A PVT Robust Analog Compute-In-Memory AI Accelerator with Integrated Activation Functions

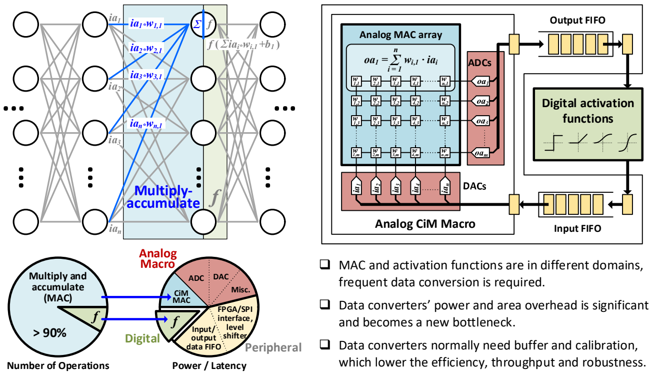
Hechen Wang, *Senior Member*, *IEEE*, Renzhi Liu, *Senior Member*, *IEEE*, Richard Dorrance, *Senior Member*, *IEEE*,
Deepak Dasalukunte, *Senior Member*, *IEEE*, Niranjan Mylarappa Gowda, Brent Carlton, *Member, IEEE*

***Abstract*— Most Analog Compute-in-Memory (ACiM) works only focus on the multiple-accumulate (MAC) operation while neglecting the activation function (AF) in the digital domain. The frequent data conversion greatly reduces the benefits obtained by analog computing. This paper proposes an efficient 8-bit in-memory MAC with hybrid capacitor ladders. Then a sparsity-aware R-2R DAC and an embedded SAR-ADC that reuses the capacitor ladders in the MAC are introduced to reduce the conversion overhead. Two on-chip AF schemes are included to further improve efficiency. Finally, differential signal path offers 1st order PVT cancellation that improves computing accuracy and reduces the need for calibration.**

***Index Terms* — AI, analog computing, charge domain computing, CMOS, compute-in-memory (CiM), machine learning accelerator, mixed-signal, multiply-accumulate operation (MAC), neural networks, PVT, static random-access memory (SRAM).**

I. Introduction

In the modern era of artificial intelligence (AI), machine learning (ML) and neural networks (NN) have seamlessly integrated into our daily lives. However, as AI becomes more embedded in our day-to-day activities, its energy consumption has escalated, turning into a considerable concern. The massive model training and daily inference tasks have become energy-intensive, contributing to substantial electricity usage that not only has a high economic impact but also exacerbates the environmental footprint. In response to this challenge, Analog Computing-in-Memory (ACiM) using highly efficient unconventional computation mechanisms has emerged as a promising solution to address the power consumption issues associated with AI [1], [2]. Analog computing, which previously faced hurdles due to its limited precision and resulting unacceptable accuracy degradation, is gaining traction gradually. As algorithms evolve to operate with lower precision to manage the ever-growing memory needs [3], they align well with today's ACiM capabilities, which efficiently support 8-bit precision for mainstream applications [4], [5]. However, some major obstacles remain for the ACiM, preventing it from being considered as a possibility for any high-volume product. Most existing ACiMs only focus on the MAC operation while neglecting the activation function (AF) and relegating it to the digital domain, as illustrated in Fig. 1. Although AF accounts for less than 10% of the total number of operations in NNs, the frequent data conversion between analog and digital domains, along with the data movement overhead, greatly reduces or even neutralizes the benefits obtained by analog computing and becomes a new bottleneck. Moreover, the need for calibration, potentially slowing down throughput and noticeable to the end-user, is another deterrent for those exploring alternatives to existing digital solutions.



❑ MAC and activation functions are in different domains, frequent data conversion is required.

❑ Data converters' power and area overhead is significant and becomes a new bottleneck.

❑ Data converters normally need buffer and calibration, which lower the efficiency, throughput and robustness.

Fig. 1. Challenges in the analog-CiM scheme.

This paper presents a multibit SRAM-based charge domain ACiM prototype, providing several techniques to address those issues. Section II covers the system architecture of the ACiM macro and design details, including efficient signed 8-bit in-memory MAC with hybrid differential capacitor ladders, sparsity-aware R-2R DAC, embedded SAR-ADC, and two on-chip AF schemes: 1) SAR-ADC LSB skipping based ReLU; 2) analog buffer based tanh, which shows the potential to bypass data converters. The measurement results are given in Section III and followed by the conclusions drawn in Section IV.

II. Architecture and Circuit Implementation

Figure 2 shows the top-level architecture of the proposed SRAM ACiM macro, consisting of 16 Digital Output (DO-) ACiM cores, two Analog Output (AO-) ACiM cores, and additional testing peripherals (shared SRAM pool, I/O interface, clocking, power delivery, etc.). The DO-ACiM contains 64 8-bit DACs, 16 kb SRAM cells, 256 8-bit MAC units, and four 8-bit ADCs. Each cycle, the DO-ACiM provides either four linear 8-bit digital outputs or four LSB skip-based ReLU AF results. The converted results are collected by the digital AXI/AHB BUS and sent to either the FPGA/SPI interface or the shared SRAM pool. In the AO-ACiM core, DAC and MAC units are doubled, while ADCs are replaced by four analog tanh AF buffers to generate analog outputs, which are directly connected to the output pads. The overall chip achieves a maximum throughput of 2.5 TOPS.

Manuscript received xx xx, xxxx; revised xx xx xxxx; accepted xx xx xxxx. Date of publication xx xx xxxx; date of current version xx xx xxxx. xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx. (*Corresponding author: Hechen Wang.*)

H. Wang, R. Liu, R. Dorrance, D. Dasalukunte, N. M. Gowda, and B. Carlton are with Intel Labs, Intel Corporation, Hillsboro, OR 97124 USA (e-mail: hechen.wang@intel.com).

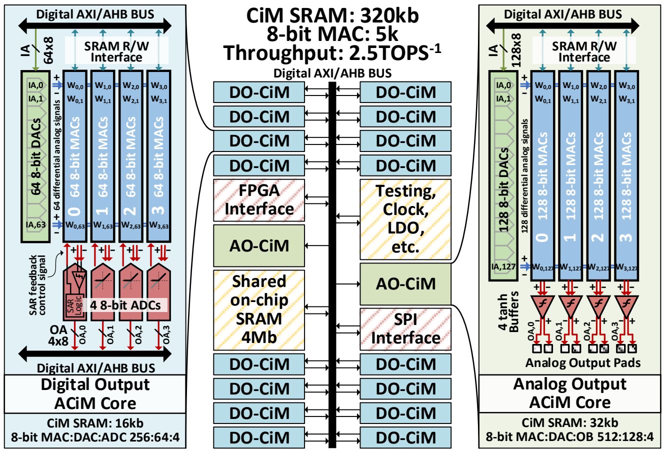


Fig. 2. ACiM macro diagram with DO-ACiM and AO-ACiM cores.

A. In-Memory Sign-Magnitude 8-bit MAC Unit

Figure 3 presents the 8-bit differential in-memory MAC unit based on a hybrid capacitor ladder for both DO-ACiM and AO-ACiM cores. Prior arts realized multibit MAC in the charge domain using binary [6] or C-2C capacitor ladders [7]. The C-2C ladder is area-efficient, but due to parasitic C_P, an extra compensation capacitor, $2C_P$, must be purposefully added to the cascaded 2C to maintain the ladder's desired 1:2 ratio. A dense in-memory layout can lead to a C_P close to or even larger than the chosen unit capacitor "C", resulting in the actual 2C on the ladder being 4-6 "C"s and leading to a more than 50% signal attenuation. By having the first two MSB capacitors binary weighted without additional $2C_P$, the attenuation is significantly alleviated, making a hybrid ladder a more efficient approach. Analog computing is inherently prone to generating errors because it is vulnerable to limited SNR and PVT variations. A differential scheme with two identical ladders serving as a pair of complementary positive and negative signal rails is employed to enable 1st-order error cancellation and enlarge the dynamic range, which improves the SNR and PVT robustness. Bit cells <6:0> are identical and contain a local in-memory "Compute & Conversion Logic" to modulate the ladder by selecting either the NN weights stored in SRAM during the MAC computing phase or the ADC feedback signal during the data conversion phase to control a pass-gate MUX in each capacitor branch on the ladder. Eight 9T SRAM cells are grouped for each bit, providing eight sets of weight to improve the in-memory weight storage volume. To isolate the original global WLs and BLs, local WLs and BLs are added to select the target weight bank in each cycle for each MAC unit. The MUX switches the signal between the DAC's input voltage ($V_{IN,P/N}$) and the differential rails' virtual ground, V_{REF} (equal to half V_{DD}). Then, in the computation phase, the product of $V_{IN,P/N}$ and NN weight is available at the output of the ladder ($V_{OUT,P/N}$). A butterfly switch alters the $V_{IN,P/N}$ and feeds to the differential rails based on the sign-bit (MSB<7>) to realize a sign-magnitude INT-8 format. With this arrangement, the error due to capacitor mismatch is aligned with typical NN weight distributions, which ensures better computation accuracy. The outputs of multiple MAC units are connected for charge summation and fed to the ADC.

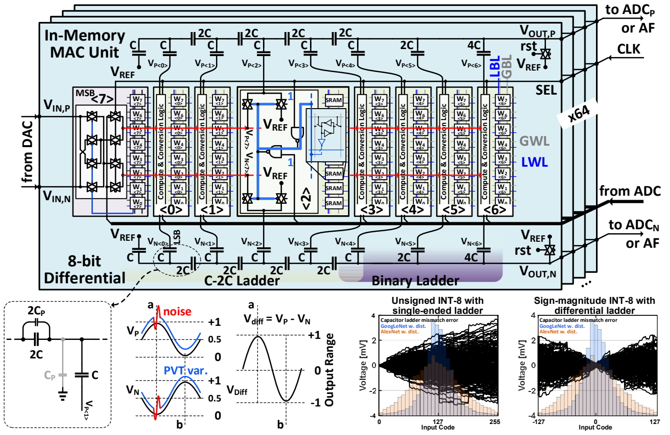


Fig. 3. In-memory sign-magnitude 8-bit MAC unit with differential hybrid capacitor ladders providing 1st order PVT, noise automatic cancellation.

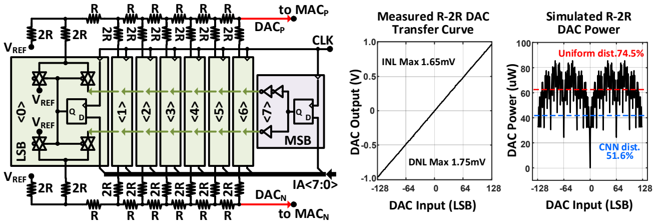


Fig. 4. Proposed sparsity-aware R-2R DAC with linearity and power plots.

B. Sparsity-Aware R-2R Resistor DAC

To mitigate the data conversion bottleneck, a sparsity-aware DAC is proposed as shown in Fig. 4. The DAC uses R-2R resistor ladders to provide enough fanout strength and lower the energy. A hybrid scheme is not necessary for the R-2R ladder due to limited parasitic resistance. Similar to the MAC, the DAC is also arranged differentially. Bit cells <6-0> are identical, and a flip-flop is used in each bit cell to achieve a higher throughput. MSB cell <7> alters the V_{DD} and GND to the differential rails based on the sign-bit of the digital input activation (DIA). Then, the MUXs on the ladder switch between the power rail (V_{DD} or GND) and the virtual ground V_{REF}, to generate the output voltages ($DAC_{P/N}$). Measurement shows that the R-2R DAC has good linearity, with a maximum DNL and INL of 1.75 mV and 1.65 mV in a 1-volt supply. Another feature of the R-2R DAC is its intrinsic sparsity awareness, as its power is code-dependent and is zero if the input is zero. As a result, if the input follows a uniform distribution, its average power is 74.5% of its peak power, and if it follows a distribution in NNs, the power can be reduced to only 50% of its peak power.

C. Embedded SAR-ADC

ADCs are the top power and area consumers in many ACiMs. This paper proposes an efficient embedded SAR-ADC that lets in-memory capacitor ladders sample and store the charge on the combined output node during MAC operation, and then the same ladders are reused for digitization. As presented in Fig. 5, the only add-on parts to build an ADC are a comparator and SAR logic block. Active signal buffers, commonly used between the CiM output and the ADC to enhance the signal, lower the impact of the parasitics, and reduce the kickback noise introduced by the comparator, are omitted in this work, as the total capacitance is large enough to ignore those effects.

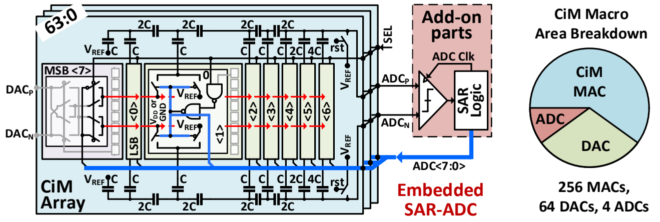


Fig. 5. Embedded SAR-ADC and CiM macro area breakdown.

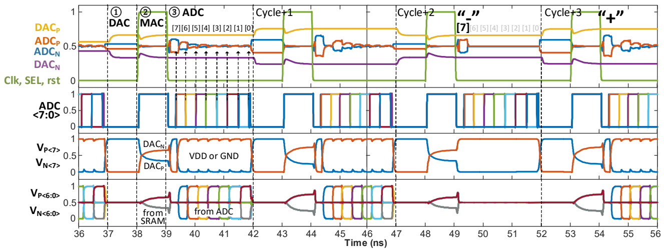


Fig. 6. Timing diagram of compute, conversion, and LSB-skipping ReLU.

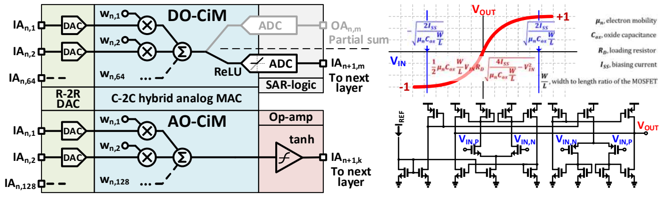
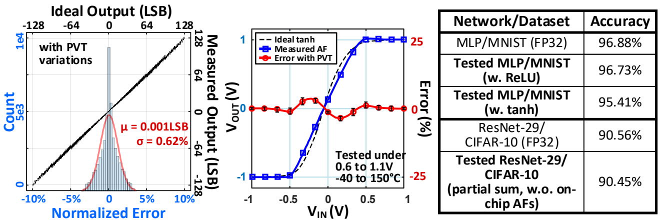


Fig. 7. Two types of on-chip AFs and differential pair tanh AF.

Furthermore, since the capacitors used for digitization are the same during MAC computing, errors generated during those two steps are automatically cancelled. Thus, no calibration or compensation is required. A simulated timing diagram is given in Fig. 6 to illustrate the operation detail. One complete cycle consists of three phases. Phase 1 starts each time after the previous ADC conversion is completed. DACs fetch new data from the input activation buffer and generate corresponding differential analog output DAC$_{P/N}$. In the meantime, the MAC unit selects one of the eight SRAM banks and connects to the Compute & Conversion Logic. Phase 2 starts at the clock-raising edge. The clock signal serves as the "rst" signal of the ladders' top plates and the "SEL" signal to decide whether the ladder is controlled by SRAM or the ADC feedback. Ladders' top plates are reset to virtual ground (V$_{REF}$). ADC feedback signal ADC<7:0> and "SEL" are forced to "1" (Fig. 3) to let the ladder be controlled by SRAM. The bottom plates on the ladder (V$_{P/N<6:0>}$) are connected to DAC$_{P/N}$ or V$_{REF}$. The polarity of DAC$_{P/N}$ is determined by the MSB of the data through V$_{P/N<7>}$. MAC computing is essentially an RC network charging process. Phase 3 starts when the charging is fully settled. "SEL" and "rst" are set to "0", and all 64 differential ladders are then controlled by the same set of SAR logic signals (ADC<7:0>). V$_{P/N<7>}$ in this phase is switched from DAC$_{P/N}$ to the power rails (V$_{DD}$ and GND). Thus, the ladder is selected between power and V$_{REF}$. The digitization result is available after eight ADC internal clock cycles.

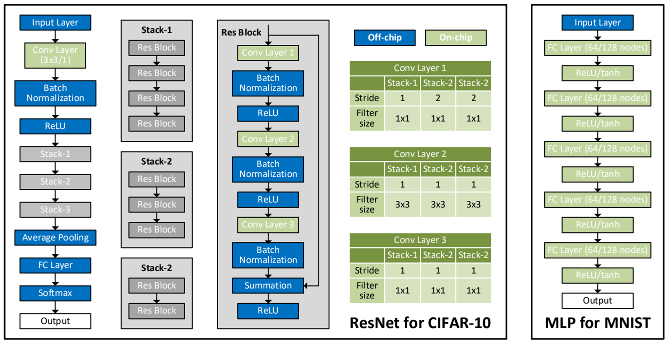
D. On-Chip Activation Functions

Figure 7 shows two options to further improve the efficiency: 1) on-chip AF to save data transfer energy and latency; 2) AF in the analog domain to remove the need for data conversions. In DO-ACiM cores, ReLU can be realized with SAR-ADC LSB skipping. Once the first SAR comparison is completed, the polarity is revealed ("Cycle 2" in Fig. 6). Then the conversion can be stopped if the result is negative or continued if positive. As the output is evenly distributed around zero, this scheme can lower the power by more than 40%. In the AO-ACiM core, an approximate hyperbolic tangent (tanh) AF is provided. The MOSFET differential pair follows a tanh-like square root transfer function (equation in Fig. 7).



Network/Dataset	Accuracy
MLP/MNIST (FP32)	96.88%
Tested MLP/MNIST (w. ReLU)	96.73%
Tested MLP/MNIST (w. tanh)	95.41%
ResNet-29/ CIFAR-10 (FP32)	90.56%
Tested ResNet-29/ CIFAR-10 (partial sum, w.o. on-chip AFs)	90.45%

Fig. 8. Measured MVM error, analog tanh AF, and NN inference results.



Conv Layer 1			
	Stack-1	Stack-2	Stack-2
Stride	1	2	2
Filter size	1x1	1x1	1x1

Conv Layer 2			
	Stack-1	Stack-2	Stack-2
Stride	1	1	1
Filter size	3x3	3x3	3x3

Conv Layer 3			
	Stack-1	Stack-2	Stack-2
Stride	1	1	1
Filter size	1x1	1x1	1x1

Fig. 9. Networks used during the ACiM chip accuracy evaluation.

III. Measurement Results

A test chip was fabricated in Intel-16 CMOS technology with the proposed ACiM macro implemented using foundry-provided SRAM collateral. Measurement results are compared with other state-of-the-art in-memory computing works listed in Table I. The total ACiM core area is 0.55 mm^2. The operation frequency is in a range of 50-240 MHz, when tuning the supply voltage from 0.6 to 1.1V. The measured maximum energy efficiency and area efficiency for 8b inputs and weights for DO-ACiM and AO-ACiM cores are 42.6 TOPS/W, 4.4 TOPS/mm^2 and 65.9 TOPS/W, 4.9 TOPS/mm2, respectively. A 55.0 TOPS/W can be achieved with the LSB skipping ReLU in DO-ACiM, and a 104.5 TOPS/W 7.0 TOPS/mm^2 can be achieved if bypassing the DACs in the AO-ACiM core, according to simulation. Analog computing always comes with errors. Figure 8 shows the measured AF curve versus the ideal tanh. The error in the ACiM core is also presented with a 50k-point matrix-vector-multiply (MVM) test. Its probability density function (PDF) and distribution indicate a standard deviation less than 0.62% over PVT variation. The impact of the errors has been evaluated by the NN inference. As shown in the table, the proposed ACiM has been evaluated with ReLU and tanh AFs in a 5-layer MLP on the MNIST dataset and ResNet-29 on the CIFAR-10 dataset, showing acceptable accuracy. Figure 9 gives the two networks used during the accuracy evaluation, with on-chip and off-chip operations marked in green and blue, respectively. 1) A 29-layer ResNet for CIFAR-10 dataset classification is used to evaluate the basic in-memory MAC operations. As the array size is smaller than the dimension in each convolutional layer (Conv), the in-memory MAC only performs a partial summation without using on-chip activation functions; 2) A MLP with one fully connected (FC) input layer and five 64-node/128-node FC hidden layers for on-chip ReLU AF and tanh AF, respectively.

Figure 10 shows the chip micrograph, µFCBGA package, the Shmoo plot, and the testing setup of the ACiM. Two sets of power supplies and clock generators are used to separate the analog and digital domains and provide more accurate power consumption readings to better calculate the efficiency numbers. The chip digital inputs and outputs are sent and received through a FPGA testing board. The analog outputs are captured from a high sample rate oscilloscope. Ten chips were tested in a testing chamber with temperature changing from -40 to 150, humidity from 0% to 98%, voltage from 0.6V to 1.1V, and additional aging test settings. No significant computation error was observed within that range, indicating good PVT robustness.

IV. Conclusions

An SRAM-based analog charge domain CiM macro is presented in Intel-16 process using hybrid capacitor ladders. Differential signal path offers 1st order PVT cancellation that improves computing accuracy and reduces the need for calibration. A sparsity-aware DAC and an embedded SAR-ADC are introduced to lower the data conversion overhead. Two AFs are included to further improve efficiency: 1) SAR-ADC LSB skipping based ReLU; 2) analog buffer based tanh.

References

[1] S. Ambrogio *et al.*, “An analog-AI chip for energy-efficient speech recognition and transcription,” Nature, vol. 620, pp. 768-775, Aug. 2023.

[2] H. Wang, “Analog chip paves the way for sustainable AI,” Nature, vol. 620, pp. 731-732, Aug. 2023.

[3] B. Keller *et al.*, “A 95.6-TOPS/W deep learning inference accelerator with per-vector scaled 4-bit quantization in 5 nm,” *IEEE J. Solid-State Circuits*, vol. 58, no. 4, pp. 1129-1141, Apr. 2023.

[4] H. Wang *et al.*, “A 32.2 TOPS/W SRAM compute-in-memory macro employing a linear 8-bit C-2C Ladder for charge domain computation in 22nm for edge inference,” in *Proc. IEEE Symp. VLSI Circuits* (*VLSI*), 2022, pp. 36-37.

[5] R. Dorrance *et al.*, “An energy-efficient Bayesian neural network accelerator with CiM and a time-interleaved Hadamard digital GRNG using 22-nm FinFET,” *IEEE J. Solid-State Circuits*, vol. 58, no. 10, pp. 2826-2838, Oct. 2023.

[6] S. -E. Hsieh *et al.*, “7.6 A 70.85-86.27 TOPS/W PVT-insensitive 8b word-wise ACIM with post-processing relaxation,” in *Proc. IEEE Int. Solid-State Circuits Conf.* (*ISSCC*), 2023, pp. 136-138.

[7] H. Wang *et al.*, “A charge domain SRAM compute-in-memory macro with C-2C ladder-based 8-bit MAC Unit in 22-nm process for edge inference,” *IEEE J. Solid-State Circuits*, vol. 58, no. 4, pp. 1037-1050, Apr. 2023.

[8] H. Fujiwara *et al.*, “A 3nm, 32.5 TOPS/W, 55.0 TOPS/mm^2 and 3.78 Mb/mm^2 fully-digital compute-in-memory macro supporting INT12 × INT12 with a parallel-MAC architecture and foundry 6T-SRAM bit cell,” in *Proc. IEEE Int. Solid-State Circuits Conf.* (*ISSCC*), 2024, pp. 572-574.

[9] J. -O. Seo *et al.*, “A 332.7 TOPS/W 5b variation-tolerant analog CNN processor featuring analog neuronal computation unit and analog memory,” in *Proc. IEEE Int. Solid-State Circuits Conf.* (*ISSCC*), 2022, pp. 258-260.

[10] H. Jiang *et al.*, “A 40nm analog-input ADC-free compute-in-memory RRAM macro with pulse-width modulation between sub-arrays,” in *Proc. IEEE Symp. VLSI Circuits* (*VLSI*), 2022, pp. 266-267.

[11] H. Wang *et al.*, “A PVT Robust 8-Bit Signed Analog Compute-In-Memory Accelerator with Integrated Activation Functions for AI Applications” in *Proc. IEEE Symp. VLSI Circuits* (*VLSI*), 2024.

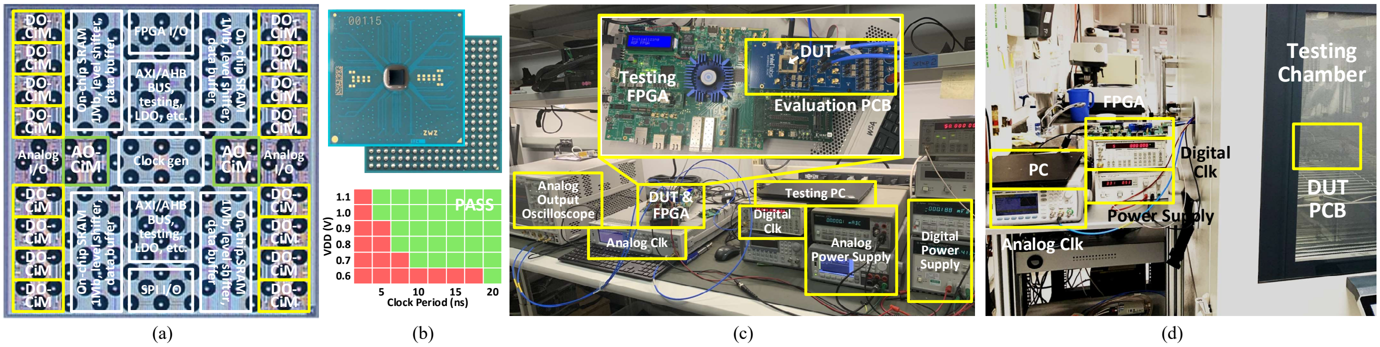


Fig. 10. (a) Testchip die photo, (b) µFCBGA package together with measured shmoo plot, (c) computation performance test setup, and (d) PVT test setup.

TABLE I PERFORMANCE SUMMARY AND COMPARISON

	ISSCC 2023 12 nm [6]	JSSC 2023 22 nm [7]	ISSCC 2024 3 nm [8]	ISSCC 2022 28 nm [9]	VLSI 2022 40 nm [10]	THIS WORK Intel-16 [11]			
						DO-ACiM		AO-ACiM	
MAC scheme	Charge/SRAM	Charge/SRAM	Digital/SRAM	Charge/6T-1C	Current/ReRAM	Charge/SRAM			
CiM core area	/	0.25 mm^2	0.0157 mm^2	0.63 mm^2 est.	0.62 mm^2	0.45 mm^2		0.10 mm^2	
CiM memory	128kb	128kb	60.75kb	30.6kb est.	64kb	256kb		64kb	
Clock	/	145-240 MHz	0.3-1.6 GHz	200 MHz	100 MHz	50-240 MHz			
Voltage	0.5-0.85 V	0.7-1.1 V	0.36-1.1 V	1.0 V	0.9	0.6-1.1 V			
I/O channels	1024 / 16	128 / 32	72 / 4	336 / 112	256 / 8	1024 / 64		256 / 8	
I/W/O bits	8 / 8 / 8	8 / 8 / 8	12 / 12 / -	Ana. / 2 / ana.	Ana. / 2 / ana.	8 / 8 / 8		8 / 8 / ana.	
TOPS	0.10	0.60-0.97	0.38-1.94	0.10	0.01	0.41-1.97		0.10-0.49	
Data conversion	Yes	Yes	No	No	No	Yes		No ADC	No
TOPS/W (8b)	71-86	16-32	52	130	7	**16-43**	**55**	**31-66**	**105**
TOPS/mm^2 (8b)	/	2.4-4.0	23.9-123.8	0.1	0.1	**0.9-4.4**		**1.0-4.9**	**7.0**
Compute error	0.41%	0.89%	No error	0.32%	4.1%	0.62% over PVT			
PVT calibration	Yes	-	No	-	-	Not required			
On-chip AF type	No	No	No	ReLU	ReLU	No	ReLU	tanh	
Network Dataset Accuracy	ResNet20 CIFAR-100 67.9%	MLP MNIST 98.14%	-	VGG-16 CIFAR-10 91.0%	VGG-8 CIFAR-10 89%	ResNet29 CIFAR-10 90.45%	MLP MNIST 96.73%	MLP MNIST 95.41%	