A 65nm Thermometer-Encoded Time/Charge-Based Compute-in-Memory Neural Network Accelerator at 0.735pJ/MAC and 0.41pJ/Update

Minxiang Gong, *Student Member, IEEE*, Ningyuan Cao, *Student Member, IEEE*, Muya Chang, *Student Member, IEEE*, Arijit Raychowdhury, *Senior Member, IEEE*

***Abstract*—This paper presents an in-memory compute macro for neural-network based controllers including inference and in-situ weight updates featuring: (1) in-memory multi-bit matrix and transposed matrix multiplication; (2) thermometer-encoded, pulse-modulated storage element for in-memory weight update; (3) adaptive bit-line analog-to-digital (A/D) conversion for enhanced area/power efficiency. The chip was fabricated in 65nm CMOS technology and measured an energy efficiency of 0.735pJ/multiply-accumulate operation (MAC) and 0.41pJ/weight update.**

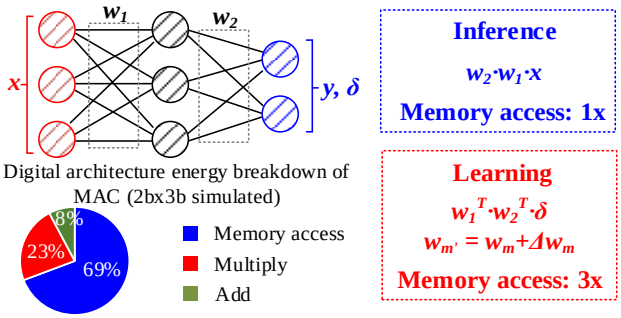
***Index Terms*—Compute-in-memory, matrix multiplication, neural network, mixed-signal, on-chip learning.**

I. Introduction

With the proliferation of internet of things (IoT) and edge intelligence (EI), the integrated circuits (ICs) for edge computing are required to maintain high energy-efficiency across a wide range of operating conditions. Such an SoC was presented in [1], where the on-board computation was optimized with the amount of data to be transferred to the cloud (communication), such that minimum operating power is obtained. The trade-off between computation and communication is enabled by a low-power neuro-inspired controller that leverages an actor-critic based reinforcement learning. Conventional model-based digital architectures are not well-suited for the task because: 1) low energy efficiency due to memory access (Fig. 1) and 2) die-to-die and environmental variations that render the state-space to be extremely large. Therefore, the model-free controller with a small number of parameters (in the order of 100) enabled by neural networks is a superior technology that can easily outperform model-based systems. Since neural networks feature data extensive vector multiplication, compute-in-memory (CIM) architectures show significant improvement of energy efficiency and performance [2]-[5]. Examples include switched-capacitor matrix multipliers [2], time-domain solvers

This work was supported by the Semiconductor Research Corporation under grant JUMP CBRIC task ID 2777.006.

Minxiang Gong, Ningyuan Cao, Muya Chang, and Arijit Raychowdhury are with Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: mxgong@gatech.edu, ncao@gatech.edu, mchang87@gatech.edu, arijit.raychowdhury@ece.gatech.edu).



	Weight/output precision	Forward	Backward	In memory weight update
[2]	6b/6b	✓	✖	✖
[3]	5b/5b	✓	✖	✖
[4]	1b/7b	✓	✖	✖
This work	3b/8b	✓	✓	✓

Fig. 1. Inference, learning and memory access of neural networks, digital architecture energy breakdown of MAC, and features of this work.

[3], binary convolution neural networks (CNN) [4] and binary output classifiers [5]. Despite advances in CIM for inference, learning remains challenging. In the learning process, the error vector at the output propagates backward and is multiplied by the transpose of the weight matrix in each layer. Thus, the weight matrix and its transpose form should both be easily accessible in an efficient CIM design, and the number of memory accesses for weight updates need to be decreased. Moreover, SRAM-based CIM designs are limited by SRAM's binary storage and the overhead of multi-bit A/D conversion [3]. Some designs support only binary multiplicands [4] [5] [6], while some use binary outputs [5] [6]. This severely limits the applicability of CIM designs.

This work presents a CIM architecture with in-memory weight update for vector/matrix multiplication. Based on SRAM, a new storage circuit is proposed to enable transpose access and in-situ weight updates. We demonstrate a thermometer-encoding scheme for storage cells that use pulse-width modulation (PWM) for weight updates. For optimized area and energy efficiency, an adaptive A/D conversion scheme is implemented with a low-resolution analog-to-digital converter (ADC) with the ability to scale up 8b output ENOB. By embedding computing in the memory, we eliminate data-movement from memory to logic and estimate a 1.77x and 4.63x energy reduction for MAC and weight update compared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2020.3027801, IEEE Transactions on Circuits and Systems II: Express Briefs

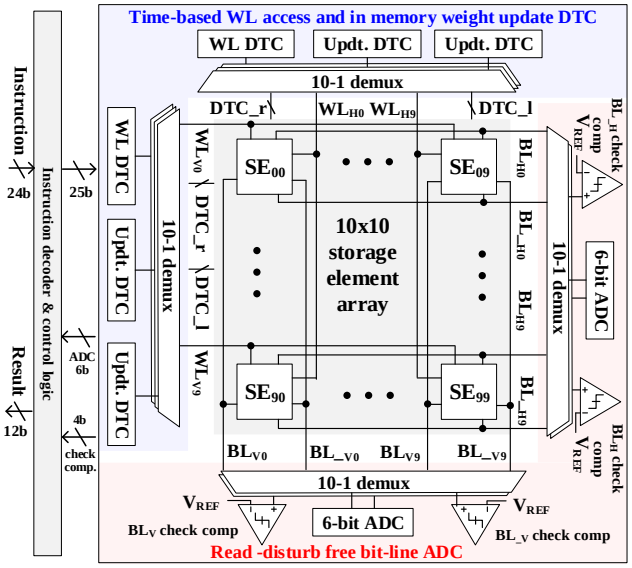


Fig. 2. CIM system architecture. (Left and bottom peripheries are for matrix (inference). Top and right peripheries are for matrix transpose (learning)).

to an iso-resolution baseline fully digital design (simulated from 65nm technology), respectively.

Rest of the paper is organized as follows: section II describes the system architecture and circuit design. Section III deals with the measurement results and comparison with state-of-the-art. Conclusions are drawn in the section IV.

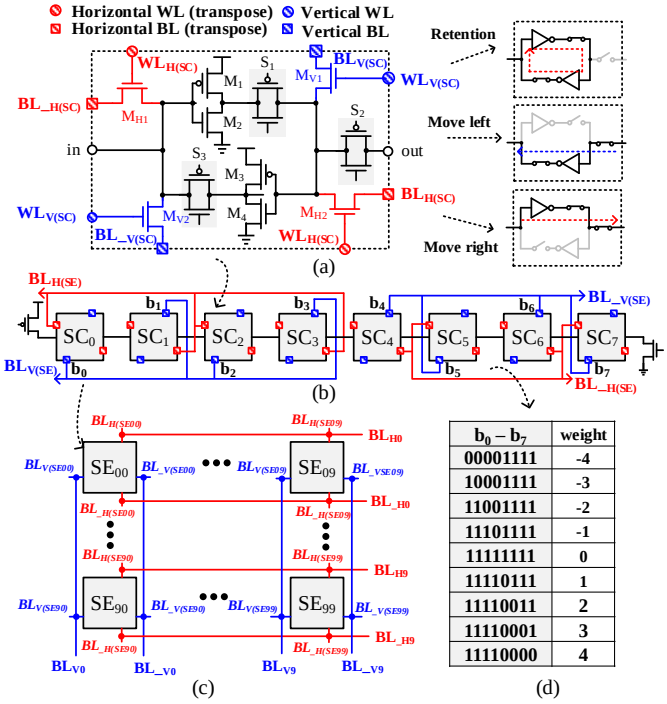
II. System Architecture and Circuit Design

A. System Architecture

The system architecture is shown in Fig. 2, which consists of storage element (SE) array, digital-to-time converters (DTC), ADC, and comparators. Instruction decoder and control logic serve as interfaces with outside circuit. At the core, 100 SEs form a 10 by 10 array with each SE vertically and horizontally connected to word-line (WL) and bit-line (BL) pairs to enable in-memory access (WL_V and BL_V for matrix; WL_H and BL_H for matrix transpose). At the periphery, three 3b DTCs in each direction, responsible for WL access, negative (minus) weight update and positive (plus) weight update. At the array output, 6b ADCs convert the differential voltage between BL and BL_ to the corresponding digital value, and comparators detect potential ADC overflow and read disturbance on BLs. The input operand ranges from 0 to 3 (2 bit), and weight ranges from -4 to 4 (3 bit). To save area, DTCs, ADCs and comparators are multiplexed across SE rows/columns.

B. Storage Circuit

To achieve in-memory multi-bit computation and weight update, a bidirectional storage cell (SC) and 3b thermometer-encoded SE are implemented. The detailed bottom-up circuit diagram and connections are shown in Fig. 3. Inside SC (Fig. 3 (a)), besides standard 6-transistor (6T) SRAM, it includes: (1) two additional WL transistors (M_{H1}, M_{H2}) and two additional output ports ($BL_{H(SC)}$, $BL\_{H(SC)}$) for transpose access (data is accessed via $WL_{V(SC)}$ or $WL_{H(SC)}$ during inferencing or learning, respectively) and (2) three transmission gates (S_1-S_3) to



$b_0 - b_7$	weight
00001111	-4
10001111	-3
11001111	-2
11101111	-1
11111111	0
11110111	1
11110011	2
11110001	3
11110000	4

Fig. 3. Bottom-up view of storage element array. (a) circuit schematic of 1b storage cell, (b) connection topology of single storage element (connections of WLs are not shown), (c) BL connections of storage element array, (d) thermometer encoding of storage element.

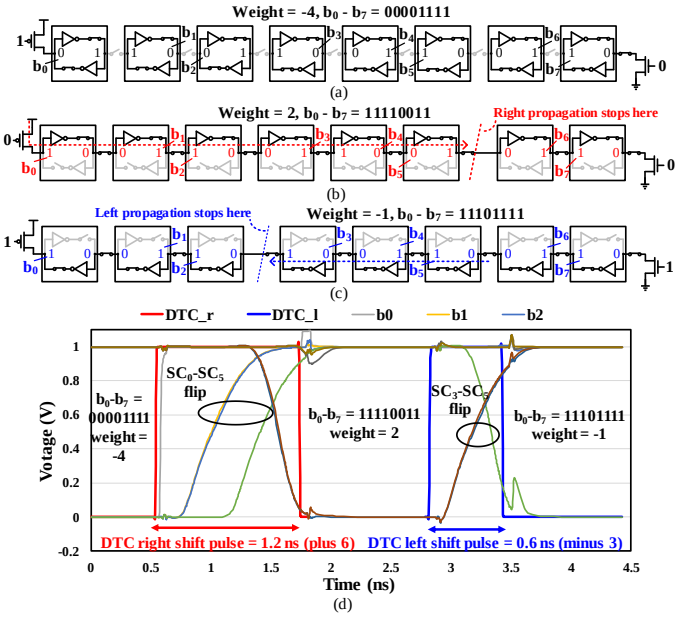


Fig. 4. Data movement in single storage element: (a) data value in SCs at initial state, (b) data value in SCs after right propagation, (c) data value in SCs after left propagation, and (d) simulated transient waveforms showing data flipping inside storage element.

facilitate data retention and in-situ data movement. In the data retention mode, S_1, S_3 are closed and S_2 is open, and either pair of WL transistors (M_{H1}, M_{H2} or M_{V1}, M_{V2}) can access the latched data. In the data movement mode, S_2 is always closed to provide data path between adjacent SCs. Depending on the direction of the data movement, either S_1 or S_3 is closed to provide unidirectional path (S_1 for right, S_3 for left). WL transistors are disabled during data movement.

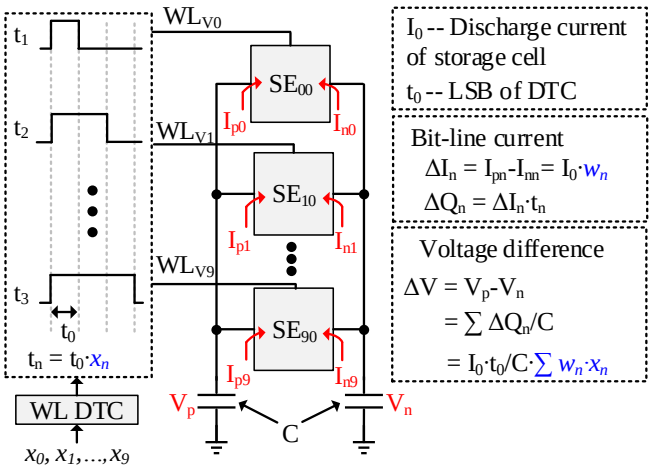


Fig. 5. Time/charge-domain vector multiplication.

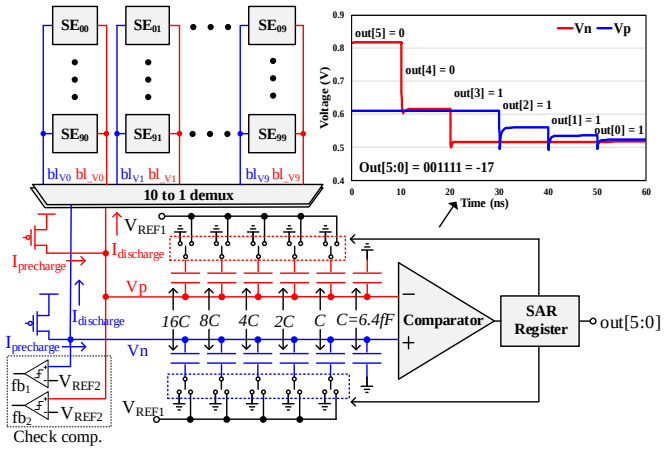


Fig. 6. Data-aware adaptive ADC, check comparators and its connections to bit-lines. (Control signals comes from the digital control logic)

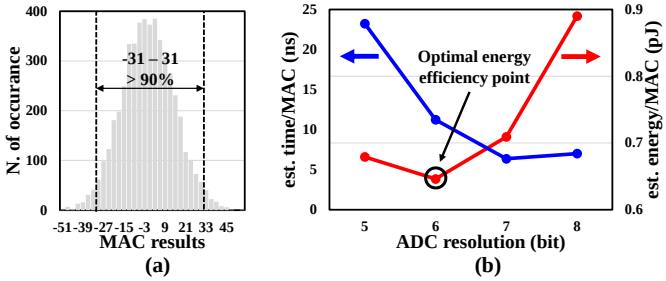


Fig. 7. (a) Simulated distribution of MAC results with uniformly distributed inputs and weights. (b) Simulated energy and time trade-off of ADC resolution with same clock frequency and unit capacitance.

A thermometer-encoded SE with 8 sequentially connected SCs enables multi-bit in-memory computation and update. The connection topology of SCs and BLs are shown in Fig. 3 (b). SC_0-SC_3 are connected to $BL_{V(SE)}$ and $BL_{H(SE)}$, and SC_4-SC_7 are connected to $BL\_{V(SE)}$ and $BL\_{H(SE)}$. $WL_{H(SC)}$ or $WL_{V(SC)}$ of all SCs are connected together, respectively. In each SC, either $BL_{V(SC)}$ ($BL_{H(SC)}$) or $BL\_{V(SC)}$ ($BL\_{H(SC)}$) is chosen to represent weight bits. The thermometer encoding of SE is shown in Fig. 3 (d), the number of '0's on $BL_{V(SE)}$ (b_0-b_3) represents the value with negative magnitude and number of '0's on $BL\_{V(SE)}$ (b_4-b_7) represents the value with positive magnitude. Value 0 is encoded by all '1's. Such a scheme facilitates charge accumulation on the BLs, by: (1) directly converting weight magnitude to BL discharge current represented by the '0's, and (2) allowing only one of the differential BLs to discharge (negative or positive number only discharge $BL_{V(SE)}$ ($BL_{H(SE)}$) or $BL\_{V(SE)}$ ($BL\_{H(SE)}$), respectively), which improves the dynamic range. The connection of the SE array is shown in Fig. 3 (c). All $BL_{V(SE)}$ or $BL\_{V(SE)}$ are connected to BL_V or BL_V of the SE array, and $BL_{H(SE)}$ or $BL\_{H(SE)}$ are connected to BL_H or BL_H for matrix transpose.

Besides the in-memory computing feature, the proposed structure also enables in-situ weight update with minimal additional hardware and control overhead. The Fig. 4 shows an example. On either ends of the SE, we have a pull-up and a pull-down transistor to provide '1's and '0's for rightward and leftward data movement. At first, the initial weight is -4 (b_0-b_7 = 00001111), and the corresponding data in each SC is shown in the Fig. 4 (a). A right shift pulse is generated by update DTC to change the weight. The pull up device turns on to insert 1 at the left side of the SE, and data in SCs will flip one by one as shown in simulated transient waves (Fig. 4 (d)). According to the pulse width, the propagation stops at certain place and all SCs returns to retention mode (Fig. 4 (b)). Therefore, the data in remaining SCs will not change, and the weight changes to 2 (b_0-b_7 = 11110011). Similarly, left propagation inserts 0 to the right side of the SE. The data will flip sequentially from right to left beginning at the place where last shift process stops (Fig. 4 (d)). According to the pulse width, left propagation will stop at certain place and weight is changed to -1 (b_0-b_7 = 11101111) (Fig. 4 (c)).

C. *Time/Charge-Domain Vector Multiplication*

The process of in-memory vector multiplication between input vector [x_0, x_1... x_9] and weight vector [w_0, w_1... w_9] is demonstrated in Fig. 5. The 2b input operands are modulated by time pulses generated by the WL DTC, and 3b weight operands are encoded by the discharging current of SEs. After the pre-charge of BL, each SE is sequentially accessed (depending on inference or learning, either WL_V or WL_H is used) with input-modulated time pulses. Thus, vector products are accumulated on capacitors of BLs.

D. *Capacitor-Based Data-Aware Adaptive ADC*

In CIM vector multiplication designs, ADC is a major power and area consumer. Therefore, the choice of ADC's architecture, resolution and control method are important to optimize the CIM performance and energy-efficiency. The proposed ADC topology and its connections to SE arrays are shown in Fig. 6. Since the operating principle of vector multiplication in proposed CIM is based on capacitive discharge, a relatively large BL capacitor is required to provide enough dynamic range. This is consistent with capacitor-based ADC which requires target internal capacitance to achieve better linearity and higher resolution. This motivates us to choose a charge-redistribution successive approximation (SAR) ADC [7] and share its differential capacitor arrays with the storage element array. Such a scheme not only increases the dynamic range, but also embeds the sampling process of ADC into the computing cycle. In the proposed design, the power and

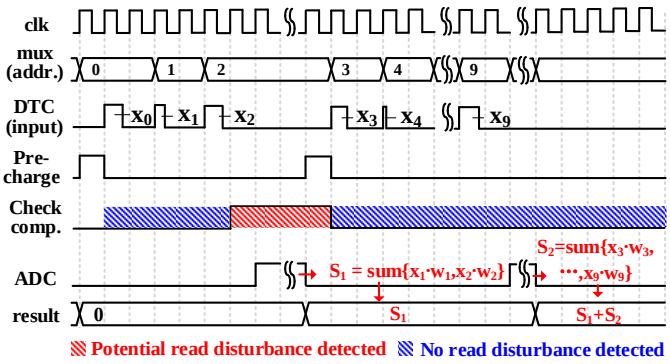
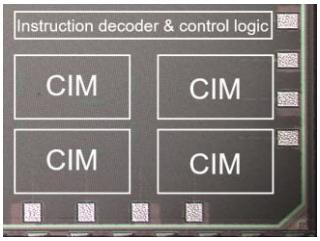


Fig. 8. Timing diagram of in memory vector multiplication process.



Technology	TSMC 65nm GP
Chip size	0.295mm^2
SC size	16.1um^2
SE size	410um^2
Package	QFN
Core supply	0.7-1V
IO supply	2.5V
Frequency	60 – 125MHz

Fig. 9. Chip shot and characteristics.

conversion time scale linearly with ADC's resolution. Design for the worst case (range from -120 to 120 with 2b input, 3b weight and 10 cells in a column/row) requires 8b resolution. However, an examination of the distribution of the MAC magnitude with uniformly distributed inputs and weights, reveals that most results are within 6b range (-32 to 31) as is shown in Fig. 7 (a). Simulated energy and time across 5-8b ADC resolution with same clock frequency and unit capacitance are shown in Fig. 7 (b). Therefore, we have chosen 6b for its best average energy efficiency. Since ADC can only handle numbers smaller than 32, if current sum of product is over 20, ADC overflow and read disturb may occur during next storage element discharging because the maximum product of one storage element is 12 (3(input) x 4(weight)). Therefore, two check comparators (BL and BL_) are used for detecting potential ADC overflow and read disturbance. To support 8b output resolution, we adopt an adaptive data conversion scheme to save energy and conversion time. An example of it is illustrated in Fig. 8. After the pre-charge, the WL DTC accesses each SE sequentially (depending on inference or learning, either WL_V or WL_H is accessed). The control logic checks comparator values after each WL access. In most cases, the vector multiplication is completed before ADC conversion. However, when the intermediate sum of products may cause ADC overflow and read disturb on SCs, check comparators will detect it. Then the control logic stalls the DTC and turns on the ADC to convert the BL voltage to its digital equivalent. After that, BLs are re-charged, and process continues to the remaining partial products in the vector multiplication. The converted numbers are accumulated in the digital domain to achieve full output resolution (8b).

III. Measurement Results and Comparison

The chip die-shot and characteristics are shown in Fig. 9. The chip has been fabricated in 65nm CMOS technology and total

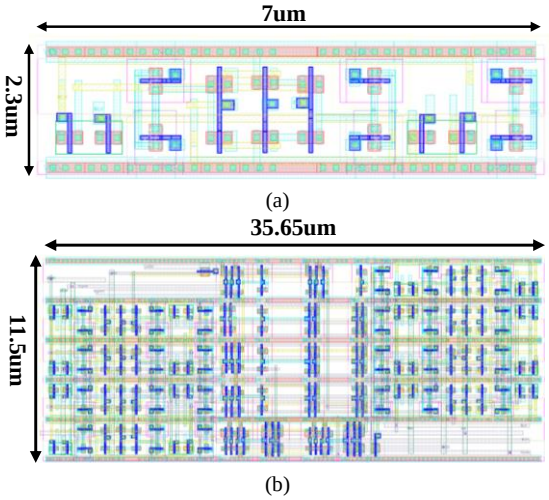


Fig. 10. (a) Storage Cell layout. (b) Storage element layout.

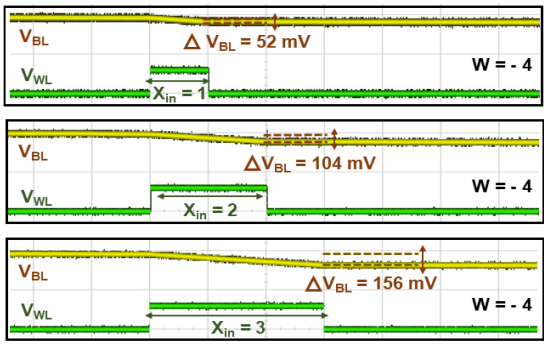


Fig. 11. Oscilloscope capture of BL voltage with varying WL DTC pulse accesses.

die area is 0.295mm^2. In the die region, we have 4 CIM blocks together with a controller. The SC and SE occupy the area of 16.1um^2 and 410um^2, respectively (Fig. 10). The proposed design embeds computing within the memory, thereby increasing energy-efficiency at the expense of silicon area per storage cell. This is well suited for our current application as an on-chip controller that fine-tune control knobs for edge device at target resolution and high energy-efficiency. An experimental example of time/charge-based multiplication is demonstrated in oscilloscope capture in Fig. 11. With varying DTC pulses (inputs operands), the voltage drop on the BL also scales proportionally. The measure of the nonlinearity which is the error between the CIM output and ideal results for randomly generated input numbers is shown in Fig. 12. We observe that CIM results are close to ideal computation results, with a measured average error of 0.6 and a max error of 3. At the same time, we observe that 90% of vector product magnitudes are distributes within the 6b ADC range (-32 to 31). The major sources of error come from the ADC's nonlinearity, DTC's nonlinearity, and the cell leakage current. However, we note that the data encoding, differential BL discharging, and conversion scheme make the design resilient to such errors. The energy efficiency in terms of energy per MAC and per update across various supply voltages is shown in Fig. 13. We measure 125MHz, 0.735pJ/MAC and 0.41pJ/update at 1V. As the supply voltage decreases, the energy/MAC also decreases. However, the error in computation increases because a lower

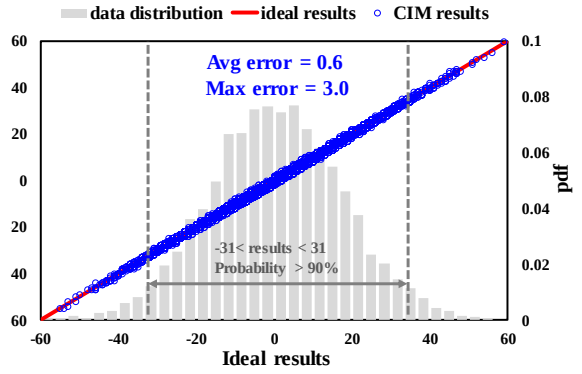


Fig. 12. CIM computing error with uniformly distributed inputs and weights.

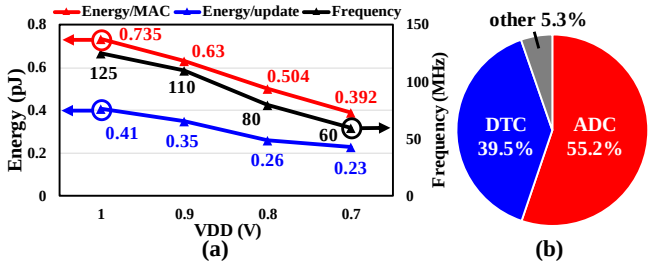


Fig. 13. (a) Measured frequency, energy efficiency of MAC and weight update, and (b) energy breakdown of MAC.

supply voltage degrades the linearity of the ADC. The measured frequency range is 60-125MHz (limited by the conversion speed of the ADC) and meets the target specification. The pie chart on the right of Fig. 13 shows total energy breakdown. DTC and ADC takes 39.6% and 55.2% energy consumption, respectively. Even after optimization, the ADC is still the major power consumer in CIM architecture, which in turn justifies the necessity to adopt the proposed adaptive data conversion scheme.

The proposed technique has been benchmarked against iso-resolution fully digital design (simulated) and state-of-the-art architectures (Table. I). Prior work that feature low-energy and low-complexity operation suitable for micro-controllers and light-weight machine learning have been selected. Compared to the state-of-the-art, we achieve high output resolution (8b). Moreover, the in-memory transposed matrix multiplication and in-situ weight update have been demonstrated. Finally, due to the optimized data conversion scheme, we achieve competitive energy efficiency of 0.735pJ/MAC and 0.41pJ/update. These measured numbers demonstrate 1.77x and 4.63x improvement respectively when compared to a simulated iso-resolution baseline fully digital design.

IV. Conclusion

In this work, we present a thermometer-encoded, time/charge-domain CIM macro for on-chip neural network-based controller. It achieves high energy efficiency by reducing memory access with the optimal data conversion scheme. We achieve competitive figures of merit and energy efficiency against state-of-the-art.

References

[1] N. Cao, B. Chatterjee, M. Gong, M. Chang, S. Sen, A. Raychowdhury, "A 65nm Image Sensor SoC Supporting Multiple DNN Models and Real-Time Computation-Communication Trade-off via Actor-Critical Neuro-Controller", *in Proc. IEEE Symp. on VLSI Technol. and Circuits*, 2020.

[2] Edward H. Lee, S. Simon Wong, "A 2.5GHz 7.7TOPS/W Switched-Capacitor Matrix Multiplier with Co-designed Local Memory in 40nm", *in Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, 2016, pp. 418–419.

[3] Thomas Chen, Jacob Botimer, Teyuh Chou, and Zhengya Zhang, "An SRAM-Based Accelerator for Solving Partial Differential Equations", *in Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, 2019.

[4] Avishek Biswas, and Anantha P. Chandrakasan, "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Powe Convolutional Neural Networks", *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217-230, Jan. 2019.

[5] Jintao Zhang, Zhuo Wang, and Naveen 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.

[6] Win-San Khwa, Jia-Jing Chen, Jia-Fang Li, et.al, "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 Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, 2018, pp. 496-497.

[7] Chun-Cheng Liu, Soon-Jyh Chang, Guan-Ying Huang, and Ying-Zu Lin, "A 10-bit 50-MS/s SAR ADC With a Monotonic Capacitor Switching Procedure", *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 731-740, Apr. 2010.

[8] Sujan Kumar Gonugondla, Mingu Kang, Naresh Shanbhag, "A 42pJ/Decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier with On-Chip Training", *in Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, 2019, pp. 490-491.

[9] Kayode Sanni, Tomas Figliolia, Gaspar Tognetti, Philippe Pouliquen, and Andreas Andreou, "A Charge-Based Architecture for Energy-Efficient Vector-Vector Multiplication in 65nm CMOS", *in Proc. IEEE Int. Symp. on Circuits and Syst. (ISCAS)*, 2018.

Table I. Comprison with Fully Digital Architecture and State-of-The-Art

	This work	Fully Digital (simulated)	[2]	[3]	[8]	[9]
Technology	**65nm**	65nm	40nm	180nm	65nm	65nm
Application	**On-chip neural controller**	/	Matrix multiplier	PDE solver	SVM	Vector multiplier
Supply voltage	**0.7-1V**	0.4V-1V	1.1V	1.8V	0.925V	1.2V
Frequency	**60 – 125MHz**	1000MHz	2500MHz	/	/	2.5MHZ
Input bit x weight bit	**2bx3b**	2bx3b	3bx6b	5bx5b	8bx8b	6bx6b
Output bit	**8b**	8b	6b	5b	4b	5b
In memory learning	**Yes**	No	No	No	No	No
In memory update (energy)	**Yes (0.41pJ)**	No (1.9pJ)	No	No	No	No
Energy efficiency	**1.36 TOPs/W**	0.77 TOPs/W	7.7 TOPs/W	0.857 TOPs/W	3.125 TOPs/W	0.284 TOPs/W
MAC energy	**0.735 pJ**	1.3 pJ	0.13* pJ	1.167* pJ	0.32 pJ	3.5 pJ

*Assuming one MAC is one operation.