nature portfolio

https://doi.org/10.1038/s41928-022-00795-x

Supplementary information

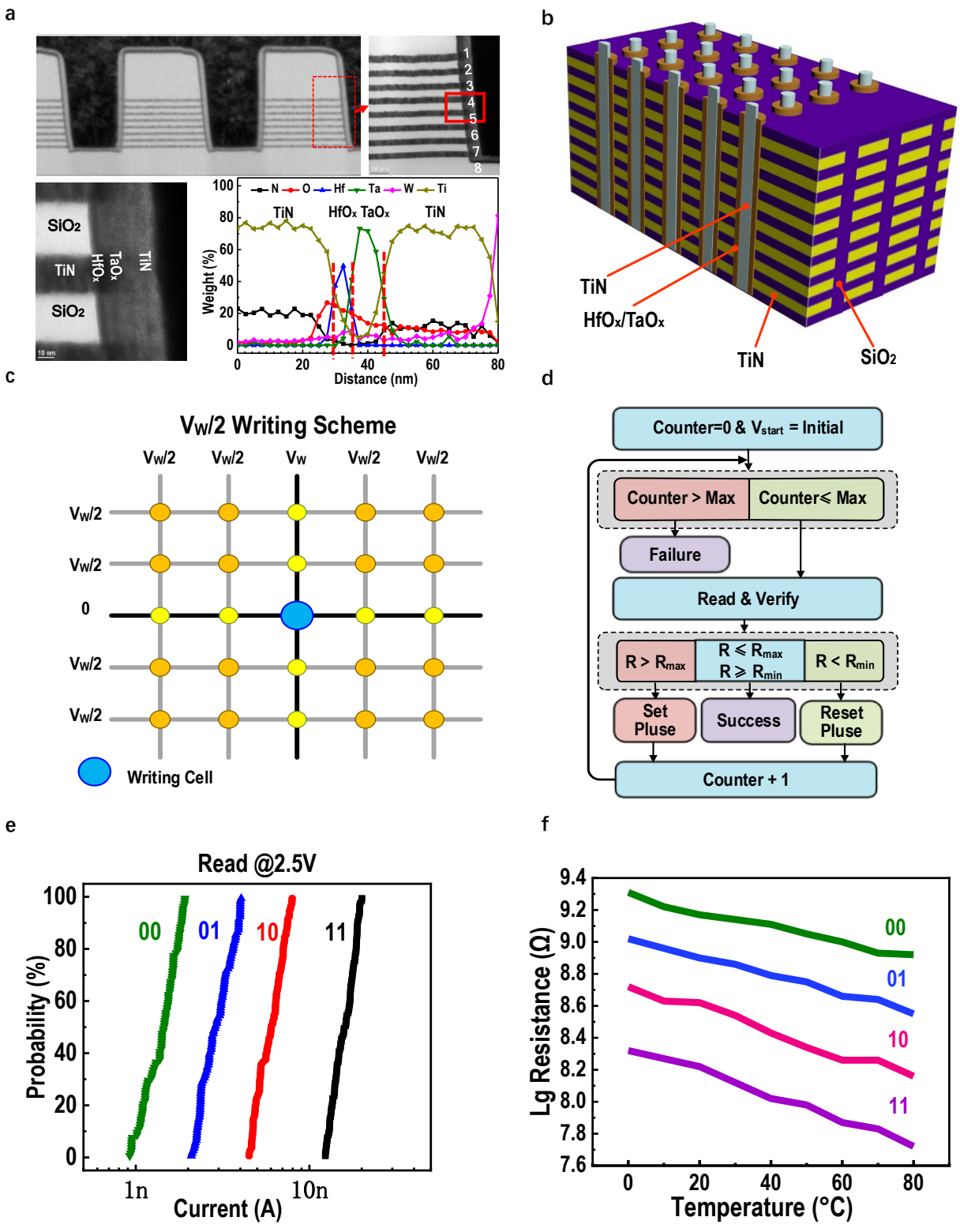
A computing-in-memory macro based on three-dimensional resistive random-access memory

In the format provided by the authors and unedited

A computing-in-memory macro based on three-dimensional resistive random-access memory

Qiang Huo[1,7], Yiming Yang[2,7], Yiming Wang[2], Dengyun Lei[3], Xiangqu Fu[1], Qirui Ren[1], Xiaoxin Xu[1], Qing Luo[1], Guozhong Xing[1], Chengying Chen[4], Xin Si[5], Hao Wu[1], Yiyang Yuan[1], Qiang Li[5], Xiaoran Li[2], Xinghua Wang*[2], Meng-Fan Chang[6], Feng Zhang*[1], Ming Liu[1]

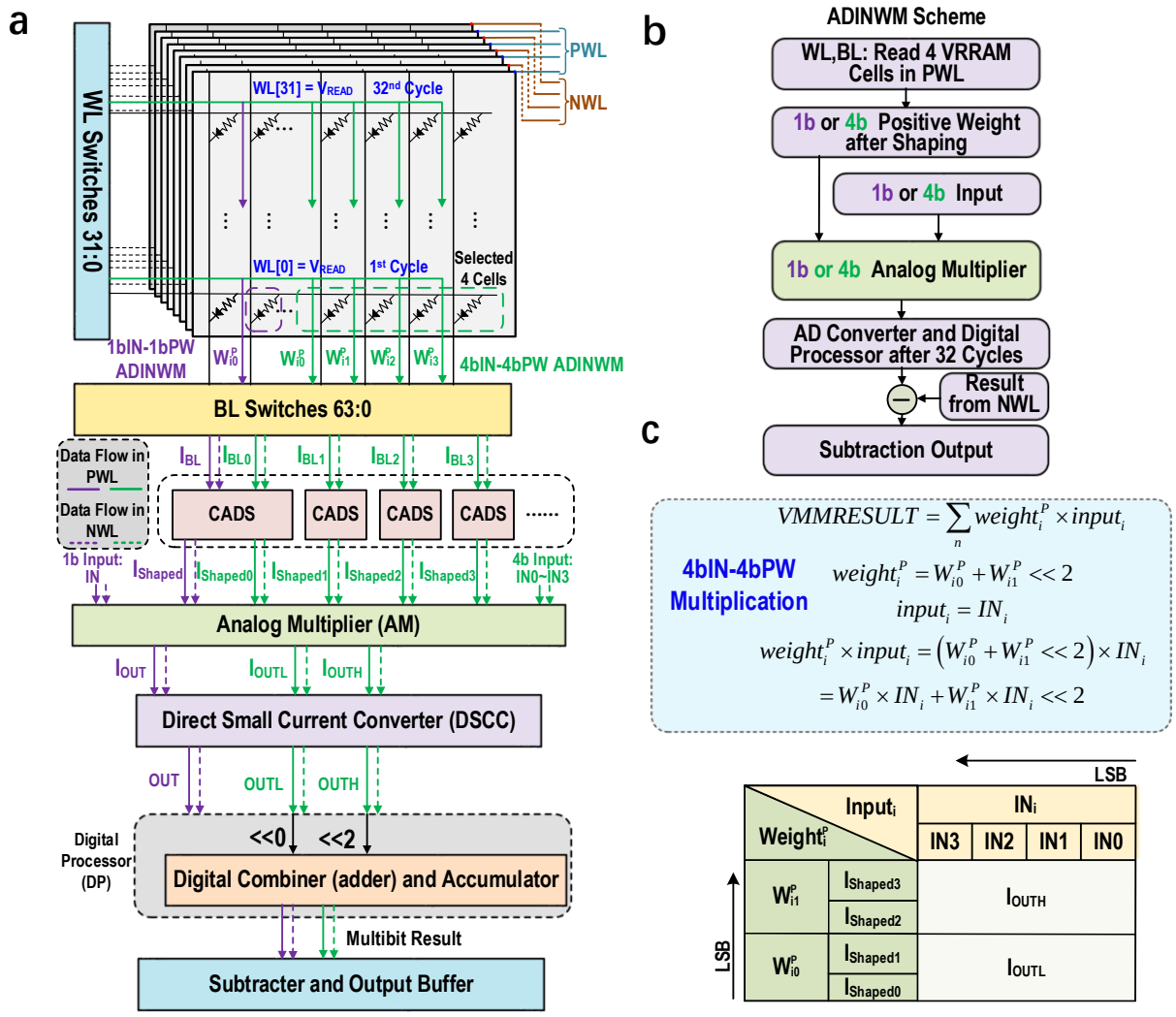
[1]Institute of Microelectronics of Chinese Academy of Sciences, Beijing, China. [2]Beijing Institute of Technology, Beijing, China. [3]School of Integrated Circuits, Guangdong University of Technology, Guangzhou, China. [4]Xiamen University of Technology, Xiamen, China. [5]University of Electronic Science and Technology of China, Chengdu, China. [6]National Tsing Hua University, Hsinchu, Taiwan. [7]These authors contributed equally. *e-mail: 89811@bit.edu.cn, zhangfeng_ime@ime.ac.cn.



Supplementary Figure 1. Description of 3D VRRAM. a, High-resolution transmission electron microscope (HRTEM) image and EDX line scan over TiN/HfO_2/TaO_2/TiN cross-section of the 3D VRRAM. **b,** Schematic of 3D VRRAM architecture. **c**, $V_W/2$ writing scheme. V_W refers to the write voltage. **d,** Flow of program-verify. **e,** Distribution of readout current I_{BL} of our multi-level self-selective (MLSS) 3D VRRAM. **f,** Drift of resistance with temperature.

Supplementary Note 1: The process flow of 3D VRRAM

The process scheme of this 3D VRRAM is shown in Fig. 1b, which is described in our previous work [S.1]. Multiple TiN(20 nm)/SiO_2 (20 nm) are deposited by PVD and PECVD, respectively. A Trench is dry etched down to the bottom SiO_2 layer to split the plane electrode (BL). After filled the trench by SiO_2, a 500 nm hole is etched down. Then, the HfO_2/TaO_x bilayer were deposited on the sidewall sequentially by ALD and sputtering. Where the HfO_2 served as the memory layer and the TaO_x served as the selective layer. The pillar electrode (WL) is extracted by sputtering the TiN/W to fill the hole. Each horizontal BL was opened by selective etching successively. The area of the memory cell was defined by the thickness of bottom electrode TiN and the perimeter of the hole. From the HRTEM image over the cross-section of the device shown in Fig. 1a, the 500 nm hole structure and 8 layers vertical memory cells can be observed clearly. EDX line scan verified the bilayer device structure and identified interface. 3D VRRAM is an interface type structure different from that of some 2D RRAM conductive filaments, which occurs on the entire device area and does not form a conductive channel with strong conductivity locally. The device has uniform distribution and the resistance can be continuously modulated because of the change of the barrier between the electrode and the dielectric material or the process of charge capture/release by defects in the material body [S.2-S.3]. Multi-level resistance states could be obtained by modulating the voltage pulses on this memory device, as shown in Fig. 1e. The flow of program-verify is shown in Fig. 1d, only when $R_{min}<R<R_{max}$, the write operation finishes with a success flag, otherwise fail. Fig. 1f mainly describes the drift of resistance with temperature.



$$VMMRESULT = \sum_n weight_i^P \times input_i$$

$$weight_i^P = W_{i0}^P + W_{i1}^P << 2$$

$$input_i = IN_i$$

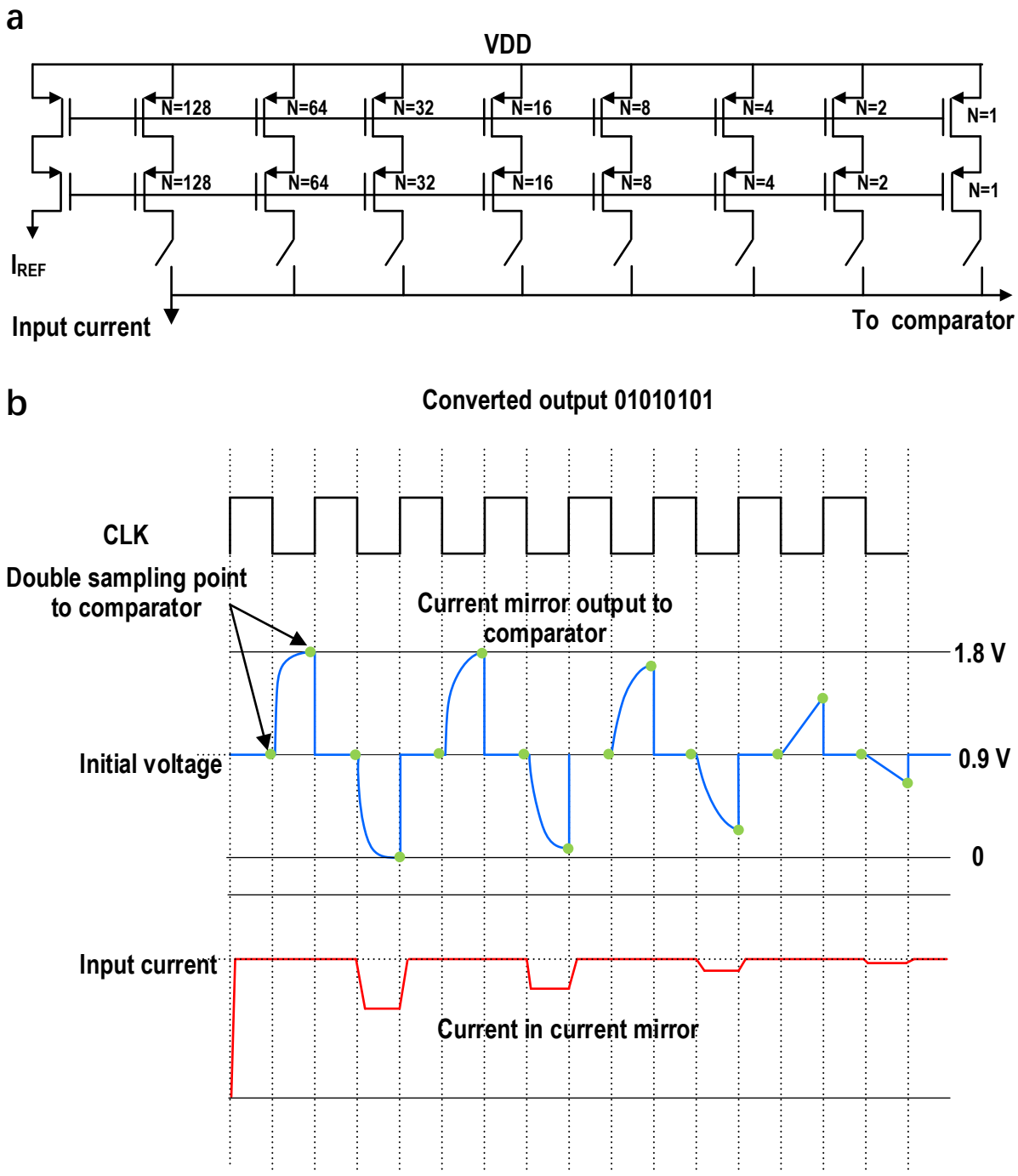
$$weight_i^P \times input_i = \left(W_{i0}^P + W_{i1}^P << 2\right) \times IN_i$$

$$= W_{i0}^P \times IN_i + W_{i1}^P \times IN_i << 2$$

Supplementary Figure 2. 3D VRRAM-based nvCIM macro using ADINWM scheme. a, Dataflow and architecture of anti-drift multibit analog IN-W multiplication (ADINWM) for 1bIN-1bPW (purple lines) and 4bIN-4bPW VMM operation (green lines). The dotted lines show the corresponding dataflow of 1bIN-1bNW and 4bIN-4bNW operation in NWL. **b,** Implementation framework of proposed ADINWM scheme for 1bIN-2bW and 4bIN-5bW operations. **c,** Formula and table of 4bIN-4bPW operation in PWL. 4bIN-4bPW operation is divided into two 4-2-bit multiplications (multiplying 2-bit input by 4-bit input) and the results correspond to two output currents (I_{OUTH} and I_{OUTL}). The currents are converted to 6 bit digital signal by direct-small-current-converter (DSCC) and then shift (OUTH shifts 2 bit to the left) and combine to complete the 4 bit multiplication operation for PW in PWL. All the multiplication results from 32 WLs are accumulated after 32 cycles in series by digital processor. The result makes difference with the result of 4bIN-4bNW operation in NWL to get the result of 4bIN-5bW.

Supplementary Note 2: Curretn digital-to-analog converter (DAC) in DSCC

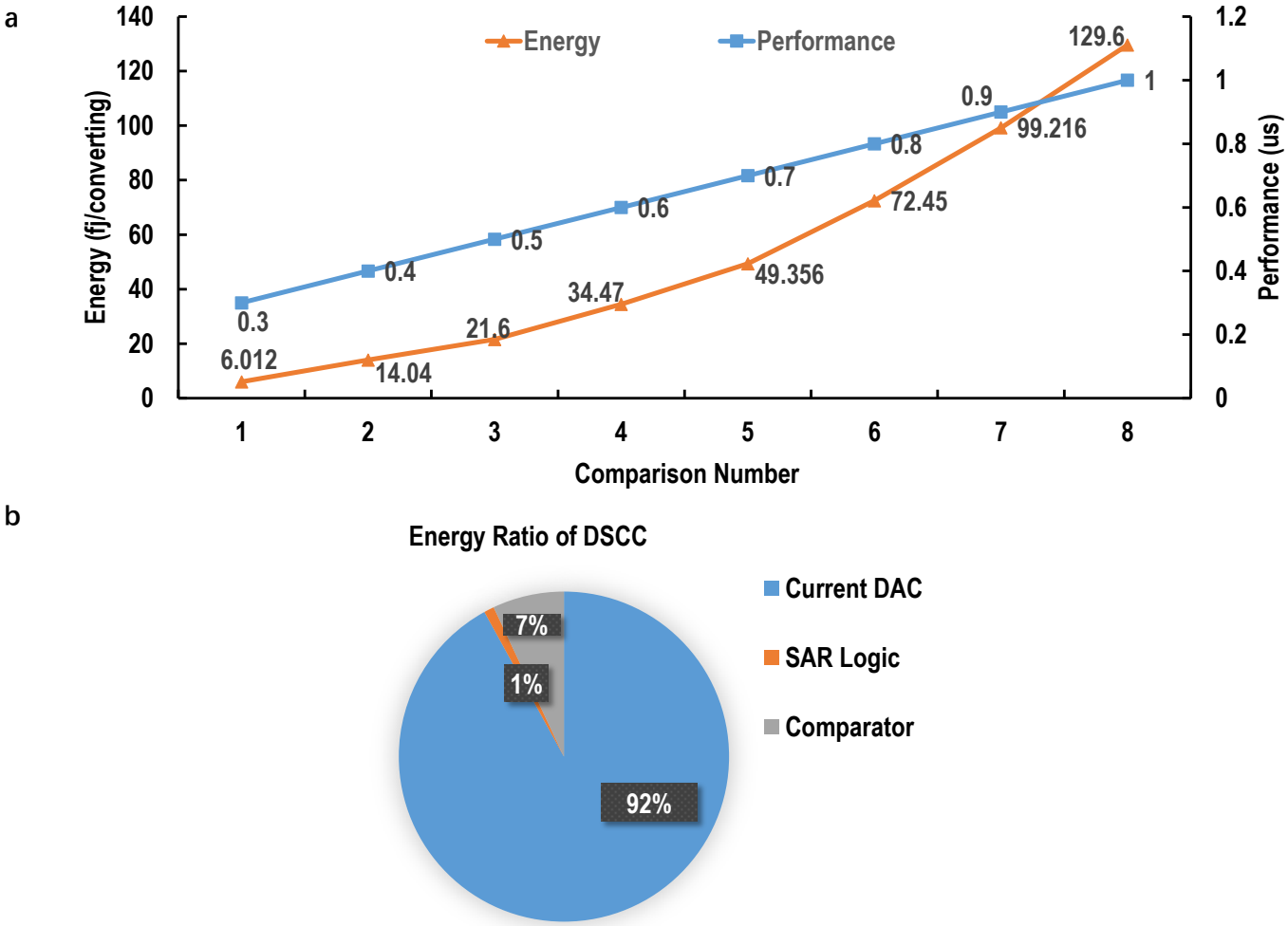
Supplementary Fig. 3(a) illustrates the schematic of current digital-to-analog converter in DSCC. DAC can convert the digital signal into analog current, which is used to compare with the input current to judge the approximation degree of the converted digital output. The current DAC consists of multiple current mirrors. The output to the comparator consists of the initial voltage and the divided voltage of pull-up DAC and pull-down current mirror of input current. Supplementary Fig. 3(b) demonstrates the waveform of the current mirror output to comparator and input current in current mirror when converted output is '01010101'. The DAC makes comparison for each beat and controls switches to adjust the output current. The difference between the input current and the DAC output current will cause the parasitic capacitance charging and discharging, thereby changing the output voltage. As the comparison bits reduces, the corresponding current difference and voltage fluctuation decreases. Double sampling means that in the first half of the clock cycle, an initial voltage of 0.9 V is applied and the gate capacitor of the comparator samples and holds it, which makes the complementary current mirror in the saturation region to reduce the voltage swing. In the second half of the cycle, the comparator samples again by adding the divider voltage. The initial voltage structure can reduce the latency of DSCC from 10 μs to 0.83 μs. The double-sampling circuit performs samplings and makes comparison twice in one clock cycle, which is more sensitive to voltage change, thereby ensuring the speed and accuracy of DSCC circuit.



Supplementary Figure 3. Internal analysis of DSCC. a, Schematic of current digital-to-analog converter (DAC). **b,** Waveform of the current mirror output to comparator and input current in current mirror when converted output is '01010101'.

Supplementary Note 3: Energy consumption and energy ratio of DSCC

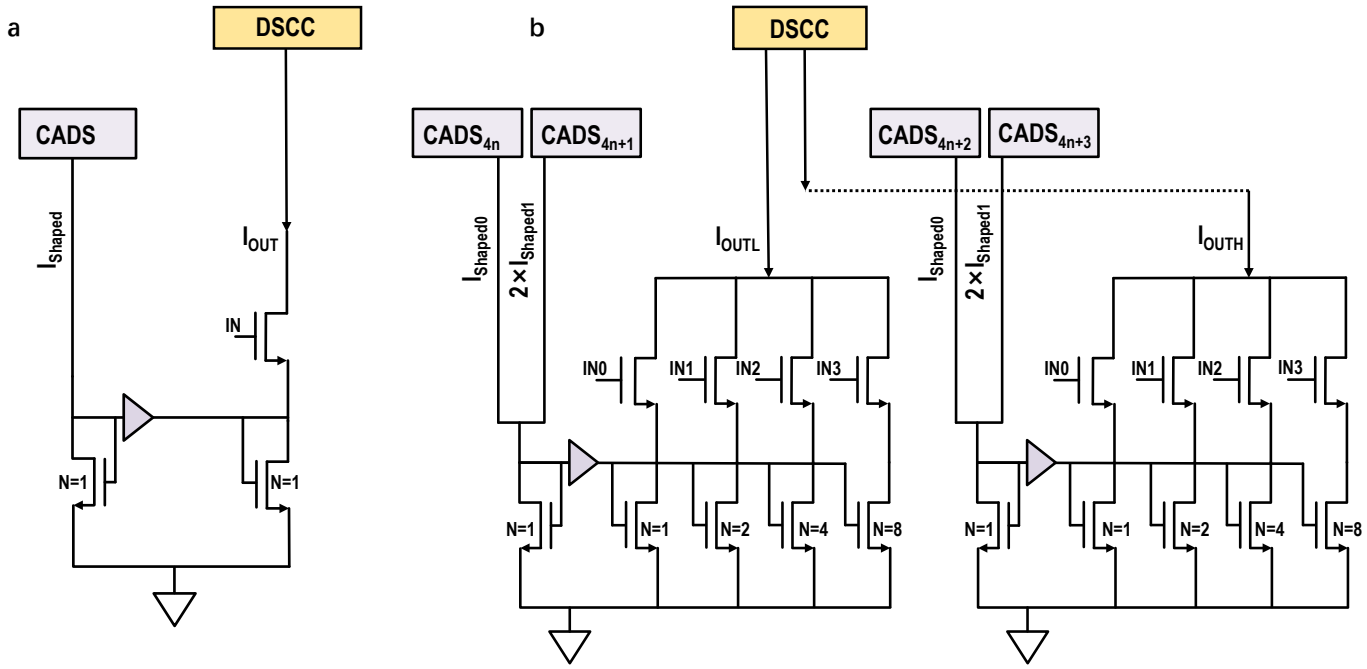
Supplementary Fig. 4(a) presents the energy consumption and performance of DSCC with the change of comparison number. The complete conversion is completed within 1 us. Supplementary Fig. 4(b) shows the energy ratio of DSCC. The current DAC and pull-down current mirror of input current form the complementary current mirrors, which occupy the main power consumption of DSCC. The power consumption of complementary current mirrors is determined by the input current in pull-down current mirror. Therefore, when the input current is small enough, the power consumption of DSCC will be greatly reduced. Because there are a lot of sparse matrix calculation in CNN, it always guarantees relatively small input current. At the same time, the resistance of the 3D VRRAM is high, and the readout current is further reduced. Therefore, the DSCC is conducive to the reduction of the power consumption of the whole system.



Supplementary Figure 4. Quantitative analysis of DSCC. a, Energy consumption and performance of DSCC as a function of comparison number which refers to the resolution of DSCC. **b,** Energy ratio of DSCC.

Supplementary Note 4: Analog multiplier (AM) for 1bIN-2bW and 4-2-bit AM operation

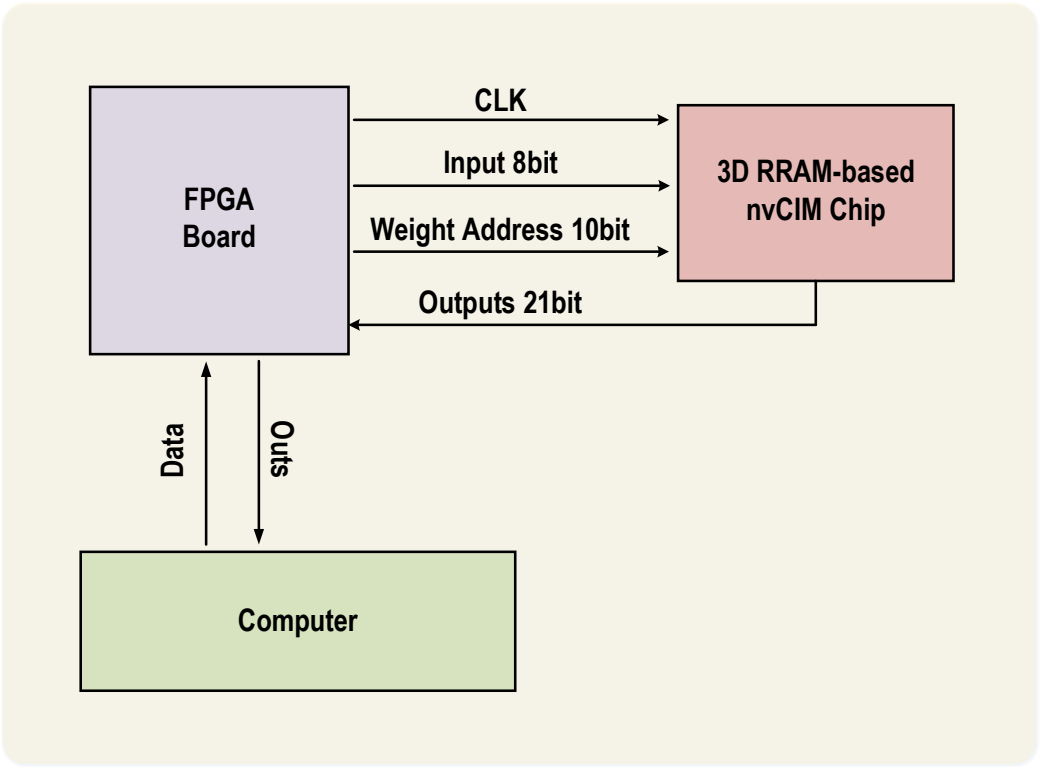
Supplementary Fig. 5(a) illustrates the schematic of 1 bit AM composed of current mirror. For 1bIN-1bPW operation, as shown in purple lines in Supplementary Fig. 2(a), the shaped current I_{Shaped} is input to AM and the current is multiplied by the input IN through current mirror to get the output currents I_{OUT}. The current is converted to 1 bit digital signal by DSCC and complete the 1 bit multiplication operation for input and PW in PWL. All the multiplication results from 32 WLs are accumulated after 32 cycles in series by digital processor. Then, make difference with the result of 1bIN-1bNW operation in NWL to get the final 1bIN-2bW VMM result. Supplementary Fig. 5(b) shows the schematic of 4-2-bit AM group for 4bIN-5bW operation. For 4bIN-4bPW operation, as shown in green lines in Supplementary Fig. 2(a), four shaped currents ($I_{Shaped0}$~$I_{Shaped3}$) are input to AM. Then these currents are multiplied by the input IN0~IN3 through current doublers (CDs) to get the output currents I_{OUTL} and I_{OUTH}. The current doubler is a current mirror variant that controls the output current by adjusting the size of the output transistor. The two currents are converted to 6 bit digital signal by DSCC and complete the 4-2-bit multiplication operation for input and PW in PWL. After 32 cycles, all the multiplication results from 32 WLs are accumulated in series by digital processor. The difference between the result of 4bIN-4bNW operation in NWL and the accumulation results is the final 4bIN-5bW VMM result. Gate precharge switch follower (GPSF) has been added to reduce T_{AC} due to the large parasitic capacitance of the output gate of current doubler.



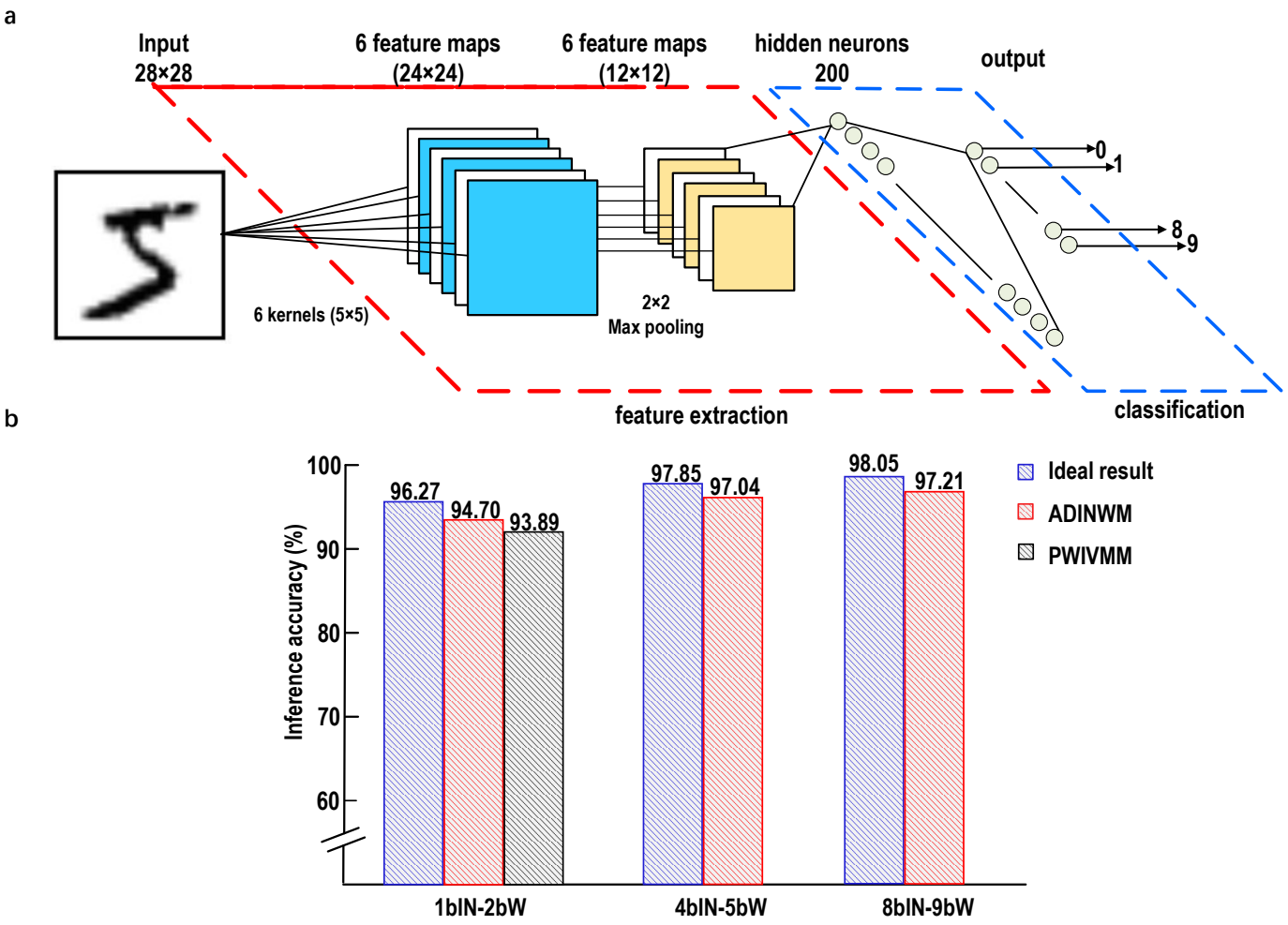
Supplementary Figure 5. Analog multiplier (AM). a, Schematic of 1 bit AM for 1bIN-2bW operation. **b,** Schematic of 4-2-bit AM group for 4bIN-5bW operation. N is the normalized width of CMOS transistor.

Supplementary Note 5: Influence of AM on system latency

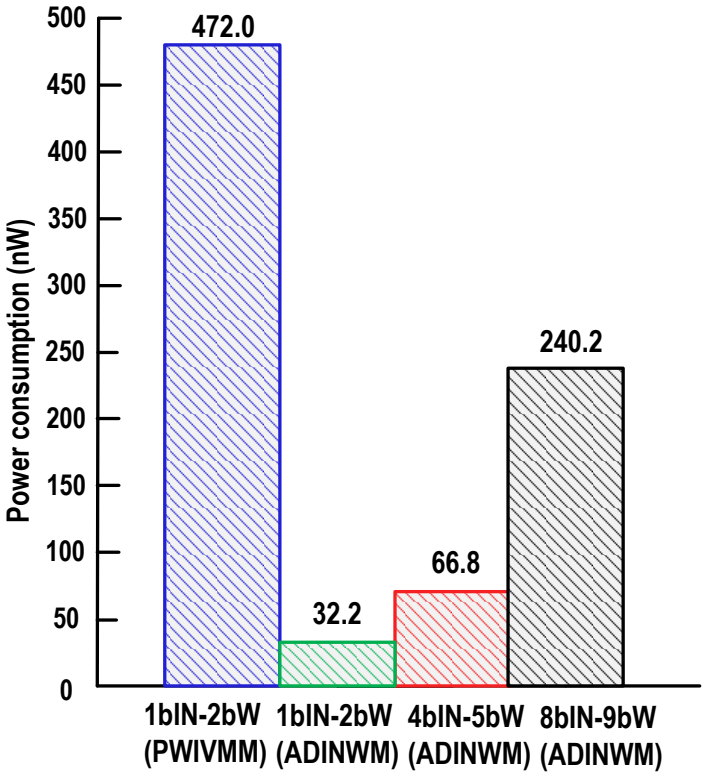
In our design, current-amplitude-discrete-shaping (CADS) circuit, DSCC and AM work in the form of pipeline, so the highest frequency of our nvCIM system is determined by the main modules. We find that the AM design has the greatest impact on the latency of the whole system. The latency of AM is reduced from 2.2 us without gate precharge switch follower structure to 1 us. The latency of DSCC is 10 us without using the initial voltage scheme, and the improved latency is 0.83 us. Therefore, AM determines the latency of the whole system. The two schemes reduce the latency from 10 us to 1 us. In other words, the frequency increases from 100 kHz to 1 MHz.



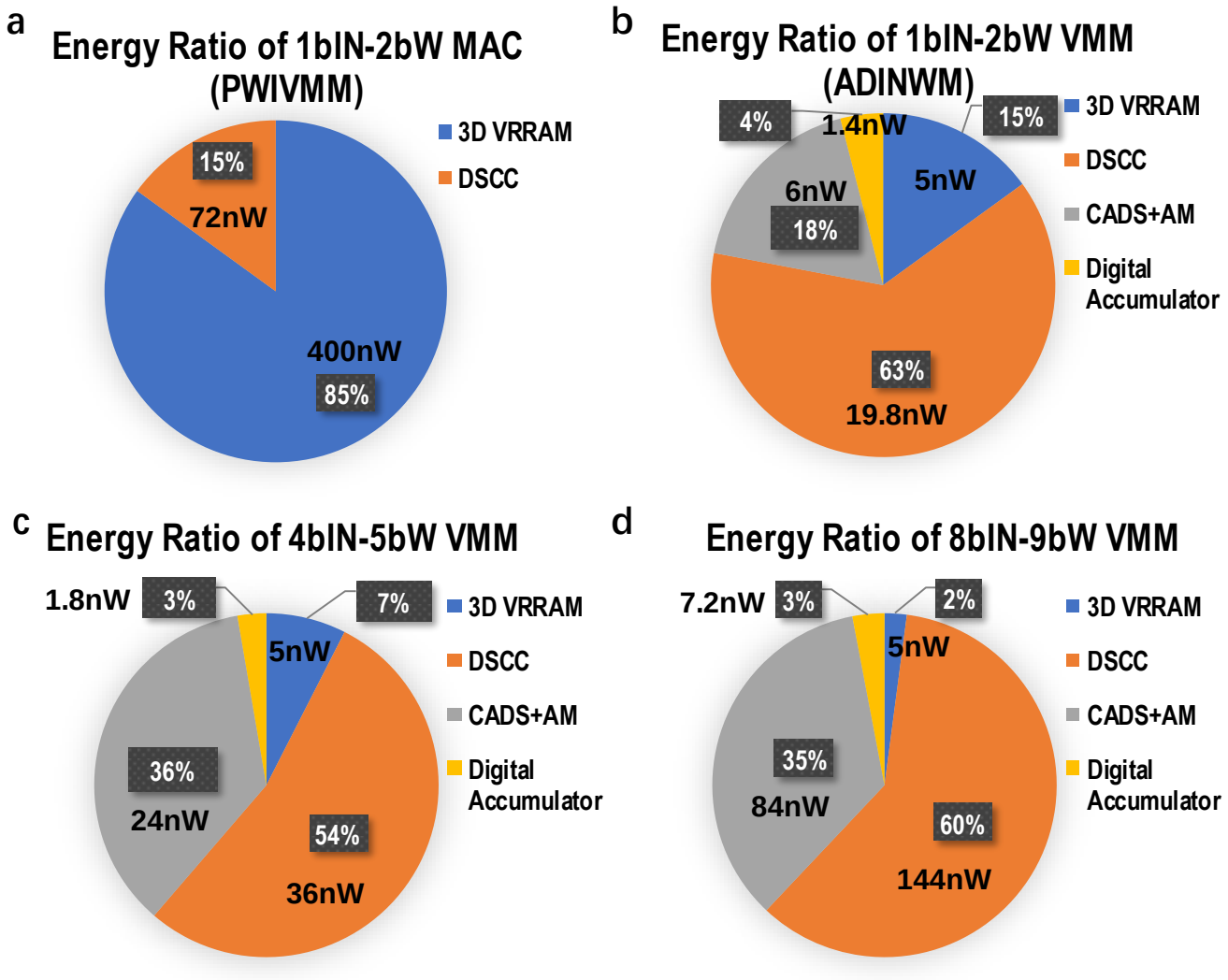
Supplementary Figure 6. Hardware automatic test platform with 3D VRRAM-based nvCIM macro under FPGA control. The convolution kernel is written into RRAM in advance. The computer sends the 8-bit data to the field-programmable gate array (FPGA). In the meantime, the address of the corresponding convolution core stored in the RRAM also needs to be sent to the FPGA. Then the FPGA sends the data and address of the corresponding weight to the chip, while to provide the clock. The chip performs multiplication and addition. After a convolution is completed, the chip outputs the result, which is read by the FPGA and sent back to the computer. The computer puts it on the result matrix, encodes and saves it, and further calculate the final results. Therefore, the code on the FPGA needs to control the pin, provide the clock, read and write data to the computer, which conduct the board control, image readout, encoding and decoding, displaying, and the saving of the results.



Supplementary Figure 7. Inference on MNIST dataset using 3D VRRAM-based nvCIM macro. a, Schematic of the self-defined CNN. Six 5×5 convolution kernels are programmed into 3D VRRAM. The rest modules are implemented by software (see Methods). **b,** Inference accuracy on the MNIST dataset. Compared with PWIVMM scheme, ADINWM scheme improves the accuracy by 0.81%.



Supplementary Figure 8. Power consumption at different configurations. 1bIN-2bW operation using PWIVMM scheme consumes more power because it includes 32 VMM operations in one cycle, while ADINWM scheme only includes one VMM operation in one cycle.



Supplementary Figure 9. Energy ratio at different configurations. a, Energy ratio of 1bIN-2bW VMM operation using PWIVMM scheme. **b,** Energy ratio of 1bIN-2bW VMM operation using ADINWM scheme. DSCC occupies the main energy consumption at this configuration, because we use the 8 bit DSCC as 1 bit DSCC. **c,** Energy ratio of 4bIN-5bW VMM operation using ADINWM scheme. **d,** Energy ratio of 8bIN-9bW VMM operation using ADINWM scheme. Because of the multi-level characteristics of our 3D VRRAM, we use four memristors to store 8 bit weights. Therefore, compared with 4bIN-5bW operation, 3D VRRAM occupies less energy consumption.

Supplementary Note 6: Power consumption calculations of the different configurations.

The power consumption, calculated by the average current and supply voltage at the same operating frequency, are based on the modules used in different working mode. The total power consumption can be calculated as follows:

$$P = \sum n_i \times I_i \times V_i$$

where I_i represents the average current derived and calculated from the actual measured value, V_i denotes the supply voltage, n_i is the number of the corresponding working modules.

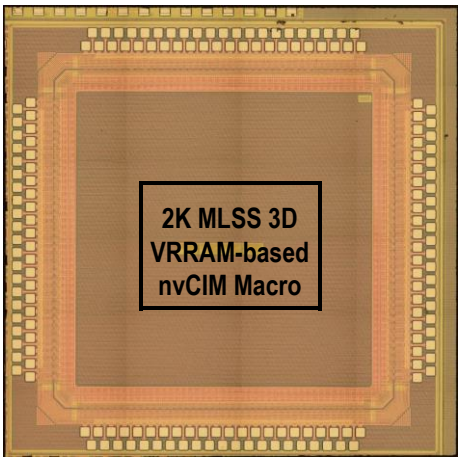
	PWIVMM (1bIN-2bW)				ADINWM (1bIN-2bW)				ADINWM (4bIN-5bW)				ADINWM (8bIN-9bW)			
	n	I (nA)	V (V)	P (nW)	n	I (nA)	V (V)	P (nW)	n	I (nA)	V (V)	P (nW)	n	I (nA)	V (V)	P (nW)
RRAM	32	5	2.5	400	1	2	2.5	5	4	0.5	2.5	5	4	0.5	2.5	5
CADS&AM					1	5	1.2	6	1	20	1.2	24	1	70	1.2	84
DSCC	1	40	1.8	72	1	11	1.8	19.8	1	20	1.8	36	3	40 (20)	1.8	144
DIGITAL					1	1.2	1.2	1.44	1	1.5	1.2	1.8	1	6	1.2	7.2
TOTAL	472				32.2				66.8				240.2			

Notes about 4bIN-5bW and 8bIN-9bW modes: According to the typical MNIST dataset provided by the literature and the sparsity parameter provided, the percentage of 0 in the input matrix is more than 80% [S.4-S.6], we did the corresponding analysis and calculated power consumption.

Supplementary Table 1. Comparison between this work and previous nvCIM works. FoM refers to figures of merit, which can be obtained from FoM = Energy Efficiency × Input Precision × Weight Precision × (Output Precision / Full Precision). Our work has achieved competitive results in computing precision, energy efficiency, FoM and inference accuracy on publication dataset.

Works	Nature'20 [S.7]	ISSCC'20 [S.8]			ISSCC'20 [S.9]	Nat. Electron'19 [S.10]			Nat. Electron'20 [S.11]	This Work (PWIVMM)	This Work (ADINWM)		
Technology	130 nm	22 nm			130 nm	65 nm			130 nm	55 nm	55 nm		
Input Precision	8b	1b	2b	4b	1b	1b			8b	1b	1b	4b	8b
Weight Precision	4b	2b	4b	4b	3b	2b			2b	2b	2b	5b	9b
RRAM	1T1R	1T1R			1T1R	1T1R			1R	MLSS Cell	MLSS Cell		
Bit Density (bit/um^2)	14	N/A			0.59	N/A			N/A	58.2	58.2		
Dimension	2D	2D			2D	2D			3D	3D	3D		
Capacity	16Kb	2Mb			158.8Kb	1Mb			1Kb	2Kb	2Kb		
Accumulation	N/A	16	16	16	N/A	9	9	9	N/A	32	32	32	32
MAC Output Precision	N/A	6b	10b	11b	N/A	2b	3b	4b	N/A	7b	7b	14b	22b
Ideal Precision	N/A	6b	10b	12b	N/A	6b	6b	6b	N/A	7b	7b	14b	22b
Power (W)	7.438m	0.84u	1.68u	1.89u	8-80u	8.03u	8.03u	7.97u	N/A	472n	32.2n	66.8n	240.2n
Read Latency	N/A	9.8ns	13.1ns	18.3ns	77us/image	4.9ns	9.8ns	14.8ns	N/A	1us	250ns	840ns	1us
Power*Tac	N/A	16.47f	43.94f	69.13f	N/A	39.35f	78.69f	117.96f	N/A	472f	8.05f	56.11f	240.2f
Energy Efficiency (TOPS/W)	11.02	121.38	45.52	28.93	78.4	50.85	25.42	16.95	19.48	135.59	62.11	29.94	8.32
Fom	N/A	242.76	364.16	424.31	N/A	33.9	25.42	22.6	N/A	271.18	124.22	598.8	599.04
MNIST Accuracy	95.83%	N/A	N/A	N/A	94.40%	89.16%	98.82%	99.03%	98.10%	93.89%	94.70%	97.04%	97.21%
CIFAR-10 Accuracy	94.08%	N/A	90.18%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	90.54%

a



b

CHIP SUMMARY	
Technology	55nm CMOS Process
RRAM	MLSS 3D VRRAM
Macro Size (Inc. CIM Macro and testmodes)	550um × 430um
Macro Mode	(1) Memory (2) CIM (MAC)
Capacity	2K
Power Consumption @VDD = 1.2V	CIM Mode (ADINWM): 32.2 nW (1bIN-2bW-7bOUT)
	CIM Mode (ADINWM): 66.8 nW (4bIN-5bW-14bOUT)
	CIM Mode (ADINWM): 240.2 nW (8bIN-9bW-22bOUT)
Energy Efficiency @VDD = 1.2V	CIM Mode (ADINWM): 62.11 TOPS/W (1bIN-2bW-7bOUT)
	CIM Mode (ADINWM): 29.94 TOPS/W (4bIN-5bW-14bOUT)
	CIM Mode (ADINWM): 8.32 TOPS/W (8bIN-9bW-22bOUT)

Supplementary Figure 10. 3D VRRAM-based nvCIM macro. a, Die photo of our chip. **b,** Summary of our chip.

Supplementary References

[S.1] Luo, Q. et al. 8-Layers 3D vertical RRAM with excellent scalability towards storage class memory applications. In 2017 *IEEE International Electron Devices Meeting, IEDM* 2.7.1–2.7.4 (IEEE, 2017).

[S.2] Park, S. et al. RRAM-based Synapse for Neuromorphic System with Pattern Recognition Function. In 2012 IEEE *International Electron Devices Meeting, IEDM* 10.2.1–10.2.4 (IEEE, 2012).

[S.3] Woo, J. et al. Resistive Memory-Based Analog Synapse: The Pursuit for Linear and Symmetric Weight Update. *IEEE Nanotechnology magazine.* 36-44 (2018).

[S.4] Zhang, T. et al. A Systematic DNN Weight Pruning Framework using Alternating Direction Method of Multipliers. *European conference on computer vis*ion *(ECCV)*, 184-199 (2018).

[S.5] Han, S. et al. EIE: Efficient Inference Engine on Compressed Deep Neural Network. *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture*, *ISCA*, 243-254. (2016)

[S.6] Lecun, Y. et al. Learning algorithms for classification: a comparison on handwritten digit recognition. *Neural Networks: The Statistical Mechanics Perspectives*, 261-276 (1995)

[S.7] Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* ***577***, 641–646 (2020).

[S.8] Xue, C. et al. A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices. *In 2020 IEEE Int. Solid - State Circuits Conference-(ISSCC)* 244-246 (IEEE, 2020).

[S.9] Liu, Q. et al. A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing. *In 2020 IEEE Int. Solid - State Circuits Conference-(ISSCC)* 500-502 (IEEE, 2020).

[S.10] Chen, W. *et al.* CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. *Nat Electron.* **2**, 420–428 (2019).

[S.11] Lin, P. et al. Three-dimensional memristor circuits as complex neural networks. *Nat. Electron.* **3**, 225–232 (2020).