A 40nm 100Kb 118.44TOPS/W Ternary-weight Compute-in-Memory RRAM Macro with Voltage-sensing Read and Write Verification for reliable multi-bit RRAM operation

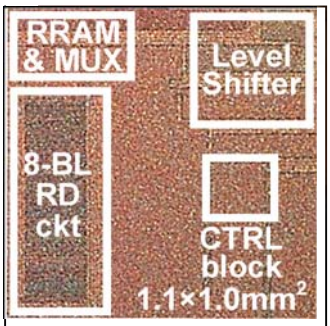
Jong-Hyeok Yoon[1,2], Muya Chang[1], Win-San Khwa[3], Yu-Der Chih[4], Meng-Fan Chang[3], Arijit Raychowdhury[1]

[1]Georgia Institute of Technology, Atlanta, GA, [2]DGIST, Daegu, Republic of Korea, [3]TSMC Corporate Research, Hsinchu, Taiwan, [4]TSMC Design Technology, Hsinchu, Taiwan

RRAM is a promising candidate for compute-in-memory (CIM) applications owing to its natural multiply-and-accumulate (MAC)-supporting structure, high bit-density, non-volatility, and a monolithic CMOS and RRAM process. In particular, multi-bit encoding in RRAM cells helps support advanced applications such as AI with higher MAC throughput and bit-density. Notwithstanding prior efforts into commercializing RRAM technology, underlying challenges hinder the wide usage of RRAM [1]. As a circuit-domain approach to address the challenges, this paper presents a 101.4Kb ternary-weight RRAM macro with 256x256 cells supporting: (1) CIM for ternary weight networks by employing voltage-based read (RD) with active feedback surmounting a low resistance ratio (R-ratio) between the high resistance state (HRS) and the low resistance state (LRS) in high-endurance RRAM, and (2) iterative write with verification (IWR) to facilitate a reliable multi-bit encoding under a narrow margin. Compared to [2] supporting CIM with binary RRAM cells, this work provides 38.44x ($=3^{3x3}/2^{3x3}$) flexibility on 3x3 filters in convolutional neural networks (CNNs), and 1.585x bit density improvement, thereby enabling advanced CIM applications with ternary weight networks.

Fig. 1 shows the proposed CIM RRAM macro supporting ternary weight networks. The proposed RRAM macro provides ternary write (WR) including the forming process, and 8-BL RD accessing up to 9 WLs simultaneously to support 3x3 convolution, a typical kernel size in convolutional neural networks (CNNs). The BL in RD is selected by the BL/SL MUX and connected with the ADC-based readout circuit that supports concurrent CIM at 8 BLs. In MAC operations with multi-bit RRAM cells, a high R-ratio is desirable to attain sufficient RD margin in particular for multi-bit encoding. However, since the CIM applications such as online learning entail frequent WRs for weight updates, the trade-off between the endurance and the R-ratio should be carefully addressed [1]. The proposed macro provides circuit solutions to these underlying challenges such as a low RD margin that is exacerbated in multi-bit encoding, and wide resistance variations that limit multi-bit encoding under a narrow encoding margin in high-endurance RRAM.

The MAC operation at the BLs of the proposed macro is shown in Fig. 2. The voltage-sensing ternary-weight CIM operation is conducted by aggregating the current of accessed RRAM cells once triggered by a 9-bit input X[t] applied to the WLs. Compared to prior designs that use current-sensing RD that suffers from logic ambiguity under a low R ratio [3-6], the proposed macro determines the MAC output by sensing the RD-BL voltage (V.RBL). The input-aware (IA) current control (Fig. 1) is used to provide the current proportional to the number of accessed RRAM cells ($\Sigma X[t]$), thereby mitigating the drastic decrease of V.RBL over the parallel resistances of accessed RRAM cells. However, the remaining nonlinearity over the composition of resistance states (Fig. 2) including the intermediate resistance state (IRS) exacerbates a narrow sampling margin at the ADC in the readout circuits. Thus, active feedback (FB) control at BLs (Fig. 3) is employed to control the current source, thereby linearizing the sampling levels in the proposed macro with 2x increase in a sampling margin. Owing to IA current control and active FB control, V.RBL can be represented as a linear combination of the V.RBL with 1 HRS, 1 IRS, and 1 LRS. (V.HRS/V.IRS/V.LRS) shown in Fig. 3. The linearized V.RBL is scanned by a 4-bit flash ADC. Since V.RBL has the same range of voltages over $\Sigma X[t]$, the reference voltages are linearly distributed and the IA ADC decoder determines the MAC output with the ternary-weight RRAM macro considering the ADC output and $\Sigma X[t]$. In case that $\Sigma X[t] > 7$ and all the accessed RRAM cells are in HRS or LRS, the MAC output is saturated when read using the IA ADC decoder. Considering the sparsity of inputs and weights in CNNs, the probability that it occurs is sufficiently low and the saturation has a negligible impact on the performance of CNNs. Thus, we use a 4-bit ADC instead of a 5-bit ADC to achieve energy efficiency without loss of accuracy.



Die micrograph.

RRAM cells always have different resistances and sensitivities to WR pulses since the conductive filament in RRAM cells cannot be fully ruptured or formed in WR as illustrated in Fig. 4. Thus, the resistance of RRAM cells in a single-cycle WR has a wide distribution of resistances incurring erroneous MAC outputs. To tighten the resistance distribution and enable multi-bit encoding under a low R ratio, IWR is employed in the proposed RRAM macro. After every WR pulse, the readout circuit detects whether the resistance of an RRAM cell reaches the target resistance by estimating the resistance based on V.RBL. In WR iterations such as IRS encoding, weak SET and RESET pulses are applied (Fig. 3). While lowering the BL voltage over iterations, the resistance is finely adjusted and eventually it settles in the target range. The overall IWR operation is shown in Fig. 4.

The MAC outputs of the proposed ternary-weight CIM RRAM macro are demonstrated in Fig. 5. The measured V.RBL represents the MAC outputs over various combinations of resistance states including the IRS. V.RBL is measured after sweeping (1) 9 cells from HRS→IRS→LRS, (2) 9 cells from HRS→LRS, and (3) 6 cells from IRS→LRS with 3 cells fixed to HRS. The maximum difference of V.RBL over the weight sweeps is 4.75 mV. These results show that a stable and repeatable CIM readout is obtained as the cells are written from any of the three states to another. For each resistance state in RRAM cells, the forming processes and WRs are accordingly conducted with IWR. IWR enables a tight IRS distribution to prevent overlap of LRS or HRS distributions. The measured μ.IRS and σ.IRS is 4.85kΩ and 204.90Ω respectively over 100 IRS RRAM cells. Considering the wider distribution of HRS resistances, the target resistance of IRS cells in WR is set closer to the LRS regime but well outside the 3σ-window ((μ.IRS-μ.LRS)/(σ.IRS+σ.LRS)=9.16). Since IRS is susceptible to read-disturb due to the absence of an upper or lower bound of resistances, the tolerance of IRS RRAM cells for read-disturb is measured with 100 IRS cells under 20k RDs and 5 IRS cells under 2-million RDs. The IRS cells successfully retain the resistance with variations of 1kΩ toward the HRS regime under 20k RDs. The measured IRS resistances under 2-million RDs demonstrate that the IRS resistance does not invade the LRS or HRS regime under extreme repetitive RD scenario. For CIM, a peak energy efficiency of 118.44 TOPS/W is measured with a ternary RRAM array. Proposed techniques help a high-endurance ternary-weight RRAM macro achieve high algorithm-level accuracy across AI benchmarks with less than 5% loss of accuracy. The comparison table with the state-of-the-art CIM architectures [2-8] in Fig. 6 shows competitive metrics while addressing key challenges essential to multi-bit CIM RRAM macro.

Acknowledgements:

JY, MC, and AR were supported by the SRC under the C-BRIC under Grant 2777.005 and 2777.006, and the ASCENT under Grant 2776.037. The authors would also like to thank TSMC for technical discussions and chip fabrication support.

References:

[1] C. Nail *et al.*, IEDM, 2016.

[2] J.-H. Yoon *et al.*, ISSCC, 2021.

[3] W.-H. Chen *et al.*, ISSCC, 2018.

[4] C.-X. Xue *et al.*, ISSCC, 2019.

[5] C.-X. Xue *et al.*, ISSCC, 2020.

[6] W. Wan *et al.*, ISSCC, 2020.

[7] J. Wang et al., ISSCC, 2019.

[8] W. He et al., IEEE SSC-L, 2020.

978-1-7281-7581-2/21/$31.00 ©2021 IEEE

2021 IEEE Custom Integrated Circuits Conference (CICC) | 978-1-7281-7581-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/CICC51472.2021.9431412

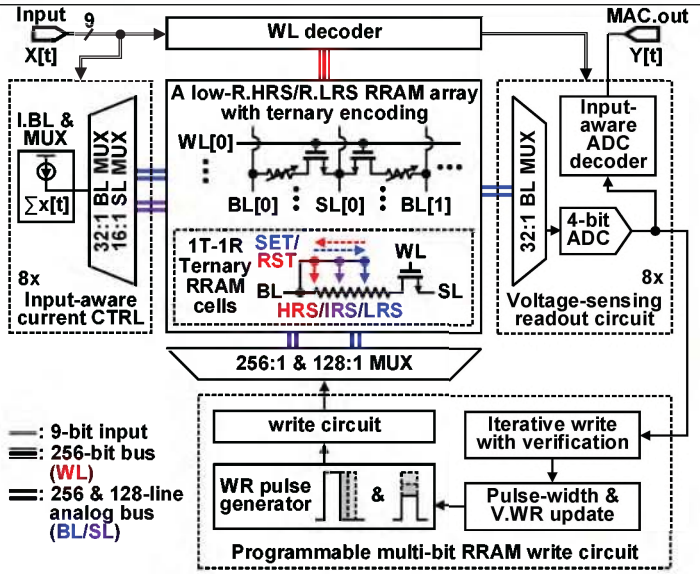


Fig. 1. Overall architecture of the proposed ternary-weight computing-in-memory RRAM macro with write verification

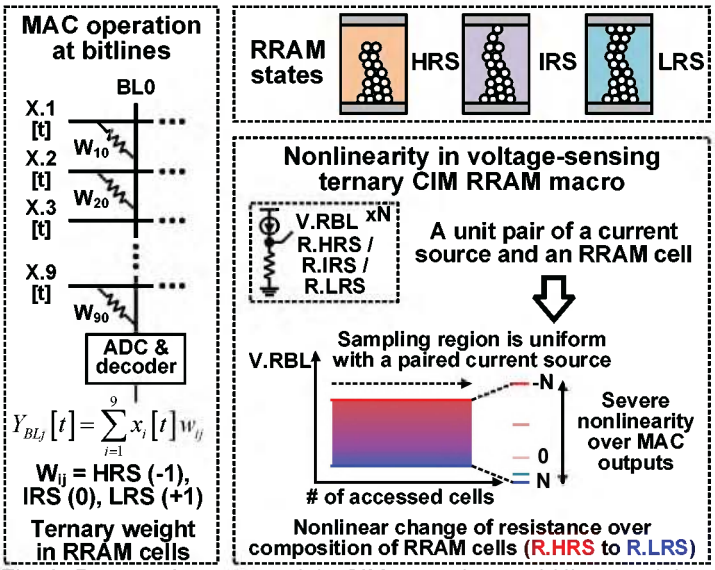


Fig. 2. Proposed ternary-weight CIM operations at bitlines and the nonlinearity in voltage-sensing ternary CIM RRAM macro

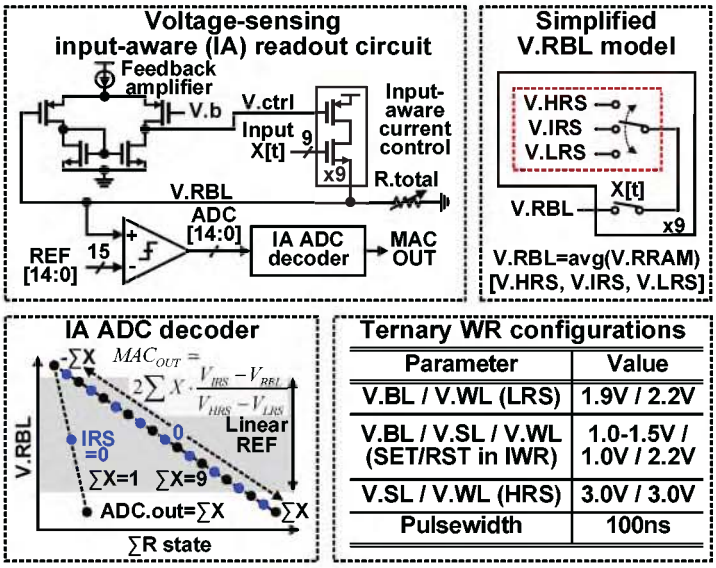


Fig. 3. Proposed voltage-sensing readout circuit for ternary-weight RRAM arrays and write configurations in ternary encoding

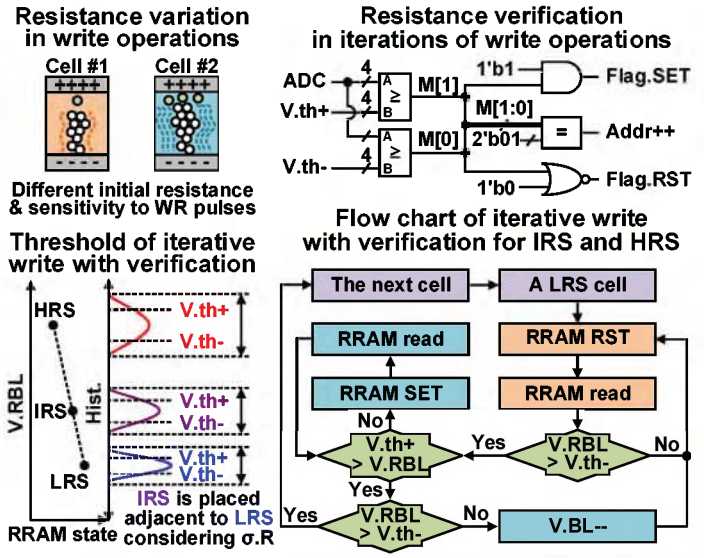


Fig. 4. Iterative write with verification in the proposed CIM architecture with multi-bit RRAM macro

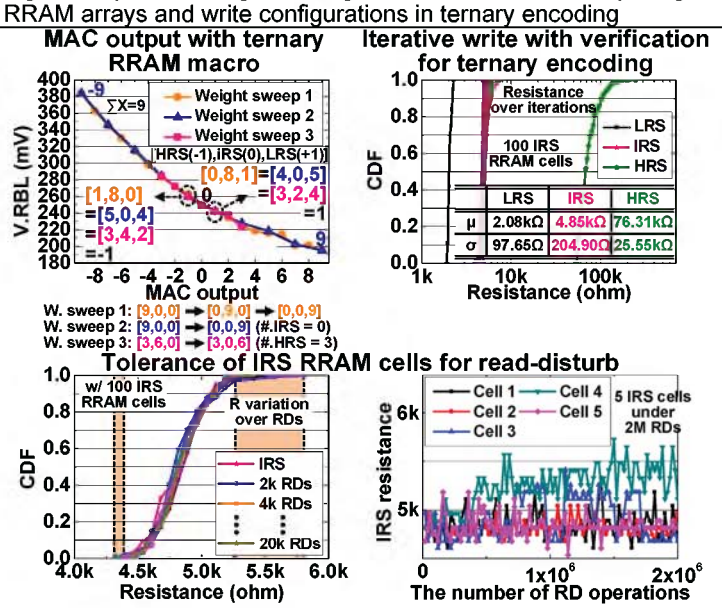
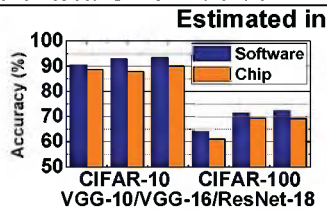


Fig. 5. Measured results of the MAC operation and multilevel encoding with write verification



Task	Network architecture	Loss of Acc.
CIFAR-10	VGG-10	-1.76 %
	VGG-16	-4.91 %
	ResNet-18	-3.39 %
CIFAR-100	VGG-10	-3.06 %
	VGG-16	-2.08 %
	ResNet-18	-3.04 %

Comparison table

	ISSCC 2018 [3]	ISSCC 2019 [4]	ISSCC 2020 [5]	ISSCC 2020 [6]	ISSCC 2019 [7]	SSC-L 2020 [8]	ISSCC 2021 [2]	This work
Technology	65 nm	55 nm	22 nm	130 nm	28 nm	40 nm	40 nm	40 nm
Memory	RRAM	RRAM	RRAM	RRAM	SRAM	RRAM	RRAM	RRAM
Supply	1.0 V	1.0 V	0.7-0.9 V	1.8 V	0.6-1.1 V	0.9 V	0.9 V	0.9 V
Sensing mode	Current	Current	Current	I&F	N/A	Voltage	Voltage	Voltage
A low R-ratio CIM structure	No	No	No	No	N/A	No	Yes	Yes
Multi-bit weight in a RRAM cell	No	No	No	No	N/A	Yes (2-bit)	No	Yes (three levels)
Resolution (Input / weight / output)	Not specified (output : 1-3 bits)	1-2 bit / 3 bits / 3 bits	1-4 bits / 2-4 bits / 6-11bits	1 bit / analog / 1 bit	Integer & floating point	1 bits / 2 bits / 1 bit	1-8 bits /1-8 bits / 20 bits	1 bit / 1.58 bits / 4 bits
Peak energy efficiency (TOPS/W)	19.2	53.17	121.38	148	0.55	51.4	56.67	118.44 w/ ternary weights

Fig. 6. Comparison table of state-of-the-art computing-in-memory RRAM macros