A 28nm Hybrid 2T1R RRAM Computing-in-Memory Macro for Energy-efficient AI Edge Inference

Wang Ye[1,3], Chunmeng Dou[1,3], Linfang Wang[1,3], Zhidao Zhou[1,3], Junjie An[1], Weizeng Li[1,3], Hanghang Gao[1,3], Xiaoxin Xu[1,3], Jinshan Yue[1], Jianguo Yang[1,3], Jing Liu[1,3], Dashan Shang[1,3], Jinghui Tian[2], Qi Liu[1,2], Ming Liu[1,2]

[1]Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China, [2]Frontier Institute of Chip and System, Fudan University, Shanghai, China, [3]University of Chinese Academy of Sciences, Beijing, China

2022 IEEE Asian Solid-State Circuits Conference (A-SSCC) | 978-1-6654-7143-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/A-SSCC56115.2022.9980669

Resistive Memory (RRAM) based computing-in-memory (CIM) can play a key role in intermittently operated AI edge devices and sensors [1-3]. By minimizing the data movement during computation and power switching, it effectively reduces the power and hence improves the real-time performance compared to conventional energy-efficient system. However, as shown in Fig. 1, previous RRAM CIM macros, based on 1T1R or 2T2R cell array, are still facing several critical challenges, including (1) large summation current induced considerable IR drop, which results in the nonlinearity between multiply-and-accumulate (MAC) signals and MAC values (MACV) [4], (2) MAC signal margin degradation due to small resistance-ratio (R-ratio), which makes the leakage of high-resistive-state (HRS) cell become non-negligible compared to that of the low-resistive-state (LRS) one [5], and (3) large hardware-cost for the analogue readout circuitry, which typically requires 2^N-1 stand-by reference signals for N-bit output precision [6].

To deal with these challenges, this work presents the first hybrid-2T1R (H2T1R) RRAM CIM macro to perform highly efficient and reliable analogue MAC. Compared to the conventional 1T1R cell array, it achieves >13X enhanced R-ratio, >80% reduced summation current, >67% smaller word-line (WL) voltage (V_{WL}) for low VDD operation and read disturb suppression, and precise multi-bit weight encoding with signal-level-cell. Besides, the proposed reference-subtracting current sense amplifier (RS-CSA) requires only N reference signals for N-bit output and halves the maximum dynamic range of the current mirror. The test-chip is fabricated using the 28nm high-k/metal-gate (HKMG) process with BEOL integrated TaOx RRAM. It can perform highly-linear analogue MAC over 32 input channels with >63% and >79% power saving for the cell array and the analogue readout circuitry, respectively.

The proposed design, as shown in Fig.2, features (1) a macro structure based on the H2T1R cell with decoupled data path and periphery circuits for memory and CIM mode to meet their different requirements, (2) weighted cell array to encode 3-bit weight using three H2T1R SLC cells and (3) a 4-bit RS-CSA with reduced reference generation cost. The H2T1R cell hybridizes one IO transistor (T1) as the selector for the memory data-path to tolerate the write voltage and another core device (T2) for the CIM data-path to improve the computing performance. The weighted cell array comprises four sub-arrays. Three sub-arrays are used to store 3-bit weight data for analogue MAC, in which different numbers of multiplier (m=1, m=2, and m=4) of T2 in the H2T1R cell are leveraged to represent different bit positions. One additional sub-array with m=4 is used as the redundancy for the most-significant-bit (MSB). In the memory mode (CIMB = 1), the RRAM cells can be addressed by the row and column address (AX and AY) to write/read (WEB = 0/1) the input/output data (DIN/DOUT). In the CIM mode (CIMB = 0), the RRAM cells are row-wisely accessed and the CIM input data (CIN) are applied through the bit-lines (BLs) by the analogue data input driver and the sub-array selector. The CIM output data (COUT) can be read out from the transpose BL (TBL) current (I_{TBL}) by RS-CSA.

Figure 3 shows the structure and operation of the H2T1R cell array. The IO T1 and core T2 are vertically aligned to minimize the bit-cell area overhead (30.3%). In order to optimize the signal ratio, T1 is half activated to maximize the change of the voltage at the node X (ΔV_X) at different resistive states. Maximized ΔV_X can be observed at the point that the square of the resistance of T1 (R_{T1}) equals to the product of the cell resistance at HRS (R_{HRS}) and LRS (R_{LRS}). Furthermore, T2 is sub-V_{TH} operated by controlling the bit-line voltage (V_{BL}) to amplify ΔV_X by its steep sub-threshold swing (S.S.). Besides, the sub-V_{TH} operating T2 can also make I_{TBL} become less subjective to the TBL clamping voltage (V_{TBL}). Compared to the conventional 1T1R, the H2T1R read can reduce the LRS current by 88%, reduce the HRS leakage current by 99.8%, lower V_{WL} by 67%, and enhance the signal ratio (R-ratio) by 13.5 times. Notice that lowered V_{WL} can not only support low VDD operation, but also reduce the voltage drop on the RRAM cell by 17% to suppress the read disturb. To perform MAC operation, the H2T1R array is row-wise accessed across different sub-array with analogue BL inputs. The V_{TBL} of the selected row is clamped to 100mV to read out the accumulated I_{TBL} and the rest of them are grounded. The typical H2T1R MAC behavior shows that the accumulated I_{TBL} is well proportional to the number of activated LRS cells with negligible signal margin degradation due to activated HRS cells, which can be attributed to suppressed IR drop by modulating cell current and lowered HRS leakage by R-ratio enhancement.

The proposed RS-CSA (Fig.4) mainly comprises a current mirror to compare the current (I_{COMP}) with the reference current (I_{REF}), four reference (REF) branches to generate proportional reference currents (I_{REF}[0:3]) controlled by SEN[0:3], a leakage current (I_{LEAK}) compensation branch to compensate the off-cell leakage in the cell array controlled by SW[0], three reference subtracting (REFS) branches that can subtract I_{REF}[0], I_{REF}[1], or I_{REF}[2] from the data-line current (I_{DL}) depending upon the switches (SW[1:3]), RS control for the timing and the subtracting logics, biasing reference voltage generator (V_{REF_N} for REF and V_{REF_P} for REFS), buffers and latches to output COUT[0:3]. In this configuration, I_{COMP} equals the difference between I_{DL} and the current of the REFS branches (I_{REFS}). A typical readout process, triggered by RD_EN, consists of one phase (PH0) for current stabling and leakage compensating and four phases (PH1-PH4) for current comparing and reference subtracting. In PH0, SW0 is triggered to compensate I_{LEAK} during I_{DL} stabling ($I_{COMP} = I_{DL} - I_{LEAK}$). In PH1-4, I_{COMP} is sequentially compared with I_{REF}[0:3] to output COUT[0:3], in which the RS control logic cumulatively subtracts the I_{REF}[i-1] at the ith phase from I_{DL} if I_{COMP}>I_{REF}[i-1] till the last one. The proposed RS-CSA only needs 4 references and halved the maximum range of I_{COMP}, which can reduce the reference power and extend linear dynamic range of the current mirror.

The properties of the H2T1R array and RS-CSA are shown in Fig.5. A clear memory window between the HRS and LRS cells can be observed at V_{WL}<VDD. Notice that because 256 core T2 devices are connected to TBL, there is a total I_{LEAK} about 1uA due to the off-cell leakage, which can be compensated by the RS-CSA. While I_{TBL} of the LRS cell linearly increases with increasing m of T2, that of the HRS cell keeps at a negligible level below 30 nA. Thanks to reduced cell currents and IR drop, the H2T1R array performs highly linear MAC operation over channels. Because of the near-V_{TH} operation of T1 and T2, the MAC power of H2T1R array can be reduced by more than 63.4% compared to its 1T1R counterpart. Besides, the RS-CSA achieves more than 79.5% power saving compared to the conventional current-mirror current sense amplifier (CM-CSA) by reducing the number of stand-by reference signals.

Figure 6 shows the measured waveform and power breakdown of the fabricated 28nm 8Kb HKMG H2T1R CIM macro. The measured latency (T_{AC}) is 66 ns for 4-bit readout, which can be reduced to 13 ns after sensing phase optimization. It consumes 31.96 uW at VDD = 0.8 V and achieves an energy efficiency of 30.34~154.04 TOPS/W at 1-bit input (IN), 3-bit weight (W), and 4-bit output (O). Compared to the previous work, this work demonstrated the first multi-bit H2T1R CIM macro supports highly linear analogue MAC with reduced reference generation cost.

Figure 7 shows a die photo and summary table .

Acknowledgement: This work was supported by the National Key R&D Program and NSFC under Grant No. 2018YFA0701500, 2018YFB2202900, 61904197, 61934005, 61825404, 61732020, 61821091, 61888102, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB44000000, and the Project of MOE Innovation Platform. The corresponding author is Chunmeng Dou.

References:

[1] W.-H. Chen *et al.*, “ CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors,” Nature Electronics, 2019:420-428.

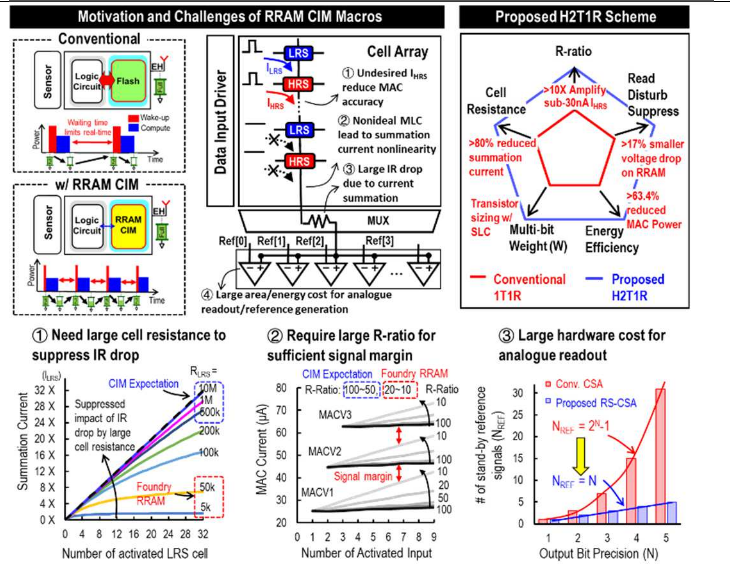
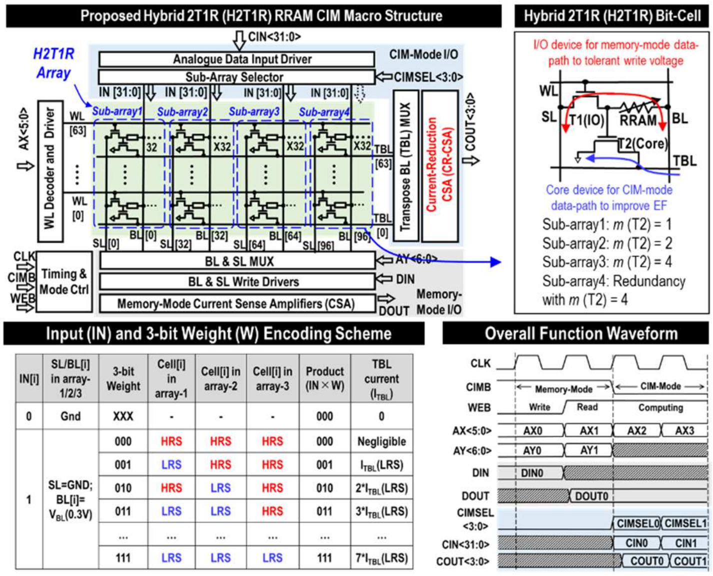

Motivation and Challenges of RRAM CIM Macros
Conventional
Sensor
Logic Circuit
Flash
EH
Waiting time limits real-time
Wake-up
Compute
Power
Time
w/ RRAM CIM
RRAM CIM
Data Input Driver
Cell Array
LRS
HRS
① Undesired I_HRS reduce MAC accuracy
② Nonideal MLC lead to summation current nonlinearity
③ Large IR drop due to current summation
MUX
Ref[0] Ref[1] Ref[2] Ref[3]
④ Large area/energy cost for analogue readout/reference generation
Proposed H2T1R Scheme
R-ratio
>10X Amplify sub-30nA I_HRS
Cell Resistance
Read Disturb Suppress
>80% reduced summation current
>17% smaller voltage drop on RRAM
Transistor sizing w/ SLC
Multi-bit Weight (W)
Energy Efficiency
>63.4% reduced MAC Power
Conventional 1T1R
Proposed H2T1R
① Need large cell resistance to suppress IR drop
Summation Current (I_LRS)
CIM Expectation
R_LRS = 10M 1M 500k 200k 100k 50k 5k
Suppressed impact of IR drop by large cell resistance
Foundry RRAM
Number of activated LRS cell
② Require large R-ratio for sufficient signal margin
CIM Expectation Foundry RRAM
R-Ratio: 100~50 20~10 R-Ratio
MACV3 MACV2 MACV1
Signal margin
MAC Current (µA)
Number of Activated Input
③ Large hardware cost for analogue readout
Conv. CSA
Proposed RS-CSA
N_REF = 2^N-1
N_REF = N
# of stand-by reference signals (N_REF)
Output Bit Precision (N)

Fig. 1. Motivation, challenges and advantages of the proposed H2T1R RRAM CIM macro.


Proposed Hybrid 2T1R (H2T1R) RRAM CIM Macro Structure
CIN<31:0>
H2T1R Array
Analogue Data Input Driver
Sub-Array Selector
CIM-Mode I/O
CIMSEL<3:0>
IN [31:0]
Sub-array1 Sub-array2 Sub-array3 Sub-array4
AX<5:0>
WL Decoder and Driver
WL [63]
WL [0]
X32
TBL [63]
TBL [0]
Transpose BL (TBL) MUX
Current-Reduction CSA (CR-CSA)
COUT<3:0>
SL[0] BL[0] SL[32] BL[32] SL[64] BL[64] SL[96] BL[96]
CLK CIMB WEB
Timing & Mode Ctrl
BL & SL MUX
AY<6:0>
BL & SL Write Drivers
DIN
Memory-Mode Current Sense Amplifiers (CSA)
DOUT
Memory-Mode I/O
Hybrid 2T1R (H2T1R) Bit-Cell
I/O device for memory-mode data-path to tolerant write voltage
WL SL T1(IO) RRAM BL T2(Core) TBL
Core device for CIM-mode data-path to improve EF
Sub-array1: m (T2) = 1
Sub-array2: m (T2) = 2
Sub-array3: m (T2) = 4
Sub-array4: Redundancy with m (T2) = 4
Input (IN) and 3-bit Weight (W) Encoding Scheme
Overall Function Waveform
CLK CIMB WEB
Memory-Mode CIM-Mode
Write Read Computing
AX<5:0> AX0 AX1 AX2 AX3
AY<6:0> AY0 AY1
DIN DIN0
DOUT DOUT0
CIMSEL<3:0> CIMSEL0 CIMSEL1
CIN<31:0> CIN0 CIN1
COUT<3:0> COUT0 COUT1

IN[i]	SL/BL[i] in array-1/2/3	3-bit Weight	Cell[i] in array-1	Cell[i] in array-2	Cell[i] in array-3	Product (IN × W)	TBL current (I_{TBL})
0	Gnd	XXX	-	-	-	000	0
1	SL=GND; BL[i]= V_{BL}(0.3V)	000	HRS	HRS	HRS	000	Negligible
		001	LRS	HRS	HRS	001	I_{TBL}(LRS)
		010	HRS	LRS	HRS	010	2*I_{TBL}(LRS)
		011	LRS	LRS	HRS	011	3*I_{TBL}(LRS)
		...	...	...	...	...	...
		111	LRS	LRS	LRS	111	7*I_{TBL}(LRS)

Fig. 2. Proposed H2T1R RRAM CIM macro, input and weight encoding method, and its operational waveforms.

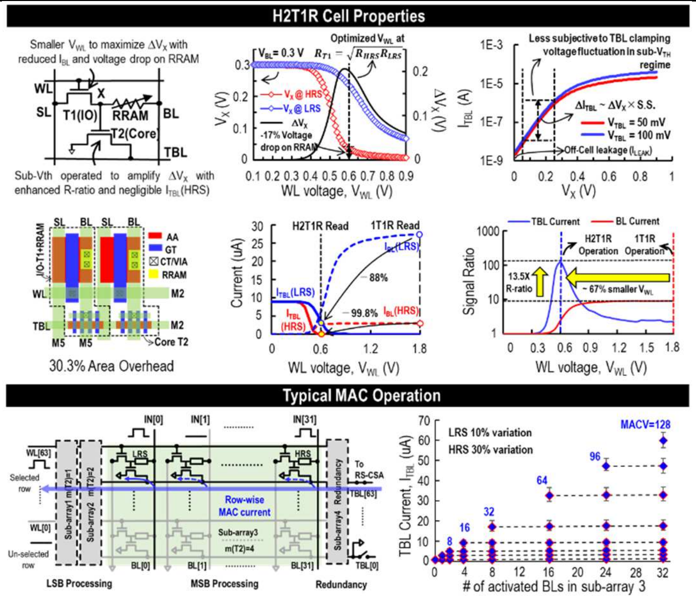


Fig. 3. H2T1R cell properties and typical MAC behaviors.

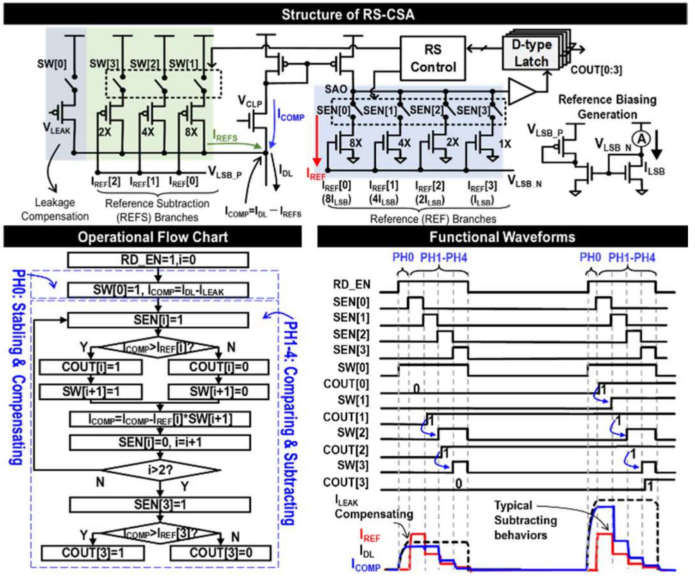


Fig. 4. Structure and operation of the proposed reference-subtracting current sense amplifier (RS-CSA).

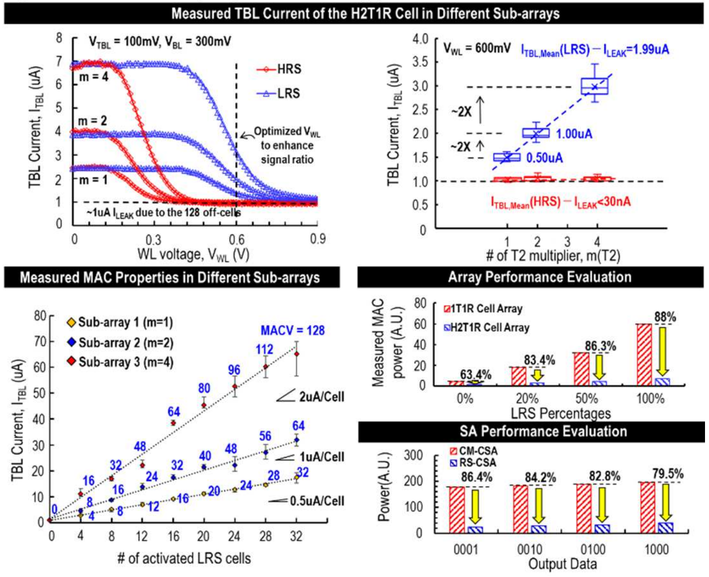
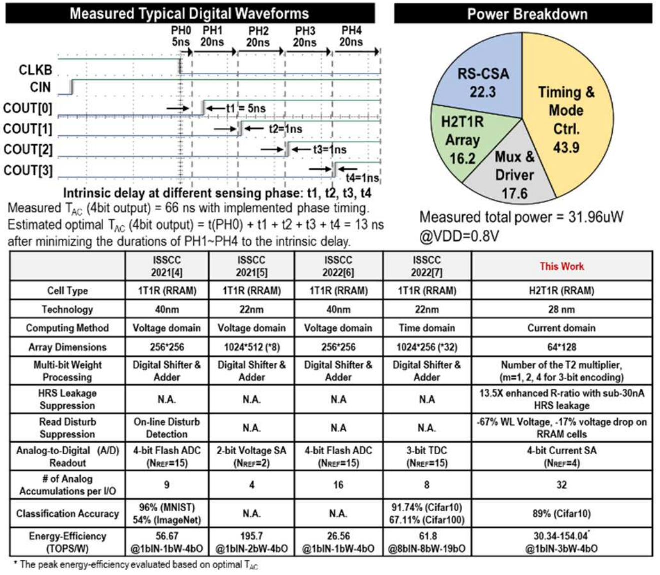


Fig. 5. Measured H2T1R cell array characteristics and performance evaluation of proposed array and sense amplifier.



Intrinsic delay at different sensing phase: t1, t2, t3, t4

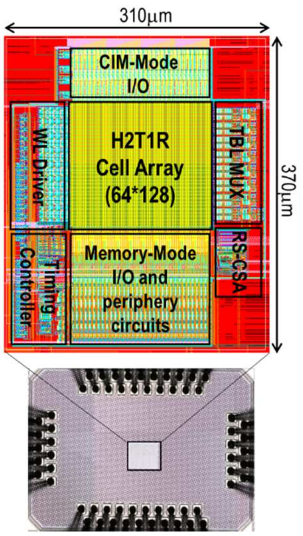
Measured T_{AC} (4bit output) = 66 ns with implemented phase timing.
Estimated optimal T_{AC} (4bit output) = t(PH0) + t1 + t2 + t3 + t4 = 13 ns after minimizing the durations of PH1~PH4 to the intrinsic delay.

Measured total power = 31.96uW @VDD=0.8V

	ISSCC 2021[4]	ISSCC 2021[5]	ISSCC 2022[6]	ISSCC 2022[7]	This Work
Cell Type	1T1R (RRAM)	1T1R (RRAM)	1T1R (RRAM)	1T1R (RRAM)	H2T1R (RRAM)
Technology	40nm	22nm	40nm	22nm	28 nm
Computing Method	Voltage domain	Voltage domain	Voltage domain	Time domain	Current domain
Array Dimensions	256*256	1024*512 (*8)	256*256	1024*256 (*32)	64*128
Multi-bit Weight Processing	Digital Shifter & Adder	Digital Shifter & Adder	Digital Shifter & Adder	Digital Shifter & Adder	Number of the T2 multiplier, (m=1, 2, 4 for 3-bit encoding)
HRS Leakage Suppression	N.A.	N.A.	N.A	N.A	13.5X enhanced R-ratio with sub-30nA HRS leakage
Read Disturb Suppression	On-line Disturb Detection	N.A.	N.A	N.A.	-67% WL Voltage, -17% voltage drop on RRAM cells
Analog-to-Digital (A/D) Readout	4-bit Flash ADC (N_{REF}=15)	2-bit Voltage SA (N_{REF}=2)	4-bit Flash ADC (N_{REF}=15)	3-bit TDC (N_{REF}=15)	4-bit Current SA (N_{REF}=4)
# of Analog Accumulations per I/O	9	4	16	8	32
Classification Accuracy	96% (MNIST) 54% (ImageNet)	N.A.	N.A.	91.74% (Cifar10) 67.11% (Cifar100)	89% (Cifar10)
Energy-Efficiency (TOPS/W)	56.67 @1bIN-1bW-4bO	195.7 @1bIN-2bW-4bO	26.56 @1bIN-1bW-4bO	61.8 @8bIN-8bW-19bO	30.34-154.04* @1bIN-3bW-4bO

* The peak energy-efficiency evaluated based on optimal T_{AC}

Fig. 6. Measured waveforms and power breakdown of the fabricated macro and the comparison table.


310µm
370µm
CIM-Mode I/O
H2T1R Cell Array (64*128)
WL-Driver
TBL MUX
RS-CSA
Timing Controller
Memory-Mode I/O and periphery circuits

Chip Summary	
Technology	28nm HKMG Logic Process
RRAM	Foundry TaOx–based RRAM
Bit-Cell	H2T1R (T1: I/O transistor, T2: core transistor)
Macro Size	310 × 370 μm^2
Array Size	64 rows × 128 columns (8Kb)
Mode	Memory/CIM
Power Supply (VDD)	0.8V (Core), 1.2V (IO)
Precision	1bIN-3bW-4bO
Energy-Efficiency	30.34-154.04 (TOPS/W)

Fig. 7. Die photo and chip summary.

Additional References:

[2] C.-X. Xue *et al*., " A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices," ISSCC 2020:244-245.

[3] C.-X. Xue *et al*., " A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices," Nature Electronics 2021:81-90.

[4] J.-H. Yoon *et al*., " A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification," ISSCC 2021:404-405.

[5] C.-X. Xue *et al*., " A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7TOPS/W for Tiny AI Edge Devices," ISSCC 2021:246-247.

[6] S D. Spetalnick *et al*., " A 40nm 64kb 26.56TOPS/W 2.37Mb/mm^2 RRAM Binary/Compute-in-Memory Macro with 4.23× Improvement in Density and >75% Use of Sensing Dynamic Range," ISSCC 2022:268-269.

[7] J.-M. Hung *et al*., " An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4 - 21.6TOPS/W for Edge-AI Devices," ISSCC 2022:182-183.