An SRAM-Based Reconfigurable Cognitive Computation Matrix for Sensor Edge Applications

Sheng-Yu Peng, *Senior Member, IEEE,* I-Chun Liu, Yi-Heng Wu, Ting-Ju Lin, Chun-Jui Chen, Xiu-Zhu Li, Yong-Qi Cheng, Pin-Han Lin, Kuo-Hsuan Hung, Yu Tsao *Senior Member, IEEE,*

***Abstract*—A reconfigurable cognitive computation matrix (RCCM) in static random access memory (SRAM) suitable for sensor edge applications is proposed in this paper. The proposed RCCM can take multiple analog currents or digital integers as the input vector and performs vector-matrix multiplication with a weight integer matrix. The RCCM can carry out 1-quadrant, 2-quadrant, or 4-quadrant multiplications in the analog domain. Therefore, the digital integers for the inputs or weights stored in the SRAM can be either signed or unsigned, providing extensive usage flexibilities. Furthermore, three commonly used activation functions, the rectified linear unit (ReLU), radial basis function (RBF), and logistic function, are available, converting multiply-accumulation outputs to single-ended currents as the computation results. The resultant output currents can be adopted as the input currents of other RCCMs to facilitate multiple-layer network implementation. A concept-proving prototype chip, including a 16×16 RCCM with 4-bit input and weight resolutions, is designed and fabricated in a $0.18\,\mu\mathrm{m}$ CMOS process. The computation accuracy that is deteriorated by process variation can be significantly improved by adopting 48 mismatch parameters after calibration. A handwritten digit recognition database, MNIST, is employed to evaluate the chip performance, achieving average efficiency of $3.355\mathrm{TOPS/W}$.**

***Index Terms*—computing in memories, artificial intelligent circuits, cognitive computation, edge computing, low-power circuit, computing-in-SRAM**

I. Introduction

CONVENTIONALLY, analog front-end (AFE) circuits, analog-to-digital converters (ADCs), and digital signal processors (DSPs) consistently consume a reasonable amount of power to ensure no drop-out data, even when the information of interest is absent. The system can be more power-efficient if a cognitive computation unit (CCU) can be employed at the sensor edge to dynamically adjust the power consumption based on the significance of input signals, as shown in Fig. 1. When the inference results from the CCU with low latency indicate that the incoming signals are not informative, the AFE circuits can be switched to a low-power mode, and the ADCs and DSPs can be kept in hibernation to save unnecessary processing power. Once the inference results reveal or predict informative sensor signals in presence, the AFE circuits can be switched to a high-performance mode, and the ADCs and DSPs can be woken up to extract the information of interest [1]–[5]. If only the coarse information is required, the ADCs and the DSPs can remain in hibernation or be saved for further power reduction.

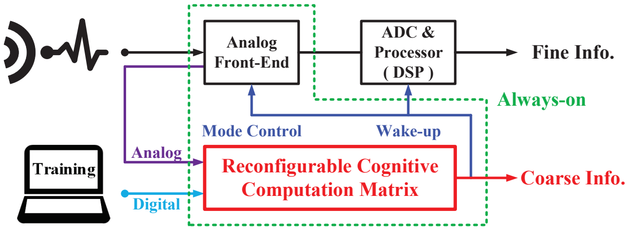


Fig. 1. A reconfigurable cognitive computation matrix facilitates dynamic power adjustment in an intelligent sensing system at the sensor edge.

Various digital CCUs for internet-of-things (IoT) or sensor applications [5]–[9] have been designed and developed near the sensor edge to reduce the power consumption for data movement. ADCs are required to be employed antecedently for sensor signal quantization. In many IoT applications with random sparse events, event-driven CCUs and level-crossing ADCs [3], [4], [10], [11] with non-uniform sampling rates are preferred to save conversion and processing power when the input signals are not of interest. Since most sensor signals are analog, many CCUs that perform cognitive computation in the analog domain have recently been proposed, avoiding the usage of ADCs for further power saving [1], [12]–[15]. However, these CCUs are designed for specific sensor signals and applications, limiting their diversities in utilization.

Many general-purposed CCUs have been proposed to perform multiplication and accumulation (MAC) operations inside static random access memories (SRAMs), which are ubiquitously accessible in CMOS technologies [16]–[33]. Binary and ternary neural networks with single-bit weights were first developed for simple networks [16]–[27]. Recently, SRAM macros capable of supporting MAC operations with multi-bit inputs and weights have been burgeoning [28]–[33]. Two weight matrices are employed to implement positive and negative weight values in [30]. Furthermore, the two's complement format can be adopted for the weight representation to improve area efficiency and reduce computation latency in [28], [29]. However, specialized processing and algorithmic adaptive readout units are required to quantize the computation output sequentially. In these approaches, the input data are encoded by either pulse intensity [16], pulse duration [21], or pulse numbers [30]. Then, they are applied on shared

S.-Y. Peng, I.-C. Liu, Y.-H. Wu, T.-J. Lin, C.-J. Chen, X.-Z. Li, Y.-Q. Cheng, and P.-H. Lin are or were with the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taiwan, e-mail: sypeng@mail.ntust.edu.tw. K.-H. Hung and Y. Tsao are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan (e-mail: yu.tsao@citi.sinica.edu.tw).

Manuscript received January 31, 2023.

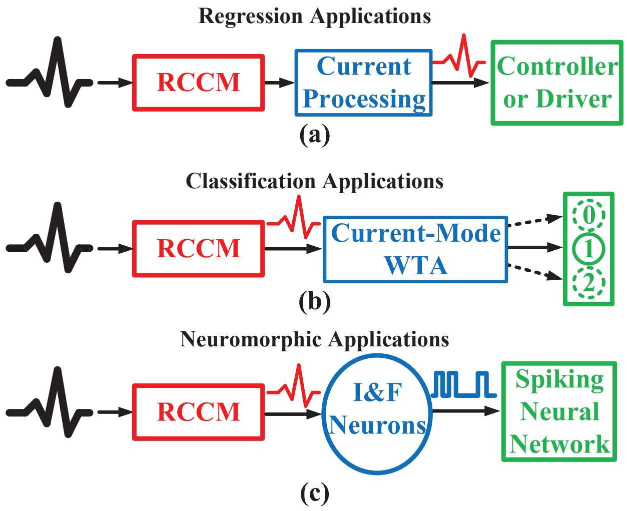


Fig. 2. The illustration of the potential employment of the proposed RCCM for (a) regression, (b) classification, and (c) neuromorphic applications.

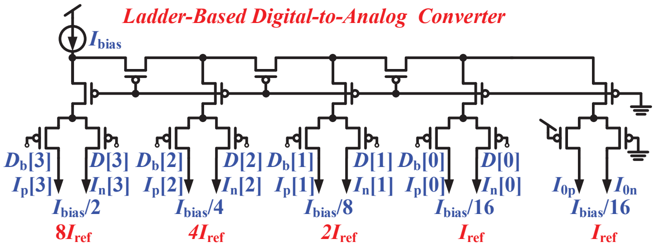


Fig. 3. A current mode ladder-based digital-to-analog converter (LBDAC).

word lines (WL) [16], [30] or shared global read bit lines (GRBL) [21], multiplied by digitally represented weight values stored in the SRAM array. The multiplication results can be accumulated on the bit lines (BL) in the current [16], [21], [27]–[30], voltage [25], [26], or charge [22], [24] mode and are subsequently quantized by ADCs. Although the MAC computations in the SRAMs mentioned above are in the analog domain, ADCs that usually dominate the power consumption [24] are still required to quantize input and output analog signals. Besides, clock signals are requisite to preset the voltage or charge on the bit lines before the MAC operation in the current and charge modes or to limit the conduction time for power saving in the voltage mode. Other arduous issues in computing-in-SRAM involve writing disturbance, area overhead, narrow linear dynamic range, low precision due to process variation, and limited reconfigurability. Therefore, designing a CCU in SRAM with diverse data representations for sensor applications is challenging.

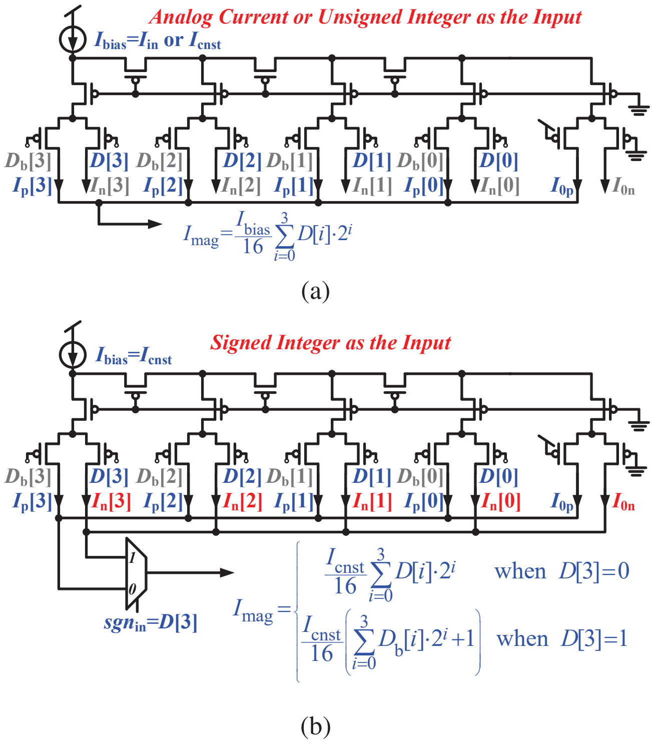
An SRAM-based reconfigurable cognitive computation matrix (RCCM) supporting various data representations is proposed in this paper. Ladder-based digital-to-analog converters (LBDACs) are meticulously amalgamated with 6T-SRAM cells in the proposed RCCM, avoiding writing disturbance and limited dynamic range. The proposed RCCM can take analog currents directly from either sensor or other RCCM chips or digital integers stored in the SRAM as the input vector and carry out 1-quadrant, 2-quadrant, or 4-quadrant vector-matrix-multiplications. Therefore, the digital integers for inputs and weights can be signed or unsigned, providing extensive usage flexibility. In addition, three commonly used activation functions, the rectified linear unit (ReLU), radial basis function (RBF), and logistic function, are available, converting multiply-accumulation outputs to single-ended currents to facilitate multiple-layer network implementation. The RCCM output currents can be followed by current processing circuitries, such as translinear circuits [34] or log-domain filters [35], to generate desired output currents for controllers or drivers directly, as shown in Fig. 2(a), composing an analog universal function generator in regression applications. Besides, a current-mode winner-take-all circuit [36] can take RCCM output currents, as shown in Fig. 2(b), generating one-hot codes for classification applications. Furthermore, integrated-and-fire (I&F) neuron circuits [37] can take the RCCM output currents and convert them into spike signals, as shown in Fig. 2(c), encoding extracted information in terms of pulse frequency or interval in neuromorphic applications. The generated spikes can be further processed in spiking neural networks. In such fashions, power-hungry circuits for data conversion can be avoided, achieving low-power intelligent sensing with low latency.

The innovation of this work involves the amalgamation of LBDACs and 6T-SRAM cells, which allows for various input and weight representations. The proposed multiplication scheme also efficiently supports signed or unsigned input and weight computation by switching connections with minimal overhead. Furthermore, the proposed algorithms enable the calibration of uploaded weight values in runtime or the offline training of the neural network to account for fabrication mismatches in the RCCM array. This paper is organized as follows. The data conversion, representation, and multiplications using LBDACs are introduced in the next section. Section III illustrates the architecture and operation principle of the proposed RCCM. Measurement results from a prototype chip are provided in section IV. The conclusion is finally drawn in section V.

II. Data Representation, Conversion, and Multiplication with LBDAC

A. Ladder-Based Digital-to-Analog Converter (LBDAC)

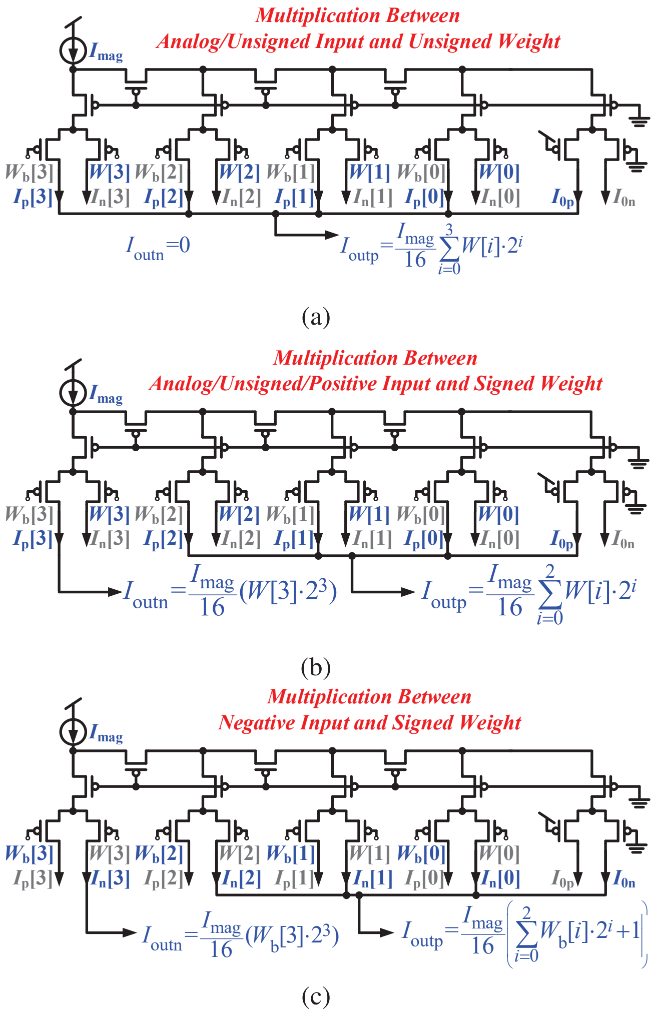
The integers for inputs and weights stored in the SRAM array are converted to currents by transistor-only ladder-based digital-to-analog converters (LBDACs) [38] in the proposed RCCM. The schematic of a 4-bit LBDAC is shown in Fig. 3, where all transistor dimensions are identical. Because of the ladder structure with matched impedances, the bias current, $I_{\rm bias}$, is split into binary weighted branch currents that flow to either the positive ($I_{\rm p}[\cdot]$) or the negative ($I_{\rm n}[\cdot]$) branches, depending on the corresponding digital bits, $D[\cdot]$ and $D_{\rm b}[\cdot]$. The branch current for the least significant bit (LSB) is the reference current, $I_{\rm ref} = {}^{I_{\rm bias}}/_{16}$, and is duplicated in $I_{\rm 0n}$.


Analog Current or Unsigned Integer as the Input
I_{bias}=I_{in} or I_{cnst}
$D_b[3]$ $D[3]$ $D_b[2]$ $D[2]$ $D_b[1]$ $D[1]$ $D_b[0]$ $D[0]$
$I_p[3]$ $I_n[3]$ $I_p[2]$ $I_n[2]$ $I_p[1]$ $I_n[1]$ $I_p[0]$ $I_n[0]$ I_{0p} I_{0n}
$I_{mag}=\frac{I_{bias}}{16}\sum_{i=0}^{3}D[i]\cdot 2^i$
Signed Integer as the Input
I_{bias}=I_{cnst}
1
0
sgn_{in}=$D[3]$
$I_{mag}=\begin{cases}\frac{I_{cnst}}{16}\sum_{i=0}^{3}D[i]\cdot 2^i & \text{when } D[3]=0\\ \frac{I_{cnst}}{16}\left(\sum_{i=0}^{3}D_b[i]\cdot 2^i+1\right) & \text{when } D[3]=1\end{cases}$

(a)

(b)

Fig. 4. The implementations of input data in different representations. (a) Input data can be represented as an analog current or an unsigned integer. (b) Input data is represented as a signed integer.


Multiplication Between
Analog/Unsigned Input and Unsigned Weight
I_{mag}
$W_b[3]$ $W[3]$ $W_b[2]$ $W[2]$ $W_b[1]$ $W[1]$ $W_b[0]$ $W[0]$
$I_p[3]$ $I_n[3]$ $I_p[2]$ $I_n[2]$ $I_p[1]$ $I_n[1]$ $I_p[0]$ $I_n[0]$ I_{0p} I_{0n}
I_{outn}=0
$I_{outp}=\frac{I_{mag}}{16}\sum_{i=0}^{3}W[i]\cdot 2^i$
Multiplication Between
Analog/Unsigned/Positive Input and Signed Weight
$I_{outn}=\frac{I_{mag}}{16}(W[3]\cdot 2^3)$
$I_{outp}=\frac{I_{mag}}{16}\sum_{i=0}^{2}W[i]\cdot 2^i$
Multiplication Between
Negative Input and Signed Weight
$I_{outn}=\frac{I_{mag}}{16}(W_b[3]\cdot 2^3)$
$I_{outp}=\frac{I_{mag}}{16}\left(\sum_{i=0}^{2}W_b[i]\cdot 2^i+1\right)$

(a)

(b)

(c)

Fig. 5. The implementations of the multiplication between input data and a weight value in different representations. (a) Multiplication between an analog input current or an unsigned integer and an unsigned weight integer. (b) Multiplication between an analog input current, an unsigned input integer, or a positively signed input integer and a signed weight integer. (c) Multiplication between a negative input integer and a signed weight integer.

B. Input Data Conversion

If the input data is an analog current from a sensor, it can be adopted as the bias current of an LBDAC. As shown in Fig. 4(a), the summation of all positive branch currents renders the input magnitude current, $I_{\rm mag}$, with the input sign bit of zero, $sgn_{\rm in} = 0$. The magnitude current can be expressed as

$$I_{\rm mag} = \frac{I_{\rm in}}{16}\sum_{i=0}^{3} D[i]\cdot 2^{i}, \tag{1}$$

where $D[\cdot]$ specifies a pre-weight value in 4-bit resolution.

Similarly, if the input data is an unsigned integer, a constant current, $I_{\rm cnst}$, is employed as the bias current, and the input magnitude can be expressed as

$$I_{\rm mag} = I_{\rm ref}\sum_{i=0}^{3} D[i]\cdot 2^{i}, \tag{2}$$

where $D[\cdot]$ represents the input data in 4-bit resolution with the input sign bit of zero, $sgn_{\rm in} = 0$.

If the input is a signed integer, the input sign bit is assigned to be the most significant bit (MSB), $sgn_{\rm in} = D[3]$. Depending on the sign bit, the magnitude current, $I_{\rm mag}$, can be either the summation of all positive or all negative branch currents, as shown in Fig. 4(b). The magnitude current can be expressed as

$$\begin{aligned} I_{\rm mag} =& I_{\rm ref}\cdot\left[\left(\sum_{i=0}^{3} D[i]\cdot 2^{i}\right)\cdot(1-D[3])\right.\\ &\left.+\left(\sum_{i=0}^{3} D_{\rm b}[i]\cdot 2^{i}+1\right)\cdot D[3]\right]. \end{aligned} \tag{3}$$

For example, a bit stream of [0101] represents a signed input integer of 5, of which magnitude can be implemented as $I_{\rm mag} = \Sigma I_{\rm p}[i] = 5I_{\rm ref}$ with $sgn_{\rm in} = D[3] = 0$. In another case of an input data of [1011] representing a signed integer of -5, the magnitude can be implemented as $I_{\rm mag} = \Sigma I_{\rm n}[i] + I_{\rm 0n} = 5I_{\rm ref}$ with $sgn_{\rm in} = D[3] = 1$.

C. Input-Weight Multiplication

Once the input magnitude current and sign bit are available, the input magnitude current, $I_{\rm mag}$, is employed as the bias current in the weight LBDAC to carry out multiplication with digital weight integers. The input sign bit, sgn_{in}, is adopted to choose output current polarities. The difference between a pair of output currents, $I_{\text{mult}} = I_{\text{outp}} - I_{\text{outn}}$, designates the multiplication result. Since the multiplication between a signed input and an unsigned weight is rarely used in artificial neural network applications, this mode is not supported in the proposed RCCM. Therefore, when the weight is unsigned, the input must be an analog current or an unsigned digital integer. In this case, as shown in Fig. 5(a), the positive output current is the summation of all positive branch currents, and the negative output current is zero. Consequently, the resultant output differential current can be expressed as

$$I_{\text{mult}} \equiv I_{\text{outp}} - I_{\text{outn}} = \frac{I_{\text{mag}}}{16} \sum_{i=0}^{3} W[i] \cdot 2^i \tag{4}$$

$$= \frac{I_{\text{ref}}}{16} \cdot \left(\sum_{i=0}^{3} D[i] \cdot 2^i \right) \cdot \left(\sum_{i=0}^{3} W[i] \cdot 2^i \right). \tag{5}$$

In the case of multiplication between an analog or unsigned input and a signed weight, the negative output current, I_{outn}, is assigned to be the positive branch current in the MSB. The positive output current, I_{outp}, is the summation of other positive branch currents, as shown in Fig. 5(b). The output differential current can be expressed as

$$I_{\text{mult}} = \frac{I_{\text{ref}}}{16} \cdot \left(\sum_{i=0}^{3} D[i] \cdot 2^i \right) \cdot \left(\sum_{i=0}^{2} W[i] \cdot 2^i - W[3] \cdot 2^3 \right). \tag{6}$$

For example, if the signed weight is denoted by a bit stream of [1110] as -2, the positive output current is given by $I_{\text{outp}} = \sum_{i=0}^{2} I_{\text{p}}[i]$, and the negative output current is assigned as $I_{\text{outn}} = I_{\text{p}}[3]$. Therefore, the resultant output current is $I_{\text{mult}} \equiv I_{\text{outp}} - I_{\text{outn}} = {}^{I_{\text{mag}}}/_{16} \cdot (6 - 8) = -2 \cdot {}^{I_{\text{mag}}}/_{16}$.

In the case of a signed input integer, the sign bit, sgn_{in}, determines whether the polarities of the output branches in the weight LBDAC should be flipped. If the input is positive, the multiplication implementation is the same as the previous case shown in Fig. 5(b). If the input is negative, the negative output current is assigned to be the negative branch current in the MSB. Besides, the positive output current is the summation of other negative branch currents, as shown in Fig. 5(c). The output differential current can be expressed as

$$I_{\text{mult}} = \frac{I_{\text{mag}}}{16} \left(\sum_{i=0}^{2} W_{\text{b}}[i] \cdot 2^i + 1 - W_{\text{b}}[3] \cdot 2^3 \right). \tag{7}$$

For example, if the input is -5 with $I_{\text{mag}} = 5I_{\text{ref}}$ and $sgn_{\text{in}} = 1$ and the signed weight integer is denoted by [1110] as -2, the positive output current is realized as $I_{\text{outp}} = \sum_{i=0}^{2} I_{\text{n}}[i] + I_{0\text{n}} = {}^{2}/_{16} I_{\text{mag}}$, and the negative output current is $I_{\text{outn}} = I_{\text{n}}[3] = 0$. Therefore, the resultant output current can be represented as $I_{\text{mult}} = I_{\text{outp}} - I_{\text{outn}} = {}^{I_{\text{mag}}}/_{16} \cdot (2 - 0) = 10 \cdot {}^{I_{\text{ref}}}/_{16}$.

D. Consolidation

A consolidated implementation, employing two cascaded LBDACs, for multiplication accommodating afore-mentioned input and weight signed/unsigned modes is illustrated in Fig. 6. The first LBDAC can take analog current as the input by replacing I_{bias} with I_{in} or take 4-bit digital input data, $D[3:0]$, in either signed or unsigned representation, providing the magnitude current, I_{mag}, for the subsequent LBDAC. Depending on the input sign mode, the input sign bit, sgn_{in}, is either 0 or $D[3]$. If the signed input integer is negative, $D[3] = 1$, and the I_{mag} is the summation of all negative branch currents; otherwise, the I_{mag} is the summation of all positive branch currents. If the weight is unsigned, the negative output current is zero, $I_{\text{outn}} = 0$, and the positive output current is the summation of all positive branch currents, $I_{\text{outp}} = I_{\text{out,MSB}} + I_{\text{out,main}}$. Finally, if the weight is signed, the output branch current in the MSB contributes the negative output current, $I_{\text{outn}} = I_{\text{out,MSB}}$, and the others are summed up to compose the positive output current, $I_{\text{outp}} = I_{\text{out,main}}$.

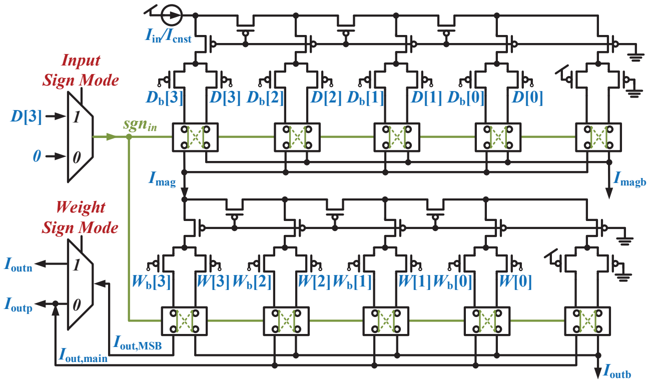


Fig. 6. A simplified consolidated block diagram of the proposed multiplier that employs two LBDACs supporting diverse data representations.

Considering an example where the input and the weight are -5 and -2, respectively, the signed input and the weight integers are denoted as [1011] and [1110], respectively. Since the input sign bit (sgn_{in}) is one, the magnitude can be implemented as $I_{\text{mag}} = \Sigma I_{\text{n}}[i] + I_{0\text{n}} = 5I_{\text{ref}}$. Because $sgn_{\text{in}} = D[3] = 1$, the polarities of the WPE output branches are flipped. According to Fig. 8(d), since $sgn_{\text{in}} = 1$ and $sgn_{\text{inb}} = 0$, $I_{\text{outn}} = I_{\text{out,MSB}} = 0$. Besides, $I_{\text{outp}} = \Sigma I_{\text{n}}[0] + I_{0\text{n}} = 2 \cdot {}^{I_{\text{mag}}}/_{16} = 2 \cdot {}^{5I_{\text{ref}}}/_{16}$. Therefore, the resulting output current is $I_{\text{mult}} = I_{\text{outp}} - I_{\text{outn}} = I_{\text{outp}} = 10 \cdot {}^{I_{\text{ref}}}/_{16}$.

III. Reconfigurable Cognitive Computation Matrix

A. Architecture

The block diagram of a version of the proposed reconfigurable cognitive computation matrix (RCCM) is shown in Fig. 7. The RCCM core consists of one column of the input processing element (IPE) array and multiple columns of the weight processing element (WPE) array. Input auxiliary control (IAC) blocks adjacent to the IPEs provide magnitude currents and input sign bits to the WPEs in the same row. Activation function (AF) blocks take currents from the WPE array and evaluate the chosen activation function in the analog domain. Three commonly used activation functions are available in the proposed RCCM, including the rectified

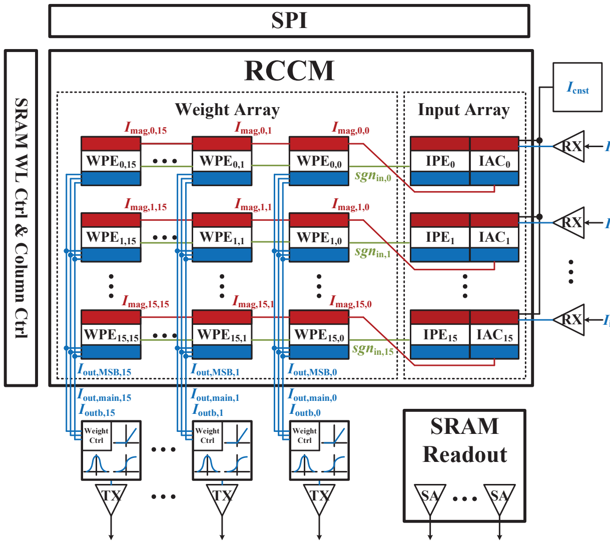


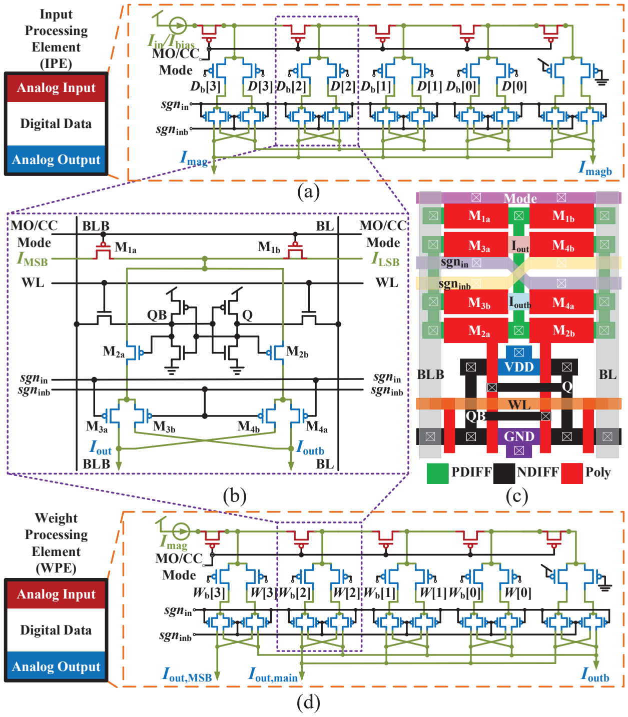
Fig. 7. The block diagram of a prototype version of the proposed reconfigurable cognitive computation matrix.

linear unit (ReLU), the radial basis function (RBF), and the logistic function. The resultant output currents from the proposed RCCM are sent to transmitter (TX) blocks, which can interface with receiver (RX) blocks in other RCCMs. The control circuitries for the SRAM array, including the serial peripheral interface (SPI), word-line selection, column headers, and sensing amplifiers, are located at the peripheries of the matrix.

Although the proposed RCCM architecture can be scaled up, the demonstrative matrix dimensions are 16 × 16 with 4-bit computing precision for digital inputs and weights, based on the trade-off between the chip fabrication cost and inference performance. More specifically, the accuracy from a fully-connected neural network of size [784 × 64 × 16 × 10], reported in [30], saturates at 4-bit precision when the MNIST dataset is employed for inference. Besides, the accuracy from a ResNet-20 neural network, reported in [29], only improves incrementally by less than 1% when the computing precision increases from 4 bits to 8 bits in the inference tasks using the CIFAR-10 and CIFAR-100 datasets. However, the power consumption and area overhead for 8-bit precision increase exponentially with lower computation efficiency when compared with 4-bit precision. Therefore, although the precision issue in existing analog-based computing-in-SRAM is not fundamentally resolved, we chose 4-bit computing precision to compromise hardware cost and system performance, as other research works [28]–[31], in implementing the presented prototype RCCM chip.

B. Input and Weight Processing Elements

Intriguingly, the schematic of an LBDAC can be sliced up into five almost identical units, as shown in Fig. 8, where dummy transistors are inserted in the MSB and LSB units. Each LBDAC slice can then be blended with a standard 6T-SRAM cell, composing a unit cell of the processing element (PE) shown in Fig. 8(b). Each PE is composed of components for analog input (M_{1a}), digital data (6T-SRAM cell), and analog output ($M_{2-4(a,b)}$), colored in red, white, and black, respectively, in PE symbols. The gate terminals of top pMOS transistors in an LBDAC shown in Fig. 3 are controlled by digital signals toggled between V_{DD} and Gnd. In the cognitive computation (CC) mode, transistor (M_{1a}) is turned on. As a result, the split input current, I_{bias} or I_{mag} in each unit, flows to either I_{out} or I_{outb}, according to the value stored in the 6T-SRAM cell and the broadcast input sign signals, sgn_{in} and sgn_{inb}. In the memory operation (MO) mode, the analog input transistor (M_{1a}) is turned off, terminating the input current.



(a)

(b)

(c)

(d)

Fig. 8. The schematics and symbols of the processing elements (PEs) in the proposed RCCM. (a) The symbol and schematic of the input processing element (IPE). (b) The schematic of a bit cell of the processing element. (c) The layout illustration of a single bit cell of the processing element. (d) The symbol and the schematic of the weight processing element (WPE).

The layout illustration of a single-bit PE cell is shown in Fig. 8(c). The layout of the 6T-SRAM bit cell is referenced on [39], [40] with the minimum dimensions. The additional seven pMOS transistors and a pMOS dummy device(M_{1b}) are then placed above the 6T-SRAM cell and sized accordingly to maximize the usage of the area with minimum widths. The additional eight pMOS transistors occupy 1.5 times the area of the 6T-SRAM cell. In contrast to digital-based computing-in-SRAM approaches, which offer high computation precision without being affected by process variation, the overhead area for bitwise multiplication and accumulating partial sums can exceed 3.5 times that of a 6T-SRAM cell [45]. Consequently, the proposed analog-based PE bit cell exhibits advantages in area efficiency.

Although the fifth-bit cell can be preset to zero without

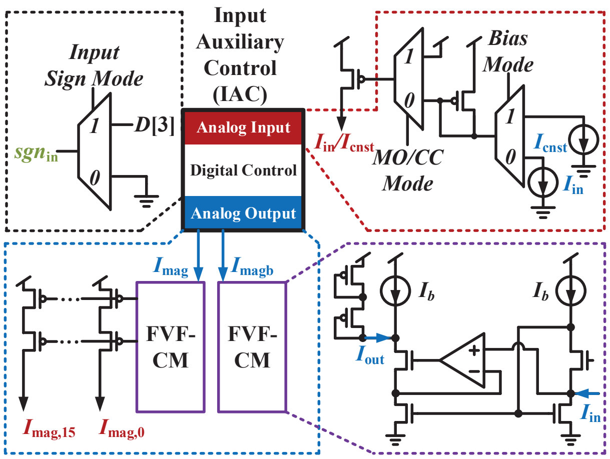


Fig. 9. The symbol and schematic of an input auxiliary control (IAC) block.

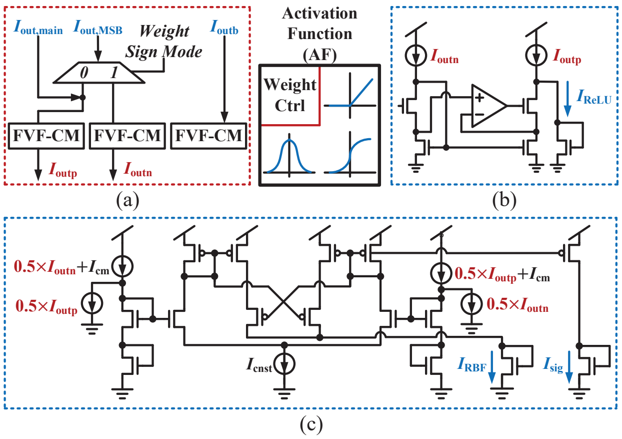


Fig. 10. The symbol of an activation function (AF) block. (a) The schematic of the weight control block. (b) The schematic of the circuit for rectified linear unit (ReLU) function evaluation. (c) The schematic of the circuit for radial basis function (RBF) and logistic function evaluation.

consuming the same 6T-SRAM layout area as other bit cells, it is still implemented by the same 6T-SRAM cell in the prototype chip implementation for layout matching consideration with slightly better usage flexibility. The fifth-bit cell can still be programmed to be 0, keeping the original functionality intact. When this cell is programmed to be 1, the maximum value of the unsigned integer to be implemented can increase from 15 to 16. The maximum value for the signed integer to be implemented can increase from 7 to 8. The range to be represented can be slightly increased.

The output current connections in input processing elements (IPEs) and those in weight processing elements (WPEs) are slightly different. In IPEs, each unit's output currents of I_{out} and I_{outb} are summed up, yielding the magnitude current, I_{mag}, and its complement, I_{magb}, as shown in Fig. 8(a). In WPEs, the I_{out} in the MSB are separated from those in rest bits, rendering output currents of $I_{\mathrm{out,MSB}}$, $I_{\mathrm{out,main}}$, and I_{outb}, as shown in Fig. 8(d).

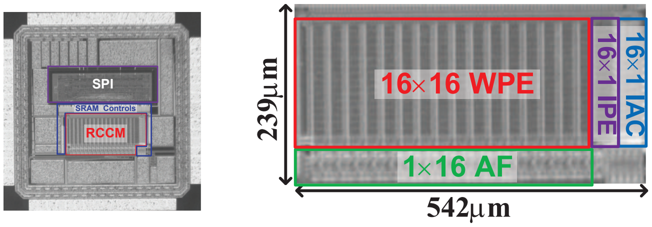
C. Input Auxiliary Control Block

The input auxiliary control (IAC) block shown in Fig. 9 delivers the magnitude current and the input sign bit from the adjacent IPE to all WPEs in the same row. The bias current, I_{bias}, for the input LBDAC can be either an externally injected input current, I_{in}, in the analog input mode or a constant current, I_{cnst}, in the digital input mode. The input sign bit, sgn_{in}, is multiplexed between the MSB and the ground according to the input sign mode. Loading the input LBDAC with low impedance for high linearity is crucial. Since the input impedance of a flipped-voltage-follower-based current mirror (FVF-CM) [41] is lower than that of general current mirrors, it is employed to load the input LBDAC. Finally, the output current is replicated by cascode *p*MOS current mirrors and broadcast to WPEs in the same row to avoid massive overhead in each column. Although the currents of I_{magb} are not utilized, they are loaded by FVF-CMs to match the impedance seen by I_{mag}.

D. Activation Function Block

The weight control circuits shown in Fig. 10(a) take the accumulated output currents, $I_{\mathrm{out,MSB}}$ and $I_{\mathrm{out,main}}$, from the WPE array and prepare the input currents, I_{outp} and I_{outn}, for the activation function circuits by adopting a multiplexer and two flipped-voltage-follower-based current mirrors (FVF-CMs). Although the currents of I_{outb} are not utilized, they are loaded by FVF-CMs to match the impedance seen by other output current branches. Two current mirrors shown in Fig. 10(b) implement the ReLU function. Besides, as shown in Fig. 10(c), an RBF and a sigmoid logistic activation function can be evaluated by a bump circuit [42] and a differential pair, respectively. The input currents of the proposed nonlinear circuit for RBF and logistic function evaluation are the differential currents of ${}^1\!/_2(I_{\mathrm{outp}}-I_{\mathrm{outn}})$ and ${}^1\!/_2(I_{\mathrm{outn}}-I_{\mathrm{outp}})$ added with an adjustable common-mode current, I_{cm}. Two diode-connected *n*MOS transistors in series convert input currents to voltages. The magnitude of I_{cm} can be a hyperparameter of the activation function that scales the input variables accordingly. The output currents of these nonlinear circuits for activation function are all single-ended. They can be fed into a current-mode winner-take-all circuit for classification or summed up directly for generative or regressive applications. These output currents can also be sent to transmitter (Tx) blocks and be adopted as the analog input currents for other RCCM chips to realize multi-layer networks.

Since the accumulation currents, I_{outp} and I_{outn}, are fed to the activation function circuits, the proposed architecture does not support partial sum directly. However, through the hardware-software co-design approach, where the inputs and weights can be specified to be unsigned, the ReLU activation function can be employed so that output currents across several RCCM chips can be accumulated, facilitating the vector-matrix multiplication operation with dimensions larger than those of a single RCCM.



(a) (b)

Fig. 11. (a) The micrograph of a prototype RCCM chip. (b) The enlarged micrograph of the proposed RCCM and its components.

E. Power and Efficiency Analysis

The proposed RCCM adopts ladder-based DACs for digital-to-analog conversion and multiplication inherently and performs the cognitive computation in the current domain, avoiding power-hungry circuitries for conventional data conversion, such as ADCs and clock generators, to achieve efficient analog-to-information conversion directly. The standby power is dominated by flipped-voltage followers and activation function circuitries located at the matrix's peripheries. As the input and output dimensions are large, the portion of the standby power can be negligible. In this case, the RCCM power consumption depends on analog input current levels. If the input data is digital, the power consumption of the RCCM depends on the input bias current level, which facilitates a direct trade-off between computation speed and system power consumption.

The power consumption of the developed RCCM chip depends on the weight values and magnitude level of the analog input currents or the chosen bias current (I_{cnst}), which can be directly traded off the computation latency. If the RCCM is in the unsigned mode and all the digital integers stored in the IPEs and WPEs are equal to half of the full scale, the current of one IAC block and one IPE can be estimated as

$$I_{\mathrm{IPE}} = 2I_{\mathrm{cnst}} + 2I_{\mathrm{FVF}}, \tag{8}$$

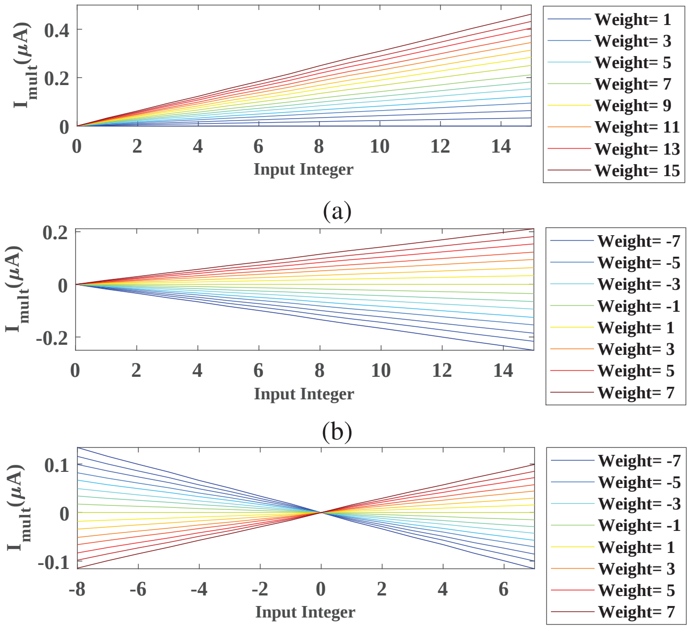
where $I_{\mathrm{FVF}} = I_{\mathrm{Amp}} + 2I_{\mathrm{b}}$ is the total current consumption of the FVF-CM, I_{Amp} is the current consumption of the amplifier, and I_{b} is the bias current of the FVF-CM. The current of the one WPE can also be estimated as

$$I_{\mathrm{WPE}} = \frac{1}{2} I_{\mathrm{cnst}}. \tag{9}$$

When the ReLU is chosen to be the activation function, the current consumption of one AF block can be estimated as

$$I_{\mathrm{AF}} = \frac{1}{4} I_{\mathrm{cnst}} \times 16 \times 2 + 3I_{\mathrm{FVF}} + I_{\mathrm{Amp}}. \tag{10}$$

In measurements, $I_{\mathrm{cnst}} = 240\,\mathrm{nA}$, $I_{\mathrm{Amp}} = 200\,\mathrm{nA}$, and $I_{\mathrm{b}} = 20\,\mathrm{nA}$, the power consumption of the prototyped RCCM, which includes 16 IACs, 16 IPEs, and 256 WPEs, is $82.9\,\mu\mathrm{W}$. The total power consumption is estimated as $164.7\,\mu\mathrm{W}$ when the ReLU is chosen as the activation function. If the MAC operation can be performed within $1.2\,\mu\mathrm{s}$, the estimated power efficiency is $2.57\mathrm{TOPS/W}$.



(a)

(b)

(c)

Fig. 12. Measured multiplication results in different sign modes. (a) Both input and weight integers are unsigned. (b) The input integer is unsigned, and the weight integer is signed. (c) Both the input and weight integers are signed.

IV. Measurement Results

A prototype chip, including a version of the proposed RCCM, has been designed and fabricated in a $0.18\,\mu\mathrm{m}$ CMOS logic process. The chip micrograph is shown in Fig. 11(a). The demonstrative RCCM comprises a 16×16 WPE array, 16×1 IPE and IAC arrays, and a 1×16 AF array, occupying an area of $542\,\mu\mathrm{m} \times 239\,\mu\mathrm{m}$, as shown in Fig. 11(b). Five SRAM bit cells are employed for each PE with the LSB padded with zero for 4-bit resolution, ranging from -8 to 7 and from 0 to 15 for signed and unsigned integers, respectively. The maximum value of the stored signed or unsigned integers can be 8 or 16 if the LSB is set to one. A serial peripheral interface, synthesized from the standard library cells, and SRAM interface circuitries, including the read/write controls, decoders, and readout sensing amplifiers, are located at the peripheries of the matrix.

A. Multiplication in Different Data Representations

The output currents from a WPE are measured in different sign modes to verify the multiplication operations in diverse digital representations. The measured multiplication output currents are shown in Fig. 12(a), (b), and (c), demonstrating 1-quadrant, 2-quadrant, and 4-quadrant multiplications, respectively. Besides, four time-varying sinusoidal currents, of which frequencies are 1×, 3×, 5×, and 7× of 194Hz, are adopted as four analog current inputs to the designed RCCM chip. The calculated four-bit weight values based on discrete cosine transformation (DCT) are loaded into the SRAM cells in the corresponding IPEs and WPEs to synthesize a time-varying square wave. The resulting current from measurements is compared with the theoretical values in Fig. 13, demonstrating

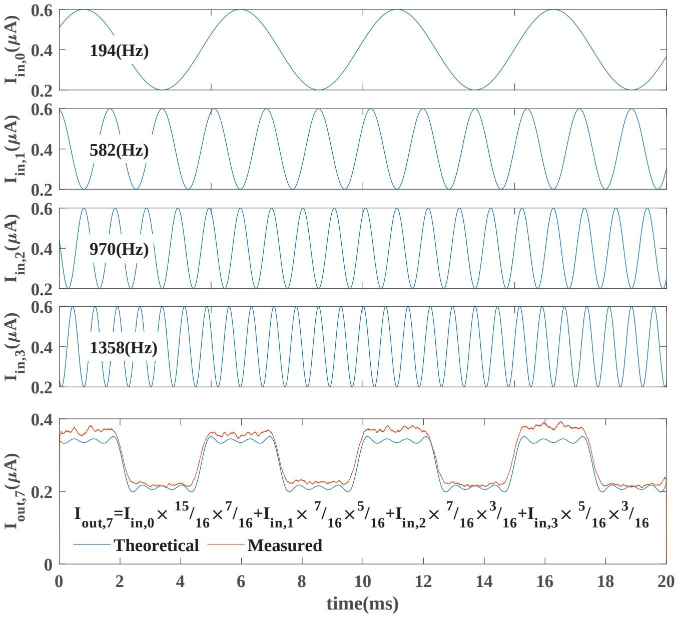
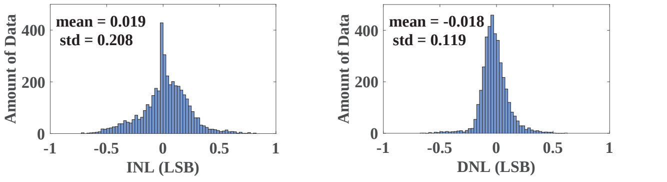


Fig. 13. The measured input and output currents demonstrate multiplication and accumulation (MAC) of the proposed RCCM with real-time analog input sinusoidal currents. Based on the discrete cosine transformation, the RCCM output current emulates a time-varying square waveform from four sinusoidal input currents with proper weight values.



(a) (b)

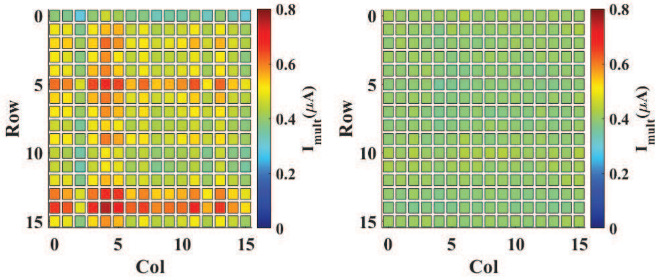
Fig. 14. The histogram of the distribution of measured INL and DNL values across the 16 × 16 weight array while sweeping the weight from 0 to 15 in the unsigned mode.

the multiplication and accumulation operations between analog input currents and unsigned weight values.

B. Process Variation and Calibration

The effects of process variation on the accuracy of the LB-DACs are evaluated by characterizing the integral nonlinearity (INL) and differential nonlinearity (DNL) across the 16 × 16 weight array while sweeping the weight from 0 to 15 in the unsigned mode. The histograms of the measured INL and DNL values are shown in Fig. 14. The means and standard deviation values are far below 0.5 LSBs, indicating the process variation is tolerable in implemented 4-bit LBDACs.

The output currents from all 16 × 16 WPEs are measured with all input and weight integers set to 15 to assess the impact of process variation on computation accuracy. With the same digital code, the output currents exhibit strong row and column dependence, as shown in Fig. 15(a). The row dependence stems from the insufficient sizing of the mirror transistors in the FVF-CM in the input auxiliary control (IAC) block shown in Fig. 9. Similarly, the column dependence is attributed to the insufficient sizing of the mirror transistors in the FVF-CMs in the activation function (AF) block shown in Fig. 10. To address this issue, the row and column mismatch ratios are characterized first. The products of characterized row and column mismatch ratios can then be employed for mismatch compensation in the corresponding WPEs. Moreover, since FVF-CMs for $I_{\rm outp}$ and for $I_{\rm outn}$ differ, the column mismatches for $I_{\rm outp}$ and $I_{\rm outn}$ are calibrated separately. As a result, the designed 256 WPEs can be calibrated using 16 parameters for the row mismatches and 32 parameters for the column mismatches, respectively. After calibration with these 48 mismatch parameters, the output currents in the unsigned weight mode are plotted in Fig. 15(b) with significantly better uniformity.



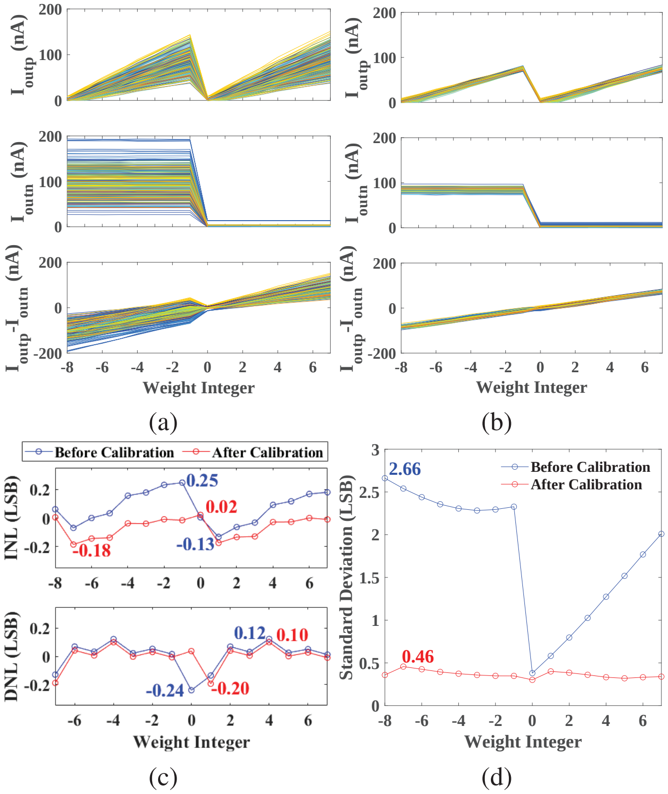
(a) (b)

Fig. 15. (a) Measured output currents from all WPEs show substantial column- and row-dependent variations. (b) After calibration, measured output currents from all WPEs exhibit better uniformity.

When the signed weight values are swept from -8 to 7, the measured output currents from all WPEs are plotted in Fig. 16(a), along with corresponding differential currents. After calibration, the output and the differential currents are plotted in Fig. 16(b) with remarkable variation reduction. The measured mean values are employed to calculate the integral nonlinearity (INL) and differential nonlinearity (DNL) for error and linearity characterization. As shown in Fig. 16(c), the adopted processing elements exhibit low average computation error and good linearity. Finally, the standard deviations with all different signed weight values before and after calibration are plotted in Fig. 16(d). The maximum standard deviation is dramatically reduced by more than 5.78 times (from 2.66 LSBs to 0.46 LSBs), validating the proposed calibration approach.

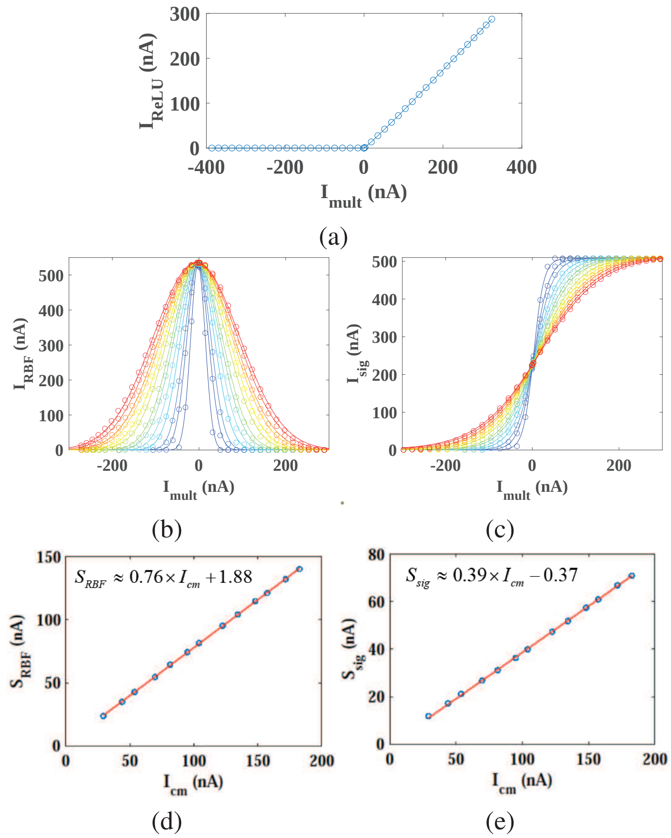
C. Activation Function Characterization

The circuits for activation function evaluation are characterized by sweeping an input current, $I_{\rm in}$, with a weight value of +7 for the positive input range and a weight value of -7 for the negative input range. The measured transfer curve from the circuit in Fig. 10(b) for ReLU function emulation is shown in Fig. 17(a). The measured transfer characteristics from the circuit for RBF and logistic function evaluation are shown in Fig. 17(b) and (c), respectively, with several levels of the input common-mode current, $I_{\rm cm}$. The transfer curves



(a) (b) (c) (d)

Fig. 16. (a) Measured output currents from all WPEs before calibration reveal considerable variations due to component mismatches. (b) Measured output currents from all WPEs exhibit significantly reduced variations after calibration. (c) The integral nonlinearity (INL) and differential nonlinearity (DNL) calculated from measured mean values characterize computation errors and circuit linearity. (d) Measured standard deviations with different weight values before and after calibration to verify the effectiveness of the calibration procedure.



(a) (b) (c) (d) (e)

Fig. 17. (a) Measured transfer characteristics of the circuit for ReLU function evaluation. (b) Measured transfer characteristics of the circuit for RBF evaluation. (c) Measured transfer characteristics of the circuit for logistic function evaluation. (d) The measured scaling coefficient S_{RBF} linearly depends on I_{cm}. (e) The measured scaling coefficient S_{sig} linearly depends on I_{cm}.

can be approximated as

$$I_{\mathrm{RBF}} \approx I_{\mathrm{cnst}} \cdot e^{-(I_{\mathrm{mult}}/S_{\mathrm{RBF}})^2} \tag{11}$$

$$I_{\mathrm{sig}} \approx \frac{I_{\mathrm{cnst}}}{1 + e^{-I_{\mathrm{mult}}/S_{\mathrm{sig}}}}, \tag{12}$$

where S_{RBF} and S_{sig} are scaling factors, exhibiting linear dependence on I_{cm} as shown in Fig. 17(d) and (e).

D. MNIST and CIFAR-10 Demonstration

A handwritten digit database, the Modified National Institute of Standards and Technology database (MNIST), is employed to demonstrate the functionality and characterize the performance of the designed RCCM chip. The samples from the MNIST database are resized to 8×8 by chopping off pixels at edges and performing average pooling within every 3×3 windows. A three-layer 64-64-16-10 fully-connected neural network is trained with the input and weight resolutions of 4 bits. Due to the limited input dimension of the prototype RCCM chip, the first two layers are implemented in computer software. The pre-trained weights in the third layer and the computation outputs from the second layer are sent to the RCCM chip through the SPI and are stored in the on-chip SRAM arrays. Output currents from ten WPE columns are measured and compared as inference results.

At first, the fully-connected network is trained without considering the mismatches due to the process variation. Compared with 95.86% from the entire network implemented in computer software, the accuracy obtained from the designed RCCM chip that implements the third layer drops to 94.03% due to the process variation. Since the power consumption depends on the input data, the histogram of measured power consumption for all 10,000 test samples is plotted in Fig. 18(a) with the mean value of $63.3\,\mu\mathrm{W}$.

With the characterized mismatch ratio parameters (R_{rw}, $R_{\mathrm{cl,p}}$, and $R_{\mathrm{cl,n}}$), the weights in integer (W_{int}) to be uploaded to the RCCM chip can be calibrated and rounded from the original weights in real (W_{real}) and clamped between -8 and 8 in runtime. The pseudocode of the runtime calibration is listed in Algo. 1. The accuracy from the weight-calibrated RCCM chip increases to 95.58%, close to the performance obtained through the software. The same experiment was performed with different supply voltages to investigate the effects of voltage variation. As shown in Fig. 18(b), the accuracy drops slightly when the supply voltage drops to $1.5\,\mathrm{V}$ and degrades significantly when the supply voltage is $1.4\,\mathrm{V}$ due to the insufficient headroom for FVF-CMs. The prototyped RCCM chip performs the same experiment at different temperatures as shown in Fig. 18(c). The accuracy variation is less than 0.5% when the temperature is swept from $0°C$ to $80°C$.

Algorithm 1: Weight calibration in runtime

```
% Update $W_{\mathrm{int}}$ from $W_{\mathrm{real}}$
For Each (row,col) do
    if $W_{\mathrm{real}}(row, col) >= 0$ then
        $W_{\mathrm{real}}(row, col) \leftarrow \frac{W_{\mathrm{real}}(row,col)}{R_{\mathrm{rw}}(row)\cdot R_{\mathrm{cl,p}}(col)}$
    else
        $W_{\mathrm{real}}(row, col) \leftarrow$
        $\frac{(W_{\mathrm{real}}(row,col)+8\cdot R_{\mathrm{rw}}(row)\cdot R_{\mathrm{cl,n}}(col))}{R_{\mathrm{rw}}(row)\cdot R_{\mathrm{cl,p}}(col)} - 8$
Rounding from $W_{\mathrm{real}}$ to $W_{\mathrm{int}}$
For Each (row,col) do
    if $W_{\mathrm{int}}(row, col) > 8$ then
        $W_{\mathrm{int}}(row, col) \leftarrow 8$
    if $W_{\mathrm{int}}(row, col) < -8$ then
        $W_{\mathrm{int}}(row, col) \leftarrow -8$
Upload $W_{\mathrm{int}}$ to the RCCM chip for inference
```

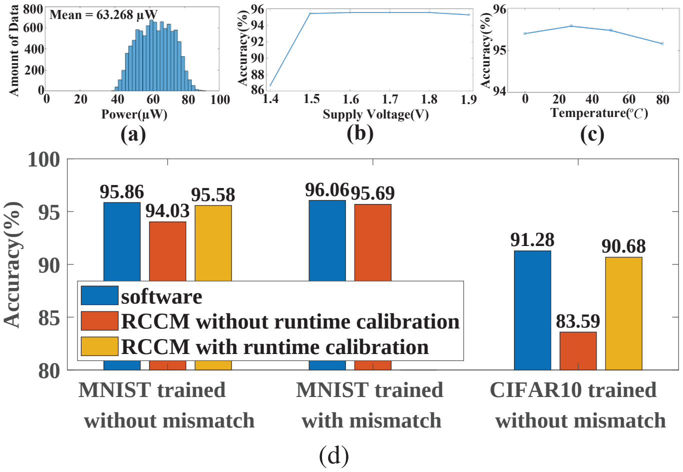


Fig. 18. (a) The histogram of the RCCM chip power consumption in inference with all MNIST test samples. (b) The MNIST accuracy dependence on the supply voltage. (c) The MNIST accuracy dependence on the temperature.(d) The accuracies of classification tasks using the MNIST and CIFAR-10 databases with different training and calibration settings.

Secondly, the network training process can include previously characterized 48 mismatch ratios. In the offline training process, the loss function and the gradient for backpropagation are calculated based on W_{real}, which are derived from existing weights in integer (W_{int}) and mismatch ratios (R_{rw}, $R_{\mathrm{cl,p}}$, and $R_{\mathrm{cl,n}}$). Subsequently, once the W_{real} values have been updated, an updated set of W_{int} is generated through a stochastic rounding process. Upon completion of the training process, the trained W_{int} values can be uploaded to the RCCM chip for inference, accounting for the process variation. In this case, the accuracy obtained from the designed RCCM chip only drops slightly from 96.06% to 95.69%, as shown in Fig. 18(d). The pseudocode of the offline training process is listed in Algo. 2.

A ResNet-20 implementation with reference to [43], consisting of 20 layers of residual blocks, is adopted to verify the chip performance on the CIFAR-10 dataset. The last layer is modified to a two-layer [64-16-10] network so that the prototyped RCCM chip can be employed to compute the last layer. The implemented architecture is summarized in Table I. The accuracy obtained from the software is 91.28%. Without considering the mismatches, the accuracy obtained from the RCCM chip drops to 83.59%. If the runtime calibration is adopted to adjust the weight values according to the characterized mismatch ratios, the accuracy improves to 90.68%, as indicated in Fig. 18(d).

Algorithm 2: Offline Training with 48 mismatch ratios

```
For Each batch do
    Update $W_{\mathrm{real}}$ through back-propagation
    % Calibrate the $W_{\mathrm{real}}$ with mismatch ratios
    For Each (row,col) do
        if $W_{\mathrm{real}}(row, col) >= 0$ then
            $W_{\mathrm{real}}(row, col) \leftarrow \frac{W_{\mathrm{real}}(row,col)}{R_{\mathrm{rw}}(row)\cdot R_{\mathrm{cl,p}}(col)}$
        else
            $W_{\mathrm{real}}(row, col) \leftarrow$
            $\frac{(W_{\mathrm{real}}(row,col)+8\cdot R_{\mathrm{rw}}(row)\cdot R_{\mathrm{cl,n}}(col))}{R_{\mathrm{rw}}(row)\cdot R_{\mathrm{cl,p}}(col)} - 8$
    Stochastic rounding from $W_{\mathrm{real}}$ to $W_{\mathrm{int}}$
    % Calculate $W_{\mathrm{real}}$ from $W_{\mathrm{int}}$ with mismatch ratios
    For Each (row,col) do
        if $W_{\mathrm{int}}(row, col) >= 0$ then
            $W_{\mathrm{real}}(row, col) \leftarrow$
             $W_{\mathrm{int}}(row, col) \cdot R_{\mathrm{rw}}(row) \cdot R_{\mathrm{cl,p}}(col)$
        else
            $W_{\mathrm{real}}(row, col) \leftarrow$
            $(W_{\mathrm{int}}(row, col) + 8) \cdot R_{\mathrm{rw}}(row) \cdot R_{\mathrm{cl,p}}(col)$
            $-8 \cdot R_{\mathrm{rw}}(row) \cdot R_{\mathrm{cl,n}}(col)$
Upload $W_{\mathrm{int}}$ to the RCCM chip for inference
```

TABLE I
IMPLEMENTED RESNET-20 ARCHITECTURE

Layer type	CNN	CNN	CNN	FC	FC
size	32 × 32	16 × 16	8 × 8	64 × 16	16 × 10
# of Layers	7	6	6	1	1
# of Filters	16	32	64	NA	NA

E. Power and Energy Distribution

When the RCCM chip is in the MO mode, the bias currents of the RCCM chip are turned of. When the chip is switched in the CC mode, the bias currents for the FVF-CMs located in the IAC and AF blocks are turned on with the standby power of $34\,\mu$W. The employed serial peripheral interface (SPI) for digital data upload is synthesized from standard library cells, operating at a $5\,$MHz clock rate with $7\,\mu$W power consumption. It was designed for functionality only and has not been optimized for speed and power consumption. The power distribution of the prototyped chip is illustrated in Fig. 19(a). Although the proposed RCCM is not designed for convolution neural networks (CNN) specifically, a 16-channel one-layer CNN with a kernel size of 4×4 can be implemented in the prototyped RCCM chip, as shown in Fig. 19(b). If

TABLE II
COMPARISON TABLE

	JSSC [21]	JSSC [30]	JSSC [28]			JSSC [29]			This Work	
Year	2019	2021	2020			2021			2023	
Technology (nm)	65	7	55			28			180	
Supported Input Type	Signed	Unsigned	Unsigned			Unsigned			Signed/Unsigned	analog current
Input Precision (bit)	7	4	1	2	4	4	4	8	4	analog current
Supported Weight Type	Signed	Signed	Signed			Signed			Signed/Unsigned	
Weight Precision (bit)	1	4	2	5	5	4	8	8	4	4b*4b
Output Precision (bit)	7	4	3	5	7	12	16	20	analog current	
Output Type	digital bits	digital bits	digital bits			digital bits			current	
Activation Functions	N/A	N/A	N/A			N/A			ReLU, RBF, Sigmoid	
Energy Efficiency (TOPS/W)	40.3∼51.3 $(6.69)^1$ $(46.83)^2$	351 $(0.53)^1$ $(8.49)^2$	72.03 $(6.73)^1$ $(13.45)^2$	37.5 $(3.50)^1$ $(35.01)^2$	18.37 $(1.72)^1$ $(34.30)^2$	68.44 $(1.66)^1$ $(26.50)^2$	33.52 $(0.81)^1$ $(25.96)^2$	16.63 $(0.40)^1$ $(25.75)^2$	3.355 $(3.355)^1$ $(53.68)^2$	
Network for MNIST	LeNet-5	784-64-16-10	ResNet20			N/A			64-64-16-10[3]	
MNIST Accuracy(%)	98∼98.3	96∼98.5	99.02	99.18	99.52	N/A			95.69	
Array Size	2kB	4kb	3.75kb			64kb			1kb	

[1] The presented energy efficiency is normalized to a 180-nm implementation, assuming energy $\propto (\text{Tech.})^2$ [21].

[2] The presented energy efficiency is normalized by the equivalent operand precision of 1 bit in a 180-nm implementation, assuming energy $\propto (\text{Tech.})^2$ [44].

[3] Only the third-layer is implemented on chip.

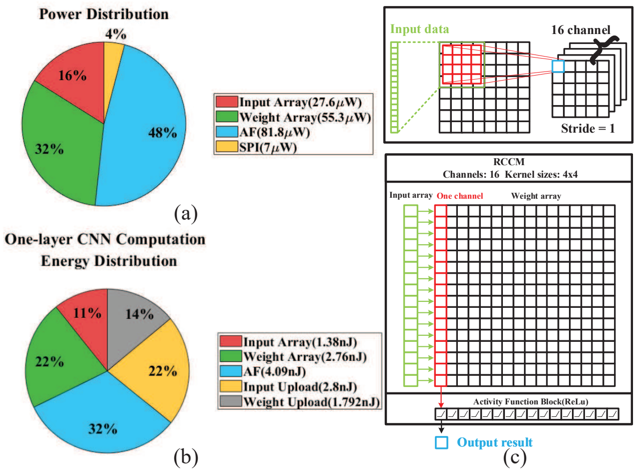


Fig. 19. (a) The power distribution of the prototyped RCCM chip.(b) The energy distribution of the CNN implementation in (c). (c) The configuration of a convolution neural network (CNN) implementation using the prototyped RCCM chip.

the image size is 8×8 and the stride is 1, the computation energy distribution of such one-layer CNN can be estimated and shown in Fig. 19(c).

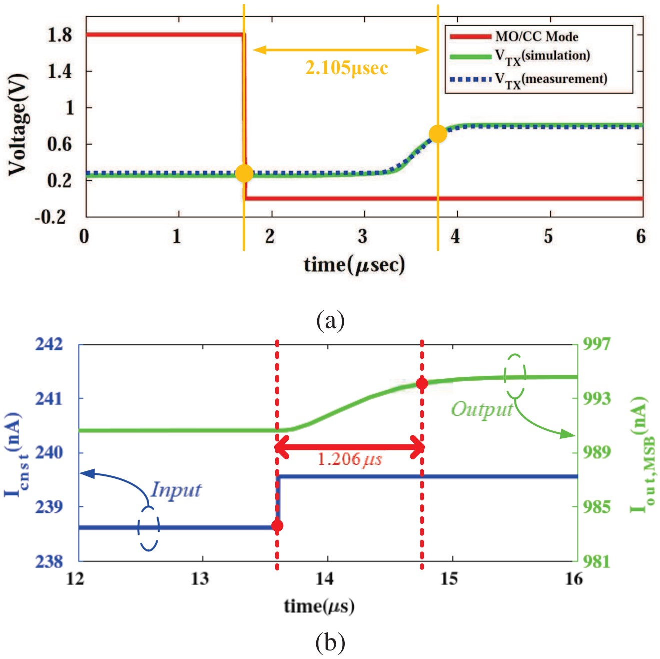


Fig. 20. (a) The measured and simulated transient responses when the RCCM chip is switched from the MO mode to the CC mode. (b) The post-layout simulation results of a step response to estimate the MAC operation time of the designed RCCM.

F. Estimated Power Efficiency and Comparison

The power efficiency of the RCCM can be characterized by the number of trillion operations per second (TOPs) per watt. The operation mode of the RCCM chip is switched from the memory operation (MO) mode to the cognitive computation (CC) mode to observe the chip response time. In the MO mode, since the transistor M_{1a} shown in Fig. 8(b) is turned off, no currents flow through LBDACs. Once the RCCM chip is switched to the CC mode, the transistor M_{1a} is activated, and the bias current I_{cnst} shown in Fig. 9 is injected into the IPE. The computed input magnitude current, I_{mag}, is duplicated and injected into the WPEs. The resultant MAC computation output currents charge diode-connected nodes (V_{TX}) in the transmitter blocks (TX), which can be observed through an on-chip buffer. The measured overall response time of the RCCM chip is 2.1 μs, which is in line with the post-layout simulation as shown in Fig. 20(a). The same post-layout simulation setting is adopted to estimate the MAC operation response time. From the post-layout simulation results shown in Fig. 20(b), the MAC operation takes $1.206\,\mu$s in the proposed RCCM. Therefore, the estimated power efficiency of the proposed RCCM is $^{256\text{Ops}}/_{1.206\,\mu\text{s} \times 63.268\,\mu\text{W}} = 3.355\text{TOPs/W}$ with a $1.8\,\text{V}$ supply voltage. Finally, the performance of the proposed

SRAM-based RCCM is compared with state-of-the-art in Table II. Assuming that the computation speed bottleneck is the parasitic capacitance, which is proportional to the square of the feature dimension, the efficiency numbers can be normalized accordingly [44]. The normalized energy efficiency achieved by the prototyped RCCM chip is comparable to that of other state-of-the-art chips fabricated in more advanced processes.

V. Conclusion

An SRAM-based reconfigurable cognitive computation matrix (RCCM) with extensive data representation flexibilities is presented in this paper. The proposed RCCM performs vector-matrix multiplication between an input vector, of which elements can be analog currents or digital integers, and a weight integer matrix. Since the RCCM can be reconfigured to carry out either 1-quadrant, 2-quadrant, or 4-quadrant multiplications, the digital integers stored in the SRAM for inputs and weights can be in the signed or unsigned format. Besides, three commonly used activation functions, the rectified linear unit (ReLU), radial basis function (RBF), and logistic function, are also available and evaluated in the analog domain, rendering single-ended currents as the computation results. Therefore, multi-layer networks can be realized by cascading multiple RCCM chips. A concept-proving prototype chip is designed and fabricated in a $0.18\,\mu\mathrm{m}$ CMOS process, including a 16×16 RCCM with 4-bit input and weight resolutions. The measurement results verify the diversity of supported data representations. The computation errors and variation induced by process variation are characterized. A calibration procedure employing 48 mismatch parameters is proposed to improve the computation accuracy and reduce variations. The chip performance is verified with a handwritten digit database, MNIST, achieving an accuracy of 95.69% with a $1.8\,\mathrm{V}$ supply voltage. The estimated average efficiency of the proposed RCCM chip is $3.355\mathrm{TOPS/W}$.

References

[1] H. Hao, J. Chen, A. G. Richardson, J. Van der Spiegel, and F. Aflatouni, “A 10.8 μ W neural signal recorder and processor with unsupervised analog classifier for spike sorting,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 2, pp. 351–364, 2021.

[2] W. Shan, M. Yang, T. Wang, Y. Lu, H. Cai, L. Zhu, J. Xu, C. Wu, L. Shi, and J. Yang, “A 510-nw wake-up keyword-spotting chip using serial-fft-based mfcc and binarized depthwise separable cnn in 28-nm cmos,” *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 151–164, 2021.

[3] Z. Wang, Y. Liu, P. Zhou, Z. Tan, H. Fan, Y. Zhang, L. Shen, J. Ru, Y. Wang, L. Ye, and R. Huang, “A 148nw reconfigurable event-driven intelligent wake-up system for aiot nodes using an asynchronous pulse-based feature extractor and a convolutional neural network,” *IEEE Journal of Solid-State Circuits*, vol. 56, no. 11, pp. 3274–3288, 2021.

[4] Y. Liu, Z. Wang, W. He, L. Shen, Y. Zhang, P. Chen, M. Wu, H. Zhang, P. Zhou, J. Liu, G. Sun, J. Ru, L. Ye, and R. Huang, “An 82nw 0.53pj/sop clock-free spiking neural network with 40us latency for alot wake-up functions using ultimate-event-driven bionic architecture and computing-in-memory technique,” in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 372–374.

[5] D. Rossi, F. Conti, M. Eggiman, A. D. Mauro, G. Tagliavini, S. Mach, M. Guermandi, A. Pullini, I. Loi, J. Chen, E. Flamand, and L. Benini, “Vega: A ten-core soc for iot endnodes with dnn acceleration and cognitive wake-up from mram-based state-retentive sleep mode,” *IEEE Journal of Solid-State Circuits*, vol. 57, no. 1, pp. 127–139, 2022.

[6] M. Eggimann, A. Rahimi, and L. Benini, “A 5 μ W standard cell memory-based configurable hyperdimensional computing accelerator for always-on smart sensing,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 10, pp. 4116–4128, 2021.

[7] Y. Lu, V. L. Le, and T. T.-H. Kim, “A 184-¡inline-formula¿ ¡tex-math notation=”latex”¿μ¡/tex-math¿ ¡/inline-formula¿w error-tolerant real-time hand gesture recognition system with hybrid tiny classifiers utilizing edge cnn,” *IEEE Journal of Solid-State Circuits*, pp. 1–13, 2022.

[8] A. R. Nair, P. K. Nath, S. Chakrabartty, and C. S. Thakur, “Multiplierless mp-kernel machine for energy-efficient edge devices,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pp. 1–14, 2022.

[9] R. Dekimpe and D. Bol, “Ecg arrhythmia classification on an ultra-low-power microcontroller,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 3, pp. 456–466, 2022.

[10] H. Chu, Y. Yan, L. Gan, H. Jia, L. Qian, Y. Huan, L. Zheng, and Z. Zou, “A neuromorphic processing system with spike-driven snn processor for wearable ecg classification,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 511–523, 2022.

[11] Y. He, F. Corradi, C. Shi, S. van der Ven, M. Timmermans, J. Stuijt, P. Detterer, P. Harpe, L. Lindeboom, E. Hermeling, G. Langereis, E. Chicca, and Y.-H. Liu, “An implantable neuromorphic sensing system featuring near-sensor computation and send-on-delta transmission for wireless neural sensing of peripheral nerves,” *IEEE Journal of Solid-State Circuits*, vol. 57, no. 10, pp. 3058–3070, 2022.

[12] S. Tannirkulam Chandrasekaran, A. Jayaraj, V. Elkoori Ghantala Karnam, I. Banerjee, and A. Sanyal, “Fully integrated analog machine learning classifier using custom activation function for low resolution image classification,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 3, pp. 1023–1033, 2021.

[13] F. Chen, K.-F. Un, W.-H. Yu, P.-I. Mak, and R. P. Martins, “A 108-nw 0.8-mm ^{2} analog voice activity detector featuring a time-domain cnn with sparsity-aware computation and sparsified quantization in 28-nm cmos,” *IEEE Journal of Solid-State Circuits*, pp. 1–10, 2022.

[14] T.-H. Hsu, G.-C. Chen, Y.-R. Chen, C.-C. Lo, R.-S. Liu, M.-F. Chang, K.-T. Tang, and C.-C. Hsieh, “A 0.8v intelligent vision sensor with tiny convolutional neural network and programmable weights using mixed-mode processing-in-sensor technique for image classification,” in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[15] H. Xu, N. Lin, L. Luo, Q. Wei, R. Wang, C. Zhuo, X. Yin, F. Qiao, and H. Yang, “Senputing: An ultra-low-power always-on vision perception chip featuring the deep fusion of sensing and computing,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 1, pp. 232–243, 2022.

[16] J. Zhang, Z. Wang, and N. Verma, “In-memory computation of a machine-learning classifier in a standard 6t sram array,” *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, 2017.

[17] W.-S. Khwa, J.-J. Chen, J.-F. Li, X. Si, E.-Y. Yang, X. Sun, R. Liu, P.-Y. Chen, Q. Li, S. Yu, and M.-F. Chang, “A 65nm 4kb algorithm-dependent computing-in-memory sram unit-macro with 2.3ns and 55.8tops/w fully parallel product-sum operation for binary dnn edge processors,” in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 496–498.

[18] R. Liu, X. Peng, X. Sun, W.-S. Khwa, X. Si, J.-J. Chen, J.-F. Li, M.-F. Chang, and S. Yu, “Parallelizing sram arrays with customized bit-cell for binary neural networks,” in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 2018, pp. 1–6.

[19] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, “A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement,” in *2018 IEEE Symposium on VLSI Circuits*, 2018, pp. 141–142.

[20] A. Agrawal, A. Jaiswal, D. Roy, B. Han, G. Srinivasan, A. Ankit, and K. Roy, “Xcel-ram: Accelerating binary neural networks in high-throughput sram compute arrays,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 8, pp. 3064–3076, 2019.

[21] 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 Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, 2019.

[22] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, “A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, 2019.

[23] J. Yang, Y. Kong, Z. Wang, Y. Liu, B. Wang, S. Yin, and L. Shi, “24.4 sandwich-ram: An energy-efficient in-memory bwn architecture with pulse-width modulation,” in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 394–396.

[24] 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 Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, 2020.

[25] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, “Xnor-sram: In-memory computing SRAM macro for binary/ternary deep neural networks,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, 2020.

[26] J. Mu, H. Kim, and B. Kim, “SRAM-based in-memory computing macro featuring voltage-mode accumulator and row-by-row ADC for processing neural networks,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2412–2422, 2022.

[27] C. Yu, T. Yoo, K. T. C. Chai, T. T.-H. Kim, and B. Kim, “A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks,” *IEEE Journal of Solid-State Circuits*, pp. 1–1, 2022.

[28] X. Si, J.-J. Chen, Y.-N. Tu, W.-H. Huang, J.-H. Wang, Y.-C. Chiu, W.-C. Wei, S.-Y. Wu, X. Sun, R. Liu, S. Yu, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, Q. Li, and M.-F. Chang, “A twin-8t sram computation-in-memory unit-macro for multibit cnn-based ai edge processors,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, 2020.

[29] X. Si, Y.-N. Tu, W.-H. Huang, J.-W. Su, P.-J. Lu, J.-H. Wang, T.-W. Liu, S.-Y. Wu, R. Liu, Y.-C. Chou, Y.-L. Chung, W. Shih, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, N.-C. Lien, W.-C. Shih, Y. He, Q. Li, and M.-F. Chang, “A local computing cell and 6t sram-based computing-in-memory macro with 8-b mac operation for edge ai chips,” *IEEE Journal of Solid-State Circuits*, vol. 56, no. 9, pp. 2817–2831, 2021.

[30] 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 Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, 2021.

[31] 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.-Y. J. Chang, “A 5-nm 254-tops/w 221-tops/mm2 fully-digital computing-in-memory macro supporting wide-range dynamic-voltage-frequency scaling and simultaneous mac and write operations,” in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[32] P.-C. Wu, J.-W. Su, Y.-L. Chung, L.-Y. Hong, J.-S. Ren, F.-C. Chang, Y. Wu, H.-Y. Chen, C.-H. Lin, H.-M. Hsiao, S.-H. Li, S.-S. Sheu, S.-C. Chang, W.-C. Lo, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, C.-I. Wu, and M.-F. Chang, “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 *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[33] B. Yan, J.-L. Hsu, P.-C. Yu, C.-C. Lee, Y. Zhang, W. Yue, G. Mei, Y. Yang, Y. Yang, H. Li, Y. Chen, and R. Huang, “A 1.041-mb/mm2 27.38-tops/w signed-int8 dynamic-logic-based adc-less sram compute-in-memory macro in 28nm with reconfigurable bitwise operation for ai and embedded applications,” in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 188–190.

[34] B. Minch, “Synthesis of static and dynamic multiple-input translinear element networks,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 2, pp. 409–421, 2004.

[35] K. Odame, M. Nyamukuru, M. Shahghasemi, S. Bi, and D. Kotz, “Analog gated recurrent unit neural network for detecting chewing events,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 6, pp. 1106–1115, 2022.

[36] S.-Y. Peng, P. E. Hasler, and D. V. Anderson, “An analog programmable multidimensional radial basis function based classifier,” *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 54, pp. 2148–2158, 2007.

[37] B. Joo, J.-W. Han, and B.-S. Kong, “Energy- and area-efficient cmos synapse and neuron for spiking neural networks with stdp learning,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 9, pp. 3632–3642, 2022.

[38] C. M. Hammerschmied and Q. Huang, “Design and implementation of an untrimmed MOSFET-only 10-bit A/D converter with -79-db THD,” *IEEE Journal of Solid-State Circuits*, vol. 33, no. 8, pp. 1148–1157, 1998.

[39] D. Balobas and N. Konofaos, “Design and evaluation of 6t sram layout designs at modern nanoscale cmos processes,” in *4th International Conference on Modelling Circuits, Systems and Technology*, 2016, pp. 7–12.

[40] Helm, Kavanaugh, Liew, Petti, Stolmeijer, Ben-tzur, Bornstein, Lilygren, Ting, Trammel, Allan, Gray, Hartranft, Radigan, Shanmugan, and Shrivastava, “A low cost, microprocessor compatible, 18.4 um/sup 2/,6-t bulk cell technology for high speed srams,” in *Symposium 1993 on VLSI Technology*, 1993, pp. 65–66.

[41] J. Ramirez-Angulo, R. Carvajal, and A. Torralba, “Low supply voltage high-performance cmos current mirror with low input and output voltage requirements,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 51, no. 3, pp. 124–129, 2004.

[42] T. Delbruck, “Bump circuits for computing similarity and dissimilarity of analog voltage,” in *IEEE Proceedings of the International Neural Network Society*, vol. 1, Oct. 1991, pp. 475–479.

[43] Y. Idelbayev, “Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch,” https://github.com/akamaster/pytorch_resnet_cifar10, accessed: 2023-05-20.

[44] 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 Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 6, pp. 2977–2981, 2022.

[45] H. Kim, T. Yoo, T. T.-H. Kim, and B. Kim, “Colonnade: A reconfigurable sram-based digital bit-serial compute-in-memory macro for processing neural networks,” *IEEE Journal of Solid-State Circuits*, vol. 56, no. 7, pp. 2221–2233, 2021.

[46] A. Basu, L. Deng, C. Frenkel, and X. Zhang, “Spiking neural network integrated circuits: A review of trends and future directions,” in *2022 IEEE Custom Integrated Circuits Conference (CICC)*, 2022, pp. 1–8.

Sheng-Yu Peng (S’02-M’09-SM’19) received B.S. and M.S. degrees in Electrical Engineering from the National Taiwan University, Taipei, Taiwan, in 1995 and 1997, respectively; a degree of Master of Science in Electrical and Computer Engineering from the Cornell University, Ithaca, NY, in 2004; and the Ph.D. degree in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta, GA, in 2008. From 2008 to 2011, he worked for GTronix and MaxLinear, respectively.

Dr. Peng joined the National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2011. Currently, he is a professor in the Department of Electrical Engineering. His research interests include interface circuits for sensors and biomedical applications, reconfigurable analog circuits and systems, power-efficient analog signal processing, and low-power machine learning algorithms. Dr. Peng received NTUST 2022, 2021, and 2013 Excellent Teaching Awards, NTUST 2022 Excellent Research Award, IEEE Taipei Section 2018 Best Master Thesis Advisor Award, and IEEE Taipei Section 2018 Best Ph.D. Dissertation Advisor Award. He also received the Best Student Paper Award at the 2016 IEEE International Ultrasonics Symposium.

I-Chun Liu received the B.S. degree in the Department of Electrical Engineering from the Fu Jen Catholic University, Taipei, Taiwan, in 2021. She is currently working toward her M.S. degree at the National Taiwan University of Science and Technology in the lab of Intelligent Sensory Microsystems with Advanced Reconfigurable Technologies. Her research interests include low-power analog integrated circuits and computation in memories for biomedical applications.

Yi-Heng Wu received the B.S. and M.S. degrees in the Department of Electrical Engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2019 and 2022, respectively. His research interests include low-power analog integrated circuits and computation in memories for biomedical applications.

Ting-Ju Lin received the B.S. degree in the Department of Electrical Engineering from the Chang Gung University, Taoyuan, Taiwan, in 2020. She is currently working toward her M.S. degree at the National Taiwan University of Science and Technology in the lab of Intelligent Sensory Microsystems with Advanced Reconfigurable Technologies. Her research interests include low-power analog integrated circuits and computation in memories for biomedical applications.

Chun-Jui Chen received the B.S. degree in the Department of Electrical Engineering from the Chang Gung University, Taoyuan, Taiwan, in 2020. He is currently working toward his M.S. degree at the National Taiwan University of Science and Technology in the lab of Intelligent Sensory Microsystems with Advanced Reconfigurable Technologies. His research interests include low-power analog integrated circuits and artificial intelligence circuit systems for biomedical applications.

Xiu-Zhu Li received the B.S. degree in the Department of Electrical Engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2022. He is currently working toward his M.S. degree at the National Taiwan University of Science and Technology in the lab of Intelligent Sensory Microsystems with Advanced Reconfigurable Technologies. His research interests include low-power analog integrated circuits and computation in memories for neural networks applications.

Yong-Qi Cheng received the B.S. degree in the Department of Electrical Engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2022. He is currently working toward his M.S. degree at the National Taiwan University of Science and Technology in the lab of Intelligent Sensory Microsystems with Advanced Reconfigurable Technologies. His research interests include low-power analog integrated circuits and computation in memories for neural networks applications.

Pin-Han Lin is pursuing his B.S. degree in the Department of Electrical Engineering at the National Taiwan University of Science and Technology, Taipei, Taiwan, and has been working with the lab of Intelligent Sensory Microsystems with Advanced Reconfigurable Technologies as an undergraduate researcher since June 2022. His research interests include low-power analog integrated circuits and computation in memories for neural networks applications.

Kuo-Hsuan Hung received the B.S. and M.S. degrees from the National Chiao Tung University and National Central University, Taiwan, in 2015 and 2017, respectively. He is currently working toward the Ph.D. degree with the Department of Biomedical Engineering, National Taiwan University. He is a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include biomedical signal processing, noise reduction, speaker recognition, and deep learning.

Yu Tsao (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. He is also a Jointly Appointed Professor with the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. His research interests include assistive oral communication technologies, audio coding, and bio-signal processing. He is currently an Associate Editor for the IEEE/ACM Transactions on Audio, Speech, and LanguagE Processing and IEEE SignaL Processing Letters. He was the recipient of the Academia Sinica Career Development Award in 2017, national innovation awards in 2018–2021, Future Tech Breakthrough Award 2019, and Outstanding Elite Award, Chung Hwa Rotary Educational Foundation 2019–2020. He is the corresponding author of a paper that received the 2021 IEEE Signal Processing Society (SPS), Young Author, Best Paper Award.