	[18]	[12]	[34]	This Work
Type	Digital	Digital	Analog /Mixed	Analog /Mixed	Analog /Mixed
Technology (nm)	28	28	65	28	55
Bitcell Area (um^2)	/	0.7215	2.79	0.257	[a]1.05
[b]Function Area (mm^2)	[c]2.9118	[d]1.7000	1.6632	0.3234	0.0760
[e]Scaled Function Area (mm^2)	9.1501	5.3422	1.2529	1.0162	0.0760
Scaled Function Area Efficiency (GOPS/ mm^2)	8.22	37.44	74.50	122.89	70.12
Die Area(mm^2)	8.97	5.04	/	/	2.02
Scaled Die Area(mm^2)	28.19	15.84	/	/	2.02
Scaled System Area Efficiency (GOPS/ mm^2)	2.67	12.63	/	/	2.64

[a]The area of 2 bitcells are calculated; [b]Function Area = Bitcell Area + Additional CIM Circuits Area; [c]Area of 16 BSOCIMs; [d]Statistics of 5 T-PIM cores using 4bit inputs; [e]Scaled Area = Area × 1.8^$\log_2(55^2$ / technology^2) [31].

average INL of the digital output data can still be calculated with the provided data in the distribution graph, which results in 2.1 LSB, a satisfying result for digital-to-digital whole chip INL [31]. The data provided in the accuracy analysis table can also be used to simulate the performance of our CIM on various neural networks and calculation algorithms. This is done by adding a random offset during accumulation of each 16 multiplication results, with the probability and size of the offset determined by referring to the accuracy analysis table.

B. CIFAR-10 RECOGNITION TEST

The CIM macro is tested in an application scenario using the CIFAR-10 dataset. We use a pre-trained lightweight network as presented in Fig. 8 for implementation. It consists of 2 convolutional layers, 3 ReLU layers, 1 flattening layer and 2 fully connected layers with a total of 26.498K features. Limited CIM macro size still remains as a common challenge in the CIM field, so many CIM designs are tested by only mapping 1 or 2 layers [20], [32], [33]. Mapping entire networks into CIM hardware still requires continuous research for the CIM community. In this work, the last fully connected layer is mapped onto the macro, while the other layers operate on the PC and FPGA. The CIM macro calculates recognition results and outputs them directly to the PC, where they can be compared to the results computed by software. Through testing, the CIM macro achieves a recognition accuracy of 92.28% using the CIFAR-10 dataset compared to baseline accuracy of 93.23%.

C. AREA STATISTIC ANALYSIS

The chip micrograph of the CIM is shown in Fig. 9 left. The die area of the chip is 2.02 mm^2, with the CIM macro

TABLE 3. Comparison to prior CIM macros.

	JSSC'22 [1]	JSSC'21 [2]	JSSC'20 [6]		JSSC'22 [15]	JSSC'23 [18]	ISSCC'23 [19]	This Work
Technology (nm)	28	7	55		65	28	4	55
Cell Structure	6T SRAM	8T SRAM	6T SRAM		8T SRAM	8T SRAM	6T SRAM	DW6T SRAM
Array Size	64kb	4kb	4kb		16kb	219kb	54kb	4kb
Input Bits	4	4	7	8	5	1	4 (±3)	4 (±3)
Weight Bits	4	4	1	8	1	2	5 (±4)	5 (±4)
ADC Output Precision (analog) Full Output Precision (digital)	5 (analog)	4 (analog)	4 (analog)		5 (analog)	[c]24b (digital)	24 (digital)	5 (±4) (analog)
Energy Efficiency (TOPS/W) (sparsity)	29.2	351	5.7	0.6	15.8	161.08 (90%)	87.4 (87.5%)	[d]7.3 (70%)
[a]Energy Efficiency @55nm	7.57	5.69	5.7	0.6	22.07	41.74	0.46	7.3
[b]FoM	121.1	91.0	39.9	38.4	110.4	83.5	29.5	146.0

[a]Energy Efficiency scaled, assuming $Energy \propto (Technology)^2$ [4]; [b]FoM = Input Bits × Weight Bits × Energy Efficiency (scaled to 55nm); [c]24b has an 24b storage, for storage under different computation conditions; [d]Each Nb × Nb is considered as 2 operations.

occupying 0.30975 mm^2 of that space. The macro's function area is 0.076012 mm^2, which includes the bitcell area and the area of additional CIM circuits. These additional CIM circuits include ADC, capacitors, row circuitry, etc. Table 2 compares our CIM macro to those of works that employ similar technologies, providing a more detailed analysis of area statistics. According to comparison, our proposed CIM macro has a small bitcell area, this is because it is modified from foundry compact bitcells. Compared to other works, the area efficiency of our work is moderate, showing decent trade-off between area, function, energy efficiency and accuracy.

D. ENERGY EFFICIENCY TEST AND STATISTIC COMPARISON

The power breakdown of the CIM is shown in Fig. 9 right. In addition to the CIM macro, the chip contains test circuits and temporary storage SRAMs for testing convenience. Tested power consumption is 726μW at 1.2 V main analog power for CIM and 0.9 V wordline activation voltage. This is an average power of the CIM macro working under different computation scenarios with an average input sparsity of 70%. The power breakdown shows that the SAR-ADC and analog computation logic circuit consume about half of total power, which is a common distribution in charge-domain analog-mixed-signal SRAM based CIMs.

We compare this work with state-of-the-art SRAM CIM macros in Table 3. The CIM macro we present achieves 7.3 TOPS/W under 1.2 V main power and 200 MHz clock frequency, it reaches the highest figure of merit (FoM) value of 146 when technology and input/weight precision is added to consideration. Compared to [1], our CIM macro exceeds similar bitwise CIMs, even with more complicated full signed computation. Compared to [2], statistics show the advantages of structure improvement compared to advanced technology process. Compared to [17], [18], [19], and [20], the advantage of analog-mixed-signal SRAM-based CIM designs over digital CIM designs when performing signed multi-bit computing operations is revealed, even with a smaller input sparsity.

IV. CONCLUSION

This paper presents a dual-wordline 6T SRAM-based analog-mixed-signal computing-in-memory macro designed for energy efficient signed multi-bit computing in sparse lightweight networks. Thanks to the DW6T bitcell structure and exquisite computation scheme, the CIM macro outputs precise results with low power and short time cycle. A test-chip is fabricated using 55nm CMOS technology. It is tested using comprehensive computing operations achieving a precision of 79.51% in a 1-FE error range of 71.88%, and an energy efficiency of 7.3 TOPS/W at ±4/±3 weight/input precision. The CIM macro is tested in an application scenario by mapping a lightweight network into the macro and preforming CIFAR-10 recognition task. Compared with other state-of-the-art CIM macros, our proposed work reaches the highest FoM value of 146 when comprehensively considering energy efficiency, technology and input/weight bit width. The CIM macro is a design with high energy efficiency and accuracy, making it worthy of reference for future analog-mixed-signal SRAM-based CIMs aiming for sparse lightweight network applications.

REFERENCES

[1] J.-W. Su et al., "Two-way transpose multibit 6T SRAM computing-in-memory macro for inference-training AI edge chips," *IEEE J. Solid-State Circuits*, vol. 57, no. 2, pp. 609–624, Feb. 2022.

[2] M. E. Sinangil, B. Erbagci, R. Naous, K. Akarvardar, D. Sun, W.-S. Khwa, H.-J. Liao, Y. Wang, and J. Chang, "A 7-nm compute-in-memory SRAM macro supporting multi-bit input, weight and output and achieving 351 TOPS/W and 372.4 GOPS," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, Jan. 2021.

[3] R. Guo, Z. Yue, X. Si, H. Li, T. Hu, L. Tang, Y. Wang, H. Sun, L. Liu, M.-F. Chang, Q. Li, S. Wei, and S. Yin, "TT@CIM: A tensor-train in-memory-computing processor using bit-level-sparsity optimization and variable precision quantization," *IEEE J. Solid-State Circuits*, vol. 58, no. 3, pp. 852–866, Mar. 2023.

[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] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang, and M.-F. Chang, "Challenges and trends of SRAM-based computing-in-memory for AI edge devices," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 5, pp. 1773–1786, May 2021.

[6] Y.-C. Chiu, Z. Zhang, J.-J. Chen, X. Si, R. Liu, Y.-N. Tu, J.-W. Su, W.-H. Huang, J.-H. Wang, W.-C. Wei, J.-M. Hung, S.-S. Sheu, S.-H. Li, C.-I. Wu, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "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.

[7] F. Tu, Y. Wang, Z. Wu, L. Liang, Y. Ding, B. Kim, L. Liu, S. Wei, Y. Xie, and S. Yin, "ReDCIM: Reconfigurable digital computing-in-memory processor with unified FP/INT pipeline for cloud AI acceleration," *IEEE J. Solid-State Circuits*, vol. 58, no. 1, pp. 243–255, Jan. 2023.

[8] S. H. H. Nemati, N. Eslami, and M. H. Moaiyeri, "A hybrid SRAM/RRAM in-memory computing architecture based on a reconfigurable SRAM sense amplifier," *IEEE Access*, vol. 11, pp. 72159–72171, 2023.

[9] S. Mirabbasi, L. C. Fujino, and K. C. Smith, "Through the looking glass—The 2023 edition: Trends in solid-state circuits from ISSCC," *IEEE Solid StateCircuits Mag.*, vol. 15, no. 1, pp. 45–62, Jan. 2023.

[10] P.-C. Wu, J.-W. Su, L.-Y. Hong, J.-S. Ren, C.-H. Chien, H.-Y. Chen, C.-E. Ke, H.-M. Hsiao, S.-H. Li, S.-S. Sheu, W.-C. Lo, S.-C. Chang, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 22nm 832Kb hybrid-domain floating-point SRAM in-memory-compute macro with 16.2–70.2TFLOPS/W for high-accuracy AI-edge devices," in *Proc. IEEE Int. Solid- State Circuits Conf. (ISSCC)*, San Fransisco, CA, USA, Feb. 2023, pp. 126–128.

[11] S. Jeong, J. Park, and D. Jeon, "A 28nm 1.644TFLOPS/W floating-point computation SRAM macro with variable precision for deep neural network inference and training," in *Proc. IEEE 48th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2022, pp. 145–148.

[12] X. Qiao, J. Song, X. Tang, H. Luo, N. Pan, X. Cui, R. Wang, and Y. Wang, "A 65 nm 73 kb SRAM-based computing-in-memory macro with dynamic-sparsity controlling," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 6, pp. 2977–2981, Jun. 2022.

[13] Z. Sun, S. Kvatinsky, X. Si, A. Mehonic, Y. Cai, and R. Huang, "A full spectrum of computing-in-memory technologies," *Nature Electron.*, vol. 6, no. 11, pp. 823–835, Nov. 2023.

[14] T.-H. Hsu, Y.-R. Chen, R.-S. Liu, C.-C. Lo, K.-T. Tang, M.-F. Chang, and C.-C. Hsieh, "A 0.5-V real-time computational CMOS image sensor with programmable kernel for feature extraction," *IEEE J. Solid-State Circuits*, vol. 56, no. 5, pp. 1588–1596, May 2021.

[15] C. Yu, T. Yoo, K. T. C. Chai, T. T. Kim, and B. Kim, "A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3466–3476, Nov. 2022.

[16] Z. Chen, Z. Yu, Q. Jin, Y. He, J. Wang, S. Lin, D. Li, Y. Wang, and K. Yang, "CAP-RAM: A charge-domain in-memory computing 6T-SRAM for accurate and precision-programmable CNN inference," *IEEE J. Solid-State Circuits*, vol. 56, no. 6, pp. 1924–1935, Jun. 2021.

[17] H. Fujiwara, H. Mori, W.-C. Zhao, M.-C. Chuang, R. Naous, C.-K. Chuang, T. Hashizume, D. Sun, C.-F. Lee, K. Akarvardar, S. Adham, T.-L. Chou, M. E. Sinangil, Y. Wang, Y.-D. Chih, Y.-H. Chen, H.-J. Liao, and T. J. Chang, "A 5-nm 254-TOPS/W 221-TOPS/mm^2 fully-digital computing-in-memory macro supporting wide-range dynamic-voltage-frequency scaling and simultaneous MAC and write operations," in *Proc. IEEE Int. Solid- State Circuits Conf. (ISSCC)*, vol. 65, San Francisco, CA, USA, Feb. 2022, pp. 1–3.

[18] J. Heo, J. Kim, S. Lim, W. Han, and J.-Y. Kim, "T-PIM: An energy-efficient processing-in-memory accelerator for end-to-end on-device training," *IEEE J. Solid-State Circuits*, vol. 58, no. 3, pp. 600–613, Mar. 2023.

[19] H. Mori, W. Zhao, C.-E. Lee, C.-F. Lee, Y.-H. Hsu, C. Chuang, T. Hashizume, H.-C. Tung, Y.-Y. Liu, S.-R. Wu, K. Akarvardar, T. Chou, H. Fujiwara, Y. Wang, Y. Chih, Y.-H. Chen, H. Liao, and T. Chang, "A 4nm 6163-TOPS/W/b 4790—TOPS/mm^2/b SRAM based digital-computing-in-memory macro supporting bit-width flexibility and simultaneous MAC and weight update," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC), San Francisco, CA, USA*, vol. 2023, San Francisco, CA, USA, 6163, pp. 132–134.

[20] X. Zhang, Y. Lu, B. Wang, and T. T. Kim, "A digital bit-reconfigurable versatile compute-in-memory macro for machine learning acceleration," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 5, pp. 1744–1748, May 2023.

[21] A. Kneip, M. Lefebvre, J. Verecken, and D. Bol, "A 1-to-4b 16.8-POPS/W 473-TOPS/mm^2 6T-based in-memory computing SRAM in 22nm FD-SOI with multi-bit analog batch-normalization," in *Proc. IEEE 48th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2022, pp. 157–160.

[22] V. T. Nguyen, J.-S. Kim, and J.-W. Lee, "10T SRAM computing-in-memory macros for binary and multibit MAC operation of DNN edge processors," *IEEE Access*, vol. 9, pp. 71262–71276, 2021.

[23] H. Jeong, S. Kim, K. Park, J. Jung, and K. J. Lee, "A ternary neural network computing-in-memory processor with 16T1C bitcell architecture," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 5, pp. 1739–1743, May 2023.

[24] P. K. Bharti, S. Jain, K. R. Pillai, S. V. Sayyaparaju, G. S. Kalsi, J. Mekie, and S. Subramoney, "Compute-in-memory using 6T SRAM for a wide variety of workloads," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Austin, TX, USA, May 2022, pp. 2963–2967.

[25] C. D. C. Arandilla, A. B. Alvarez, and C. R. K. Roque, "Static noise margin of 6T SRAM cell in 90-nm CMOS," in *Proc. UkSim 13th Int. Conf. Comput. Model. Simul.*, Mar. 2011, pp. 534–539.

[26] 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.

[27] S. Cheon, K. Lee, and J. Park, "A 2941-TOPS/W charge-domain 10T SRAM compute-in-memory for ternary neural network," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 5, pp. 2085–2097, May 2023.

[28] 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.

[29] K. Xiao, X. Cui, X. Qiao, J. Song, H. Luo, X. Wang, and Y. Wang, "A 28nm 32Kb SRAM computing-in-memory macro with hierarchical capacity attenuator and input sparsity-optimized ADC for 4b MAC operation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 6, pp. 1816–1820, Jun. 2023.

[30] J. Hannig, H. Iyer, and P. Patterson, "Fiducial generalized confidence intervals," *J. Amer. Stat. Assoc.*, vol. 101, no. 473, pp. 254–269, Mar. 2006.

[31] 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.

[32] Y. K. Lee, D. H. Ko, S. Cho, M. Yeo, M. Kang, and S.-O. Jung, "Split WL 6T SRAM based bit serial computing in memory macro with high signal margin and high throughput," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, early access, doi: 10.1109/TCSII.2023.3337754.

[33] J. Lou, F. Freye, C. Lanius, and T. Gemmeke, "An energy efficient all-digital time-domain compute-in-memory macro optimized for binary neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 71, no. 1, pp. 287–298, Jan. 2024.

[34] X. Si et al., "A local computing cell and 6T SRAM-based computing-in-memory macro with 8-b MAC operation for edge AI chips," *IEEE J. Solid-State Circuits*, vol. 56, no. 9, pp. 2817–2831, Sep. 2021.

ZUPEI GU received the B.E. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing. His research interests include computing-in-memory and intelligent edge computing.

SHUKAO DOU received the B.S. degree in electronic science and technology from Hunan University, Hunan, China, in 2018. He is currently pursuing the Ph.D. degree with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China. His research interests include analog computing-in-memory and analog-to-digital converters.

HENG YOU (Member, IEEE) received the B.S. degree in electronic science and technology from the University of Science and Technology of China, Hefei, China, in 2016, and the Ph.D. degree in circuits and systems from the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China, in 2021. In 2021, he joined Nanjing Institute of Intelligent Technology, China. His current research interests include compute-in-memory, low-power SRAM, and sub/near-threshold digital circuit design.

YI ZHAN (Member, IEEE) received the Ph.D. degree from Keio University, Yokohama, Japan, in 2012. He is currently a Professor with the Institute of Microelectronics (IME) of the Chinese Academy of Sciences, Beijing, China. He is also participating and in charge of a research supported by the National Key Research and Development Program of China. His research interests include artificial intelligence, the AIoT edge computing, computing-in-memory, low-power acoustic, and image microsystems. He has received the Japanese Government (Monbukagakusho) MEXT Scholarship from Keio University.

SHUSHAN QIAO (Member, IEEE) received the Ph.D. degree from the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China, in 2008. From 2008 to 2010, he was a Research Associate with the Institute of Microelectronics of the Chinese Academy of Sciences, where he was an Associate Professor, from 2011 to 2017, and has been a Professor, since 2018. He has also been a Professor with the University of Chinese Academy of Sciences, Beijing, since 2018. His current research interests include artificial intelligence, ultra-low-power processors, intelligent microsystems, and communication chips.

YUMEI ZHOU received the B.S. degree from the Department of Radio Electronics, Tsinghua University, Beijing, China, in 1985. Since 1997, she has been a Researcher with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing. Her research interests include integrated circuit design technology, reliability technology, low-power circuit design, and high-speed interface circuits.