16.1 A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7TOPS/W for Tiny AI Edge Devices

Cheng-Xin Xue*[1], Je-Min Hung*[1], Hui-Yao Kao[1], Yen-Hsiang Huang[1], Sheng-Po Huang[1], Fu-Chun Chang[1], Peng Chen[1], Ta-Wei Liu[1], Chuan-Jia Jhang[1], Chin-I Su[2], Win-San Khwa[2], Chung-Chuan Lo[1], Ren-Shuo Liu[1], Chih-Cheng Hsieh[1], Kea-Tiong Tang[1], Yu-Der Chih[2], Tsung-Yung Jonathan Chang[2], Meng-Fan Chang[1,2]

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]TSMC, Hsinchu, Taiwan

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9365769

Battery-powered tiny-AI edge devices require large-capacity nonvolatile compute-in-memory (nvCIM), with multibit input (IN), weight (W), and output (OUT) precision to support complex applications, high-energy efficiency (EF$_{MAC}$), and short computing latency (t_{AC}) for multiply-and-accumulate (MAC) operations. Due to the low read-disturb-free voltage of nonvolatile memory (NVM) devices and the large parasitic load on the bitline, most existing Mb-level nvCIM macros use a current-mode read scheme [1-5] and only achieve a low IN-W precision (binary to 4b).

As shown in Fig. 16.1.1, the above requirements impose serious challenges: (1) Long input latency due to input schemes using serial WL pulse-counts [6] or a fully-decoded WL pulse-width; (2) Limited system-level accuracy for high-precision MAC operations due to small signal margin caused by process variation and pattern-dependent bitline current under various input bit configurations; and (3) Large power consumption in readout circuit due to DC current consumption in the current-summation of partial MAC values (pMACV) with respective place values. To overcome these challenges we developed: (1) An asymmetric group-modulated input (AGMI) scheme, in which a 8b-input is split into 3 sub-groups (2b-3b-3b) to reduce computing latency, while maintaining sufficient signal margins for the most significant bits (MSBs); (2) A weighted current-to-voltage signal stacking (WCVSS) converter to translate the dataline current (I$_{DL}$) of partial MAC operations into voltage-mode signals with respective place values; and (3) A hybrid-precision voltage-mode readout scheme with voltage-mode sense amplifier (VSA) to reduce energy consumption and shorten the t_{AC} for multibit MAC readout operations, while maintaining the sensing margin for MSBs.

The proposed 22nm 4Mb ReRAM macro, fabricated using foundry SLC 1T1R ReRAM technology, is the first nvCIM macro supporting 8b-input and 8b-weight MAC operations. Among silicon-verified nvCIMs, it achieved the fastest t_{AC} (4.9-14.8ns) and best EF$_{MAC}$ (195.7-11.91TOPS/W) with precision from binary IN-W to 8bIN-8bW-14bOUT. Our nvCIM macro delivers MAC operations comprising 4 sets of 8bIN and 8bW in the channel direction. Note that 4 accumulations per column (bitlines/BL) provide sufficient signal margin against process variation. 8bW data is stored in 8 ReRAM cells on the same row (wordline/WL) across 8 columns in 2's complement format. Each set of columns comprises memory cells, column multiplexors (column MUX), a WCVSS converter, a VSA, and digital shift-and-add circuits (DSAC).

Figure 16.1.2 shows the operation of AGMI on one column for the LSB of weights (W[0]), in which an 8b-input (IN[7:0]) is split into 3 grouped signals (IN[7:6], IN[5:3], and IN[2:0]), generating three sequential WL pulses (WLP2~WLP0) with corresponding pulse-widths. The pulse-width of WLP2 (T$_{WLP2}$) includes one precharge and BL current-development phase (P-BLPD, T$_{BL-PD}$) and up to 3 time-units (T$_{U-WCVSS1}$) for WCVSS operations. Each pulse-width of WLP1 and WLP0 (T$_{WLP1}$ = T$_{WLP0}$ = T$_{BL-PD}$ + 7·T$_{U-WCVSS2}$) includes 1 T$_{BL-PD}$ and up to 7 T$_{U-WCVSS2}$. Note that T$_{U-WCVSS1}$ is 7/3× longer than T$_{U-WCVSS2}$ to ensure sufficient signal margin for the pMACV of MSBs (IN[7:6]) of inputs. In the WLP2 phase, the BL is precharged to the read voltage and then each activated memory cell (MC) (at WL=1) generates a memory cell current (I_{MC}), based on stored weight data (HRS or LRS). The BL current (I_{BL}) is the sum of the four I_{MC} representing the partial MAC result of 2b-input, 1b-weight, and 4 accumulations. In the WLP1 and WLP0 phases, each I_{BL} is the partial MAC result of 3b-input, 1b-weight, and 4 accumulations.

Figure 16.1.3 shows the operation of WCVSS: comprising of three current-scaling transistors (P1, P2, P3), two current-mirror inhibitors (PS1, PS2), two bias PMOS (BP0, BP1), four switches (SW0, SW1, SW2, SW3), two initial transistors (N0, N1), and two capacitors for voltage stacking (C$_C$, C$_S$). In the WLP2 phase, N1 is on (S1=1) to keep node STACK at 0V, while SW0 is on to build the P1-P2 current-mirror circuit for down-scaling the dataline current (I_{DL}) to I_{WDL0} (=0.5·I_{DL}) with a scaling ratio of 0.5. Here, EN1=1 and SW0 is turned on up to 3 times to sample I_{WDL0} and then charge the node SUM connected to capacitor C$_C$ and the input capacitor (C$_0$) of VSA. The voltage at node SUM (V$_{SUM-P2}$, maximum value is 8/9·V$_{DD}$) is determined by I_{WDL0} sampled during the input phase. At the end of the WLP2 phase, VSA reads out the V$_{SUM-P2}$ and turns on N0 (S0=1) to reset C$_0$ and C$_C$ to 0V. In the WLP1 phase, the generation of voltage at node SUM (V$_{SUM-P1}$) is similar to that of WLP2 phase, but with up to 7 SW0/EN1 pulses to sample I$_{WDL0}$ and no readout operation at the end. In the WLP0 phase, SW1 is off, while SW3 is on to downscale I_{DL} through current-mirror pairs P1-P3 with a 1/16 scaling ratio, resulting I_{WDL1} = 1/16·I_{DL}. Here, EN2=1 and SW2 is turned on up to 7 times to sample I_{WDL1} and charge the node STACK connected to capacitors C$_C$ and C$_S$, which make node STACK has matching capacitance at node SUM. Then, I_{WDL1} charges capacitor C$_S$ to generate stacking voltage (V$_{STACK}$, maximum value is 1/9·V$_{DD}$) at node STACK. The value of V$_{STACK}$ is simultaneously coupled to node SUM through capacitor C$_C$, such that the V$_{SUM-P0}$ of WLP0 is the sum of V$_{SUM-P1}$ and V$_{STACK}$. At the end of the WLP0 phase, VSA reads out the V$_{SUM-P0}$.

Figure 16.1.4 shows the hybrid-precision voltage-mode readout scheme with VSA. The V$_{SUM-P2}$ resulting from the WCVSS of the WLP2 phase is divided into 16 levels for full-precision sensing. The single VSA is activated twice (SAEN=1) at the end of the WLP2 phase to read V$_{SUM-P2}$ and generate a 4b digital output (SAOUT1[3:0]) for the pMACV result of $IN_0[7:6]xW_0[0]+IN_1[7:6]xW_1[0]+IN_2[7:6]xW_2[0]+IN_3[7:6]xW_3[0]$. At the end of WLP0 phase, the same VSA is activated twice to read V$_{SUM-P0}$ with reduced-precision of 4b (full-precision is 8b) and generate a 4b digital output (SAOUT2[3:0]) for the pMACV result of $IN_0[5:0]\cdot W_0[0] + IN_1[5:0]\cdot W_1[0] + IN_2[5:0]\cdot W_2[0] + IN_3[5:0]\cdot W_3[0]$. The 1st-level of DSAC (L1-DSAC) combine SAOUT1[3:0] and SAOUT2[3:0] to generate a 6b pMACV (DSACout0[5:0]) for $IN_0[7:0]\cdot W_0[0] + IN_1[7:0]\cdot W_1[0] + IN_2[7:0]\cdot W_2[0] + IN_3[7:0]\cdot W_3[0]$. The 2nd-level of DSAC (L2-DSAC) combines 8 DSACout of 8 computing columns (DSACout7 - DSACout0 for W[7:0]) with place value (DSACout7[5:0]·(-128) + DSACout6[5:0]·(64) + … + DSACout0[5:0]·(1)) to output a 14b DOUT (DOUT[13:0]) for the MAC operation of 8bIN-8bW-4Accumulations.

Figure 16.1.5 shows the performance of the proposed schemes. Under 8b-input, the proposed AGMI scheme reduces computing latency by 5.59-54.25×, compared to previous serial WL pulse-counts or the fully-decoded WL pulse-width schemes extended to 8b-input. This work has achieved a much higher signal margin for MSBs (with full precision) and LSBs (with reduced precision) of pMACVs compared to those of previous WL pulse-counts and WL pulse-width input schemes to achieve high macro-level read yield, which considers the resistance variation and pattern-dependent variation in signal values. Compared to previous work [5], this work improved EF$_{MAC}$ by 1.66× under 4bIN-4bW mode with 1024 operations.

Figure 16.1.6 shows the measurement results from a 22nm 4Mb ReRAM nvCIM macro fabricated using foundry ReRAM devices. The shmoo test for 8bIN-8bW-14bOUT confirmed a 14.8ns t_{AC} at a 0.8V V$_{DD}$. The measured EF$_{MAC}$ was 47.26TOPS/W at 4bIN-4bW-10bOUT and 11.91TOPS/W at 8bIN-8bW-14bOUT. Compared to previous work [5], despite having twice the memory capacity, this work improved the FoM (EF$_{MAC}$ × Input-precision × Weight -precision × Output-ratio / Computing latency) by 3.16-4.1× for binary up to 4bIN-4bW configurations. Figure 16.1.7 shows a die photo and summary table.

Acknowledgement:
The authors would like to thank MOST-Taiwan, TSRI, NTHU-TSMC JDP for financial and manufacturing support.

References:

[1] W.-H. Chen et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, Feb. 2018.
[2] R. Mochida et al., "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," *VLSI Circuits*, pp. 175-176, 2018.
[3] C.-X. Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," *ISSCC*, pp. 388-389, 2019.
[4] Q. Liu et al., "A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," *ISSCC*, pp. 500-501, 2020.
[5] 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*, pp. 244-245, 2020.
[6] Q. Dong et al., "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," *ISSCC*, pp. 242-243, 2020.

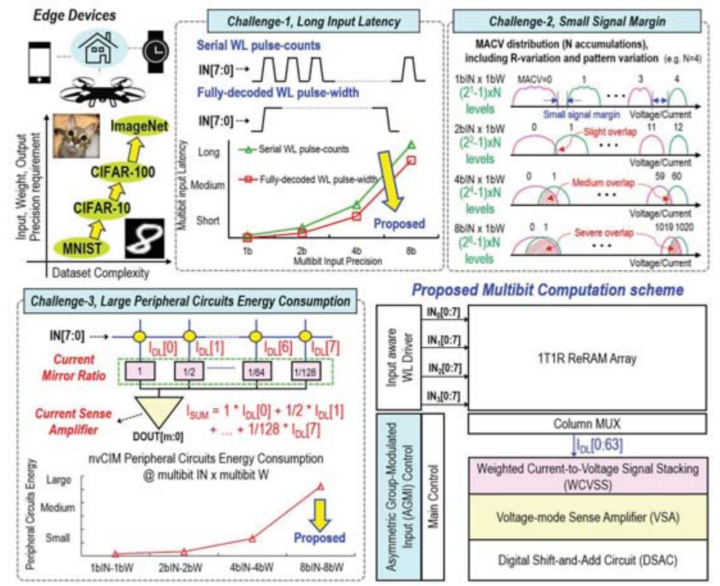
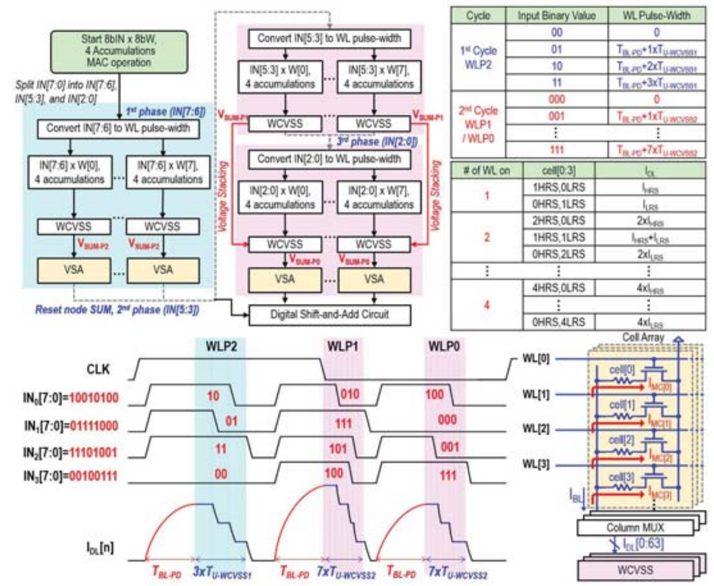

Edge Devices
ImageNet
CIFAR-100
CIFAR-10
MNIST
Input, Weight, Output Precision requirement
Dataset Complexity
Challenge-1, Long Input Latency
Serial WL pulse-counts
IN[7:0]
Fully-decoded WL pulse-width
Multibit input Latency
Long
Medium
Short
Serial WL pulse-counts
Fully-decoded WL pulse-width
Proposed
1b
2b
4b
8b
Multibit Input Precision
Challenge-2, Small Signal Margin
MACV distribution (N accumulations), including R-variation and pattern variation (e.g. N=4)
1bIN x 1bW (2^1-1)xN levels
Small signal margin
2bIN x 1bW (2^2-1)xN levels
Slight overlap
4bIN x 1bW (2^4-1)xN levels
Medium overlap
8bIN x 1bW (2^8-1)xN levels
Severe overlap
Voltage/Current
Challenge-3, Large Peripheral Circuits Energy Consumption
IN[7:0]
Current Mirror Ratio
1
1/2
1/64
1/128
Current Sense Amplifier
DOUT[m:0]
I_SUM = 1 * I_DL[0] + 1/2 * I_DL[1] + ... + 1/128 * I_DL[7]
nvCIM Peripheral Circuits Energy Consumption @ multibit IN x multibit W
Peripheral Circuits Energy
Large
Medium
Small
1bIN-1bW
2bIN-2bW
4bIN-4bW
8bIN-8bW
Proposed
Proposed Multibit Computation scheme
Asymmetric Group-Modulated Input (AGMI) Control
Input aware WL Driver
IN_0[0:7]
IN_1[0:7]
IN_2[0:7]
IN_3[0:7]
1T1R ReRAM Array
Column MUX
I_DL[0:63]
Weighted Current-to-Voltage Signal Stacking (WCVSS)
Voltage-mode Sense Amplifier (VSA)
Digital Shift-and-Add Circuit (DSAC)
Main Control

Figure 16.1.1: Challenges and proposed scheme for nvCIM.


Start 8bIN x 8bW, 4 Accumulations MAC operation
Split IN[7:0] into IN[7:6], IN[5:3], and IN[2:0]
1st phase (IN[7:6])
Convert IN[7:6] to WL pulse-width
IN[7:6] x W[0], 4 accumulations
IN[7:6] x W[7], 4 accumulations
WCVSS
V_SUM-P2
VSA
Reset node SUM, 2nd phase (IN[5:3])
Convert IN[5:3] to WL pulse-width
IN[5:3] x W[0], 4 accumulations
IN[5:3] x W[7], 4 accumulations
V_SUM-P1
3rd phase (IN[2:0])
Convert IN[2:0] to WL pulse-width
IN[2:0] x W[0], 4 accumulations
IN[2:0] x W[7], 4 accumulations
V_SUM-P0
Voltage Stacking
Digital Shift-and-Add Circuit
WLP2
WLP1
WLP0
CLK
IN_0[7:0]=10010100
IN_1[7:0]=01111000
IN_2[7:0]=11101001
IN_3[7:0]=00100111
I_DL[n]
T_BL-PD
3xT_U-WCVSS1
7xT_U-WCVSS2
WL[0]
WL[1]
WL[2]
WL[3]
Cell Array
cell[0]
cell[1]
cell[2]
cell[3]
I_BL
Column MUX
I_DL[0:63]
WCVSS

Cycle	Input Binary Value	WL Pulse-Width
1st Cycle WLP2	00	0
	01	$T_{BL-PD}+1xT_{U-WCVSS1}$
	10	$T_{BL-PD}+2xT_{U-WCVSS1}$
	11	$T_{BL-PD}+3xT_{U-WCVSS1}$
2nd Cycle WLP1 / WLP0	000	0
	001	$T_{BL-PD}+1xT_{U-WCVSS2}$
	⋮	⋮
	111	$T_{BL-PD}+7xT_{U-WCVSS2}$

# of WL on	cell[0:3]	I_{DL}
1	1HRS,0LRS	I_{HRS}
	0HRS,1LRS	I_{LRS}
2	2HRS,0LRS	$2xI_{HRS}$
	1HRS,1LRS	$I_{HRS}+I_{LRS}$
	0HRS,2LRS	$2xI_{LRS}$
⋮	⋮	⋮
4	4HRS,0LRS	$4xI_{HRS}$
	⋮	⋮
	0HRS,4LRS	$4xI_{LRS}$

Figure 16.1.2: Multibit MAC operations using AGMI scheme.

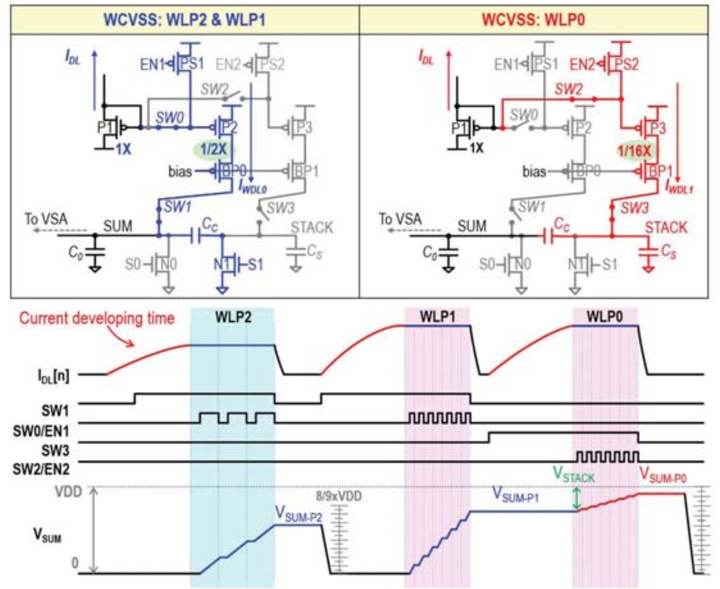


Figure 16.1.3: Operations of WCVSS converter.

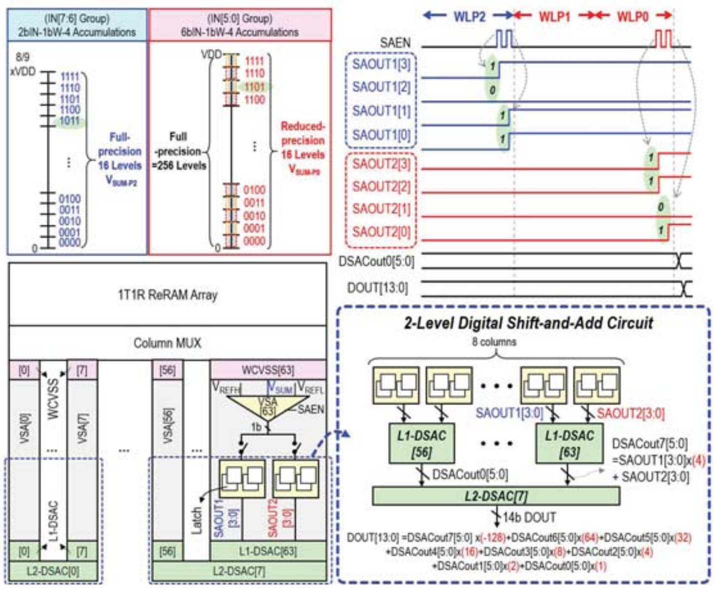

(IN[7:6] Group) 2bIN-1bW-4 Accumulations
(IN[5:0] Group) 6bIN-1bW-4 Accumulations
8/9 xVDD
VDD
Full-precision 16 Levels V_SUM-P2
Full-precision =256 Levels
Reduced-precision 16 Levels V_SUM-P0
1111
1110
1101
1100
1011
0100
0011
0010
0001
0000
WLP2
WLP1
WLP0
SAEN
SAOUT1[3]
SAOUT1[2]
SAOUT1[1]
SAOUT1[0]
SAOUT2[3]
SAOUT2[2]
SAOUT2[1]
SAOUT2[0]
DSACout0[5:0]
DOUT[13:0]
1T1R ReRAM Array
Column MUX
WCVSS
WCVSS[63]
V_REFH
V_SUM
V_REFL
VSA[63]
SAEN
1b
VSA[0]
VSA[7]
VSA[56]
L1-DSAC
Latch
SAOUT1[3:0]
SAOUT2[3:0]
L1-DSAC[63]
L2-DSAC[0]
L2-DSAC[7]
2-Level Digital Shift-and-Add Circuit
8 columns
L1-DSAC [56]
L1-DSAC [63]
DSACout7[5:0] =SAOUT1[3:0]x(4) + SAOUT2[3:0]
DSACout0[5:0]
L2-DSAC[7]
14b DOUT
DOUT[13:0] =DSACout7[5:0] x(-128)+DSACout6[5:0]x(64)+DSACout5[5:0]x(32) +DSACout4[5:0]x(16)+DSACout3[5:0]x(8)+DSACout2[5:0]x(4) +DSACout1[5:0]x(2)+DSACout0[5:0]x(1)

Figure 16.1.4: Hybrid-precision voltage-mode readout scheme.

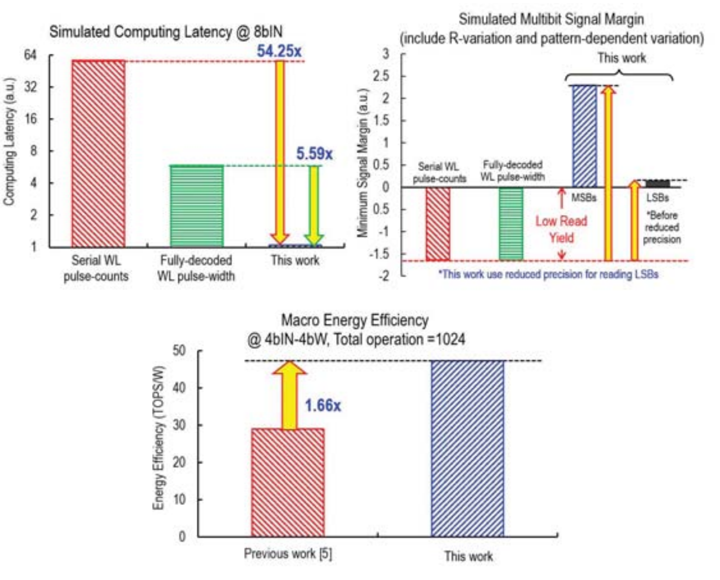


Figure 16.1.5: Simulated performance of proposed scheme.

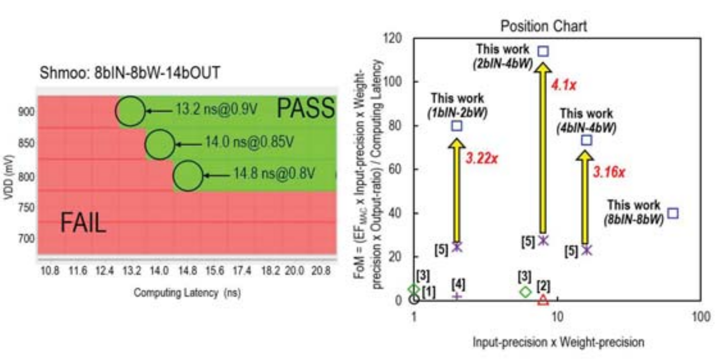
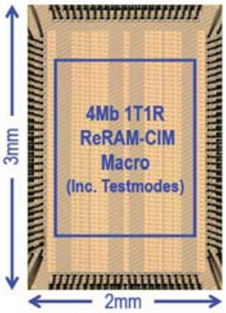


Figure 16.1.6: Measurement results.

16



Chip Summary		
Technology	22nm CMOS Logic Process	
ReRAM	Foundry 1T1R SLC ReRAM	
Testchip Size	2mm x 3mm (Inc.IO pad and testmodes)	
Capacity	4Mb (8 Sub-bank)	
Sub-bank	1024 rows x 512 columns	
Performance @ VDD=0.8V		
CIM-mode Computing Latency (ns)	1bIN-2bW-4bOUT	4.9
	4bIN-4bW-10bOUT	10.3
	8bIN-8bW-14bOUT	14.8
Throughput (GOPS)	1bIN-2bW-4bOUT	417.96
	4bIN-4bW-10bOUT	99.42
	8bIN-8bW-14bOUT	35.59
Energy Efficiency (TOPS/W)	1bIN-2bW-4bOUT	195.7
	4bIN-4bW-10bOUT	47.26
	8bIN-8bW-14bOUT	11.91

Figure 16.1.7: Die photo and chip summary.