Edge Continual Training and Inference with RRAM-Gain Cell Memory Integrated on Si CMOS

Shuhan Liu[1]*, Robert M. Radway[1], Xinxin Wang[1], Filippo Moro[2], Jean-Francois Nodin[2], Koustav Jana[1], Shuting Du[3], Luke R. Upton[1], Wei-Chen Chen[1], Jian Chen[1], Haitong Li[3], Francois Andrieu[2], Elisa Vianello[2], Priyanka Raina[1], Subhasish Mitra[1], H.-S. Philip Wong[1]*
[1]Department of EE, Stanford University, CA, USA. [2]CEA-Leti, Univ. Grenoble Alpes, France. [3]Purdue University, IN, USA.
(*E-mail: shliu98@stanford.edu, hspwong@stanford.edu)

***Abstract*—**This research presents the design and experimental validation of a novel RRAM-Gain Cell joint memory to facilitate efficient continual learning in edge devices, addressing the challenges of resource-constrained environments while supporting adaptive AI model updates. HfO_2 RRAM and Indium Tin Oxide (ITO) gain cell are monolithically integrated on 130 nm Si CMOS technology, enabling high-speed training and low-standby-power inference for edge devices. High-bandwidth on-chip data transfer can have bandwidth that is 90× state-of-the-art HBM3E and 211× PCIe 7.0, enabled by high-density monolithic 3D interconnections and high-speed transfer circuits within the integrated joint memory macro. The ALD ITO FET exhibits positive V_{TH} of 0.67 V, excellent SS of 65 mV/dec, high on-current of 20 μA/μm, and low off-current of 5×10^{-18} A/μm, as extracted from > 5,000 s retention. The joint memory macro consumes 78% less standby power and 95% less training energy for MobileBERT compared to SRAM with iso-capacity.

I. Introduction

To achieve artificial general intelligence, it is essential to develop the capability for adaptive learning from the environment to handle real-world dynamics effectively. Continual learning (Fig. 1) [1] can accumulate knowledge without catastrophic forgetting, which is particularly beneficial for edge devices that frequently interact with environmental data. Implementing edge AI is challenging with current memory technologies, particularly when balancing the divergent requirements for training and inference within strict energy constraints. Previous work for edge models small enough to fit within the chip utilized separate SRAM and RRAM macros for training and inference respectively [2], resulting in area overhead and data transfer delays. In this work, we monolithically integrate two on-chip memory technologies, oxide semiconductor (OS) gain cell and Resistive RAM, into a compact joint memory cell on a Si CMOS platform (Fig. 2). The gain cell is used as training memory capitalizing on its high speed and infinite write endurance [3][4], while the RRAM serves as the inference memory due to its low standby power and non-volatility. High-bandwidth on-chip data transfer within the memory macro is achieved by high-density monolithic 3D vertical connections and a novel high-speed transfer circuit. The transfer bandwidth is 90× that of state-of-the-art (SoTA) HBM3E [5][6] with the same form factor. The joint memory macro offers 78% less standby power and 95% less MobileBERT training energy than SRAM, and 62% less inference energy than RRAM and SRAM separate macros.

II. RRAM-Gain Cell Memory

RRAM is well-suited for inference at the edge due to its low standby power and non-volatility, though its low endurance makes it unsuitable for training. Conversely, the gain cell is well-suited for a training memory, offering high speed and infinite endurance. Here we designed and fabricated an RRAM-gain cell joint memory, in which each cell has a 1T1R2OS structure (Fig. 2): 1T1R with a front-end of line (FEOL) Si FET and a middle-EOL RRAM, and 2T gain cell with 2 back-EOL OS FETs. The joint memory cell size is determined by the footprint of a 2T OS gain cell as it is stacked on top of the 1T1R. For each bit cell, the 2T gain cell and 1T1R share the write bitline (WBL) to implement in-memory-macro data transfer. Peripheral circuitry is shared between the gain cell and RRAM, further reducing overall macro area to 0.52× compared to separate memory macros.

A. Integration Experiment Details

W/L = 1000 nm/ 500 nm Si I/O transistor is fabricated in STMicroelectronics 130 nm CMOS technology with BEOL fabrication up to metal M4 on a 200 mm wafer. TiN/Ti/HfO_2/TiN RRAM stack is integrated between metal layers M4 and M5 at CEA-Leti, with 5 nm HfO_2 film as the RRAM switching material and 5 nm Ti as the top electrode (Fig. 3). Alignment marks are patterned with M5 for gain cell integration at Stanford. After etching the Si_3N_4 top passivation layer to open the via for connection, the ALD ITO FET gain cell is fabricated above the M5 and Si_3N_4 passivation layer (Fig. 3). Ti/Pt gate metal and Ni/Au source/drain metal are deposited by e-beam evaporation at room temperature. 10 nm of HfO_2 gate dielectric is deposited by ALD at 200℃, followed by an optimized 2 nm of ALD ITO film at 200℃ for a good channel-dielectric interface. This entire gain cell fabrication flow uses process temperature under 200℃, which is compatible with BEOL integration with Si CMOS [7].

B. Measurement of Integrated Joint Memory Cell

The ALD ITO FET fabricated on the CMOS chip shows excellent characteristics without observable integration degradation as in [3][4]. Fig. 4 and 5 show the ITO FET has positive V_{TH} of 0.67 V and excellent SS of 65 mV/dec, which enables the operation voltage to be < 2 V. It also has high on-current of 20 μA/μm and mobility of 20 $cm^2V^{-1}s^{-1}$, enabling the gain cell to achieve high-speed performance. The ALD ITO gain cell has measured retention of > 5,000 s with

extracted off-current of 5×10^{-18} A/μm under a standby WWL voltage of -0.5 V applied (Fig. 6). Fig. 7 shows the Si access transistor characteristics and RRAM DC switching curve with 80 consecutive cycles. The optimal set gate voltage is 2.2 ~ 2.4 V giving small standard deviation 10 μS (Fig. 8(a)), and the BL voltage is constrained < 2 V to avoid disturbing the gain cell. The resistance distribution for 100-cycle AC pulse programming and the programming condition are shown in Fig. 8(b). RRAM LRS is programmed to be 20 kΩ with a read current of 10 μA for a 0.2 V read voltage to match the read current of the gain cell and thus the read circuit can be shared. Fig. 9 shows small relaxation with the optimal set gate voltage of 2.3 V. Fig. 10 shows that the gain cell is not disturbed by the shared WBL with RRAM, due to small WBL-to-storage-node coupling and extremely low off-current of the ITO FET.

III. High-Bandwidth Data Transfer

Continual learning can adapt itself to different tasks through interaction with the environment data, which requires incremental updates to the weight parameters, and thus frequent data transfer between the training memory and inference memory. High-bandwidth on-chip data transfer can be achieved through high-speed in-memory-macro transfer circuit and high-density vertical monolithic 3D connections.

A. High-speed In-memory-macro Transfer Circuit

To implement in-memory-macro data transfer, a high-speed peripheral transfer circuit is designed by connecting the sense amplifier (SA) to the WBL circuit, which bypasses the memory I/O interface. Fig. 12 shows the circuit schematic and the timing diagram. Selection signals in the data I/O block enable switching between three operational modes: data input, data output, and internal data transfer. A two-stage current SA is adopted for high-speed sensing, and it is shared between the RRAM and the gain cell to reduce area overhead. In the first stage, the current difference between the read current signal and the reference current is converted to a voltage difference. In the second stage, the voltage difference is further amplified to a full-swing voltage signal. Meanwhile, data-output or data-transfer is also activated. The write domino circuit starts with the pre-charge signal disabled, and then transfers either SA-output or data-input to WBL with level shifting from 0.9V to 1.8V. Transistor M_{SA} and M_{SAB} are sized to W/L = 800 nm/250 nm, balancing the need for a strong drive to handle weak SA/SAB voltages with the requirement for low capacitance to enable fast sensing. They are placed above the selecting transistor to avoid Miller effect unbalance during switching due to inverted SA and SAB voltage. Simulation results in Fig. 11 with TSMC 40 nm technology show that the designed circuit transfers data with 1 ns delay. The transfer energy of 30.6 fJ/bit is approximately two orders of magnitude lower than HBM, which typically consumes around 3-4 pJ/bit with > 60% from the data movement itself [8].

B. High-density Monolithic 3D Vertical Connection

With the high-speed transfer circuit, the data transfer rate is 1 Gbps for each BL. Assume the RRAM-GC macro has the same array configuration (subarray size of 512×64) and macro size (0.85 mm^2 for 3Mb) in [9]. SoTA HBM3E [5][6] delivers transfer rate of 9.8 Gbps per pin and bandwidth of 1.2 TB/s with 1024 pins for a form factor of 11mm×11mm. With the same form factor, the 52 MB joint RRAM+GC memory macro has 866,304 parallel BLs and can achieve 108 TB/s bandwidth, which is 90× higher than SoTA off-chip HBM3E, 211× higher than SoTA PCIe 7.0 (Fig. 13(b)). Despite the lower transfer rate per pin (Fig. 13(a)), the high bandwidth results from leveraging the high-density connections with on-chip monolithic 3D integration.

IV. Edge Continual Training and Inference

Edge devices are strictly energy-constrained, and not activated very often, so standby power is critical. RRAM has low standby power and can even be powered-off [10] with the data maintained in the non-volatile memory. With weights transferred from gain cell training memory to RRAM inference memory, the gain cell doesn't need to be refreshed, and thus doesn't add to the standby power. Even compared to SRAM with power-reduction techniques [11], the RRAM-GC joint memory still has 78% standby power saving (Fig. 14). During MobileBERT [12] training, the total memory energy is 94% less (Fig. 15(b)) compared to SRAM [13], attributed to small memory leakage energy during computation operations. If 50% weight adaptivity is assumed for continual learning [14], the training energy can be further reduced, as the dominant part of the energy consumption (write energy) decreases (Fig. 15). During MobileBERT inference, gain cell can be reused as buffer memory (Fig. 16(a)). Based on ScaleSim [15], The RRAM-GC joint memory has 47% energy saving compared with separate RRAM+GC macros, and 62% energy saving compared with separate RRAM+SRAM macros in both weight-stationary and output-stationary dataflow, due to low gain cell leakage energy (Fig. 16).

V. Conclusion

The RRAM-gain cell joint memory synergistically combines the non-volatility of RRAM and high endurance of gain cell without incurring area or delay overhead thanks to the monolithic 3D integration. Using gain cell for training and RRAM for inference, this joint memory has high data transfer bandwidth, low standby power, and low active energy for continual learning. This work offers a potential single joint memory solution for edge AI and motivates further research on combining application-domain specific device technologies using monolithic 3D integration that underpins the N3XT 3D vision for future computing hardware technologies [7].

Acknowledgement: Supported in part by DoD/Eccalon, SRC PRISM Center, SRC CHIMES Center, Stanford NMTRI, Stanford SystemX Alliance, SGF, and ERC consolidator grant DIVERSE (101043854). The authors also thank Kasidit Toprasertpong, Jeffrey Yu, Tathagata Srimani, and Sumaiya Wahid for discussions and assistance.

Reference: [1] Hadsell, Raia, et al. Trends in cognitive sciences (2020). [2] Kosta, Adarsh, et al. 2022 DATE. [3] Liu, Shuhan, et al. 2023 IEDM. [4] Liu, Shuhan, et al. IEEE TED (2024). [5] Samsung [6] Micron [7] Radway, R. M., et al. 2021 IEDM. [8] Kim, Kyungryun, 2024 VLSI-SC. [9] Upton, Luke R., et al. ESSCIRC 2023. [10] Wu, Tony F., et al. 2019 ISSCC. [11] Teman, Adam, et al. IEEE JSSC (2011). [12] Sun, Zhiqing, et al. 2020 ACL. [13] Chiu, Yi-Wei, et al. IEEE TCS (2014). [14] Adel, Tameem, et al. 2020 ICLR. [15] Samajdar, Ananda, et al. 2020 ISPASS.

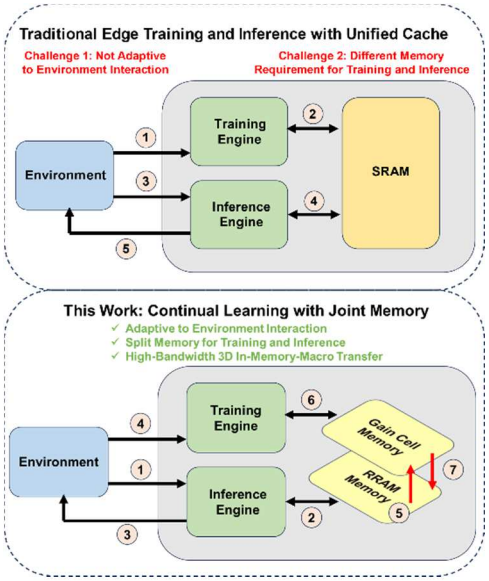

Fig.1 Continual learning can adaptively fine-tune the model based on the inference feedback with the re-training cycle 1-7. Joint memory facilitates continual learning at the edge by using different memories for training and inference and high-bandwidth data transfer.

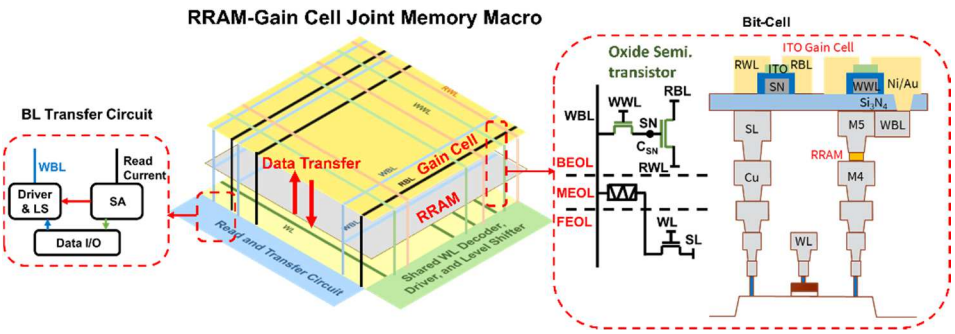

Fig.2 RRAM-Gain Cell (GC) joint array with 3D-stacked bit-cell, in-memory-macro data transfer circuit, and shared peripheral circuit.

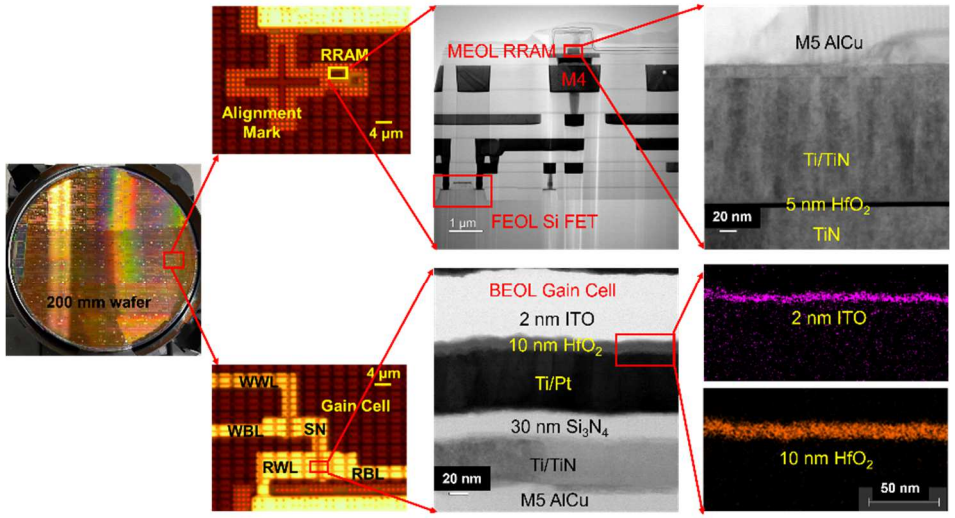

Fig.3 Photo of the 200 mm wafer and optical microscope, TEM, and EDS images of fabricated RRAM-gain cell joint memory. Top row shows the 1T1R structure and bottom row shows the 2T OS gain cell integrated on the top of the same chip.

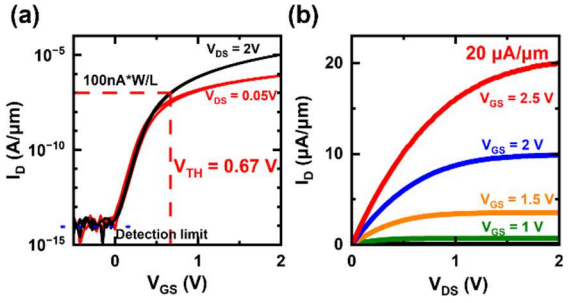

Fig.4 Measured transfer and output curve of L = 1 μm ALD ITO FET fabricated on the integrated chip. Positive V_{TH} of 0.67 V and high on-current are achieved.

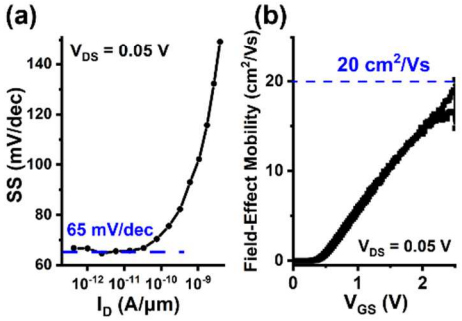

Fig.5 Extracted (a) SS and (b) field-effect mobility from measured transfer curve of L = 1μm ALD ITO FET

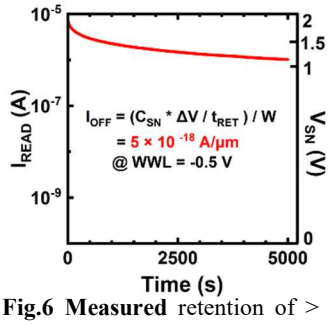

Fig.6 Measured retention of > 5,000 s of gain cell fabricated on the integrated chip.

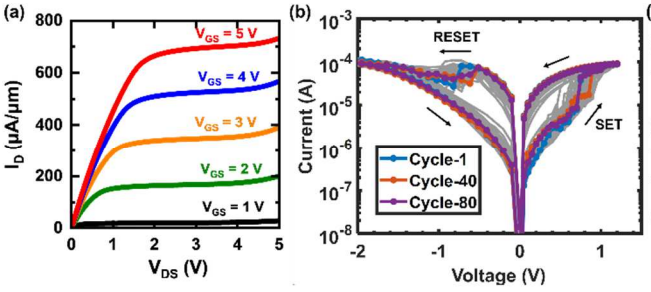

Fig.7 Measured (a) Si access transistor output curve and (b) 80-cycle RRAM DC switching curve for RRAM in joint memory.

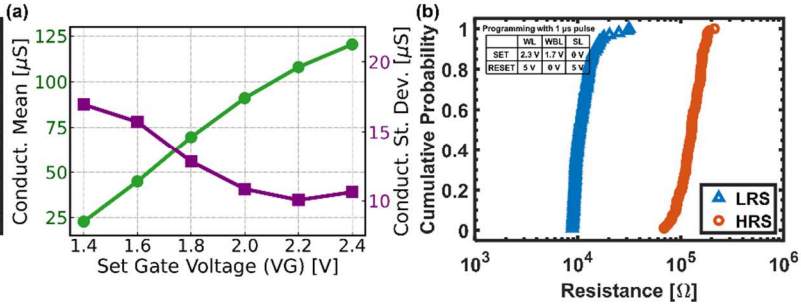

Fig.8 (a) Set gate voltage optimal range is 2.2-2.4 V. (b) **Measured** resistance distribution after 100-cycle AC programming of RRAM of the joint memory. Insert: voltage pulse programming condition.

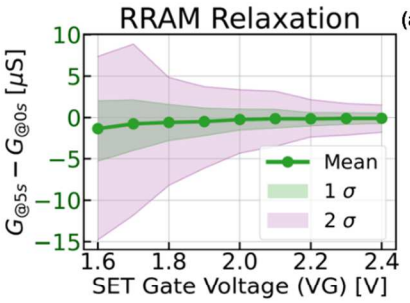

Fig.9 Measured small RRAM relaxation effect with set gate voltage > 2.2 V.

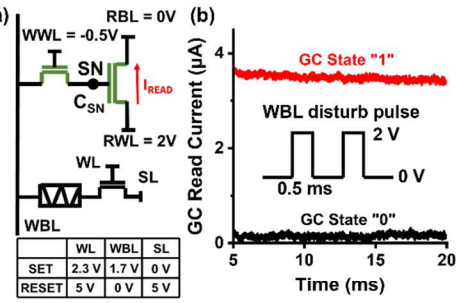

Fig.10 Measured (b) GC read current no transient change under WBL disturb pulse from (a) RRAM programming and GC write on the same column for the joint memory.

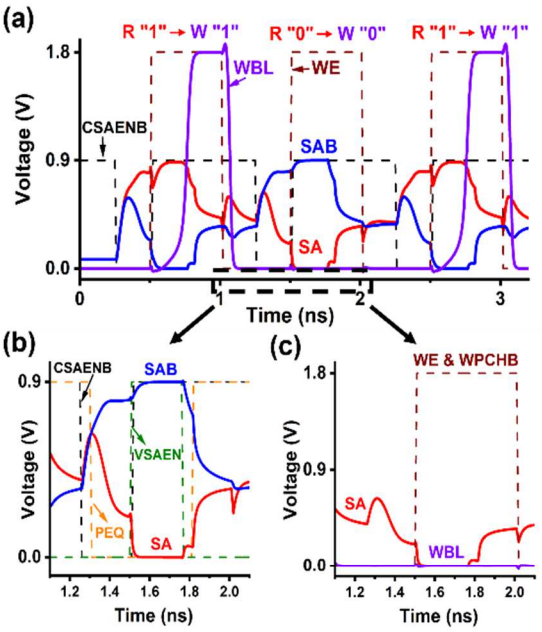

Fig.11 Simulation results of (a) three consecutive operations of data transfer for different data with each in 1 ns. Details of (b) read and (c) write-back within the 1 ns. Simulated with TSMC 40 nm.

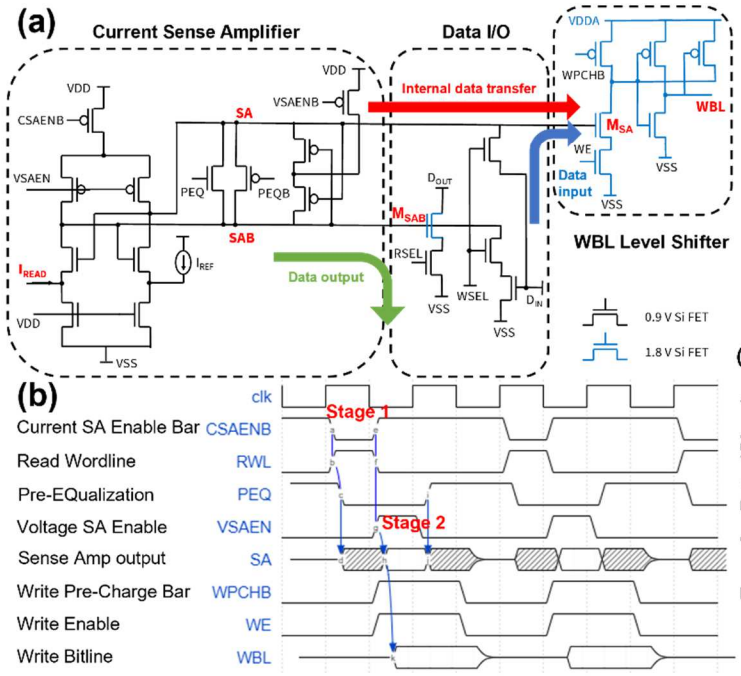

Fig.12 (a) Circuit schematic and (b) two-stage timing diagram of the designed transfer circuit with sense amplifier, data I/O, and level shifter. Data transfer is performed at the second stage.

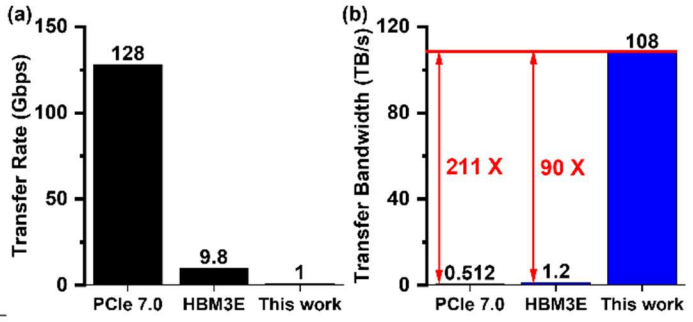

Fig.13 Despite (a) the lower transfer rate per pin, RRAM-GC joint memory still has (b) higher transfer bandwidth by leveraging the high-density connections with on-chip monolithic 3D integration.

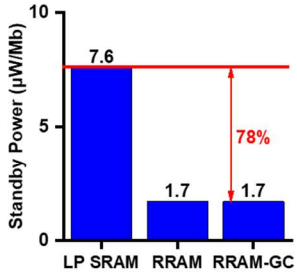

Fig.14 78% lowered standby power compared to low-power (LP) SRAM with power reduction techniques.

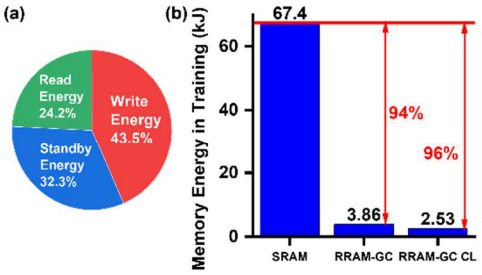

Fig.15 (a) Energy breakdown for the RRAM-GC joint memory during MobileBERT training. (b) 94% - 96% (w/o and w/ continual learning) memory energy is reduced compared to SRAM.

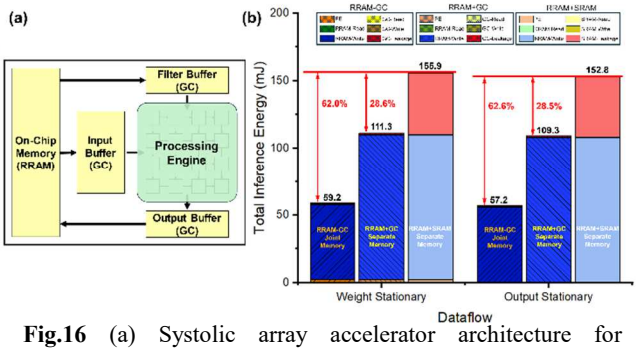

Fig.16 (a) Systolic array accelerator architecture for MobileBERT inference. (b) The RRAM-GC joint memory has 47% and 62% energy saving compared with separate RRAM+GC macros and RRAM+SRAM macros, respectively