Analog Multilevel eDRAM-RRAM CIM for Zeroth-Order Fine-tuning of LLMs

Mufeng Chen, Luqi Zheng, Jian-Yu Lin, Peide D. Ye, and Haitong Li
Elmore Family School of Electrical and Computer Engineering, Purdue University
West Lafayette, USA
haitongli@purdue.edu

2025 IEEE International Memory Workshop (IMW) | 979-8-3503-6298-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/IMW61990.2025.11026966

***Abstract*—Zeroth-order fine-tuning eliminates explicit backpropagation and reduces memory overhead for large language models (LLMs), making it a promising approach for on-device fine-tuning tasks. However, existing memory-centric accelerators fail to fully leverage these benefits due to inefficiencies in balancing bit density, compute-in-memory capability, and endurance-retention trade-off. We present a reliability-aware, analog multi-level-cell (MLC) eDRAM-RRAM compute-in-memory (CIM) solution co-designed with zeroth-order optimization for language model fine-tuning. An RRAM-assisted eDRAM MLC programming scheme is developed, along with a process-voltage-temperature (PVT)-robust, large-sensing-window time-to-digital converter (TDC). The MLC-eDRAM integrating two-finger MOM provides $12\times$ improvement in bit density over state-of-the-art MLC design. Another $5\times$ density and $2\times$ retention benefits are gained by adopting BEOL In_2O_3 FETs.**

***Index Terms*—Compute-in-memory (CIM), eDRAM, RRAM, MLC, oxide semiconductors, LLM fine-tuning.**

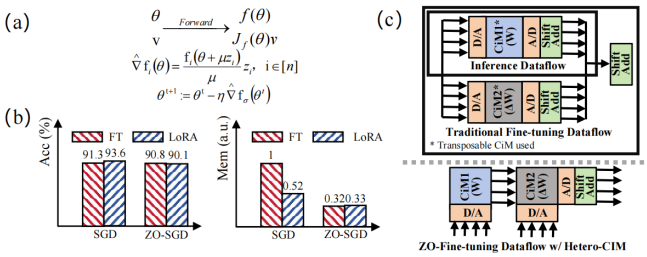

Fig. 1. (a) Zeroth-order optimization: A single forward pass computes both the output and the Jacobian-vector product via weight perturbation, enabling efficient gradient updates without backpropagation. (b) Accuracy and memory cost comparison of zeroth-order (ZO) vs. backpropagation-based fine-tuning for the OPT-1.3B LLM on the SST-2 dataset, evaluated under full fine-tuning (FT) and low-rank adaptation (LoRA). (c) Fine-tuning tasks demand higher memory and peripheral overhead compared to inference. Overview of the algorithm-architecture-technology co-design solution in this work.

I. Introduction

As large language models (LLMs) continue to proliferate, the need for efficient on-device fine-tuning has grown increasingly critical to address domain adaptation and privacy requirements across broad applications. Conventional backpropagation-based fine-tuning inflates memory overhead with intermediate activations and gradients, leading to excessive on-chip and off-chip memory accesses. By contrast, zeroth-order (ZO) forward gradient methods, as recently demonstrated on both traditional language models and LLMs [1], remove explicit backpropagation steps and thereby reduce memory usage while maintaining comparable accuracy. As depicted in Fig. 1(a), zeroth-order forward gradient optimization employs weight perturbation to generate model activation outputs, and the Jacobian vector product (JVP) needed for approximate weight gradients can be computed in a single forward pass—making it amenable to energy-efficient hardware implementations [2].

While the zeroth-order scheme serves as the algorithmic foundation for the algorithm-hardware co-design in this work, LLM fine-tuning remains memory-bound and challenging for hardware acceleration. This highlights the need for tailored memory-centric designs that can handle memory-intensive updates and computations efficiently and reliably. Compute-in-memory (CIM) addresses this need by co-locating computation and storage while lowering data movement and enabling parallel operations. However, CIM designs face major challenges for LLM fine-tuning applications. These challenges include **(1)** supporting dataflows beyond simple weight-stationary, which complicates endurance requirements in non-volatile memory (NVM) based designs, and **(2)** storing and handling transposed weights on chip. Although SRAM-based CIM offers low latencies, it often lower bit density, while RRAM-based CIM improves storage density at the cost of higher write energy and endurance constraints. Prior studies [3]–[5] have explored hybrid CIM macros with NVM and SRAM. These designs may not produce optimal bit density as CIM macros due to **(1)** area overhead with the separate CIM kernels using each memory technology (e.g., RRAM and SRAM) and dedicated peripherals, and **(2)** limited MLC capabilities. On the other hand, high-density embedded DRAM (eDRAM) gains new interests in eDRAM-CIM designs [6], [7] leveraging coarse–fine write steps on bitlines, yet they could incur longer write times with multi-stage programming and limited retention in large arrays due to low internal capacitance from shared read/write ports.

In this work, we address these challenges with an algorithm-hardware co-design solution for LLM fine-tuning, by uniting the zeroth-order forward gradient scheme with a heterogeneous eDRAM-RRAM CIM architecture. We perform fine-tuning experiments with both traditional stochastic gradient descent (SGD) and zeroth-order SGD on an Nvidia A800 GPU, using Open Pretrained Transformer (OPT-1.3b) on the

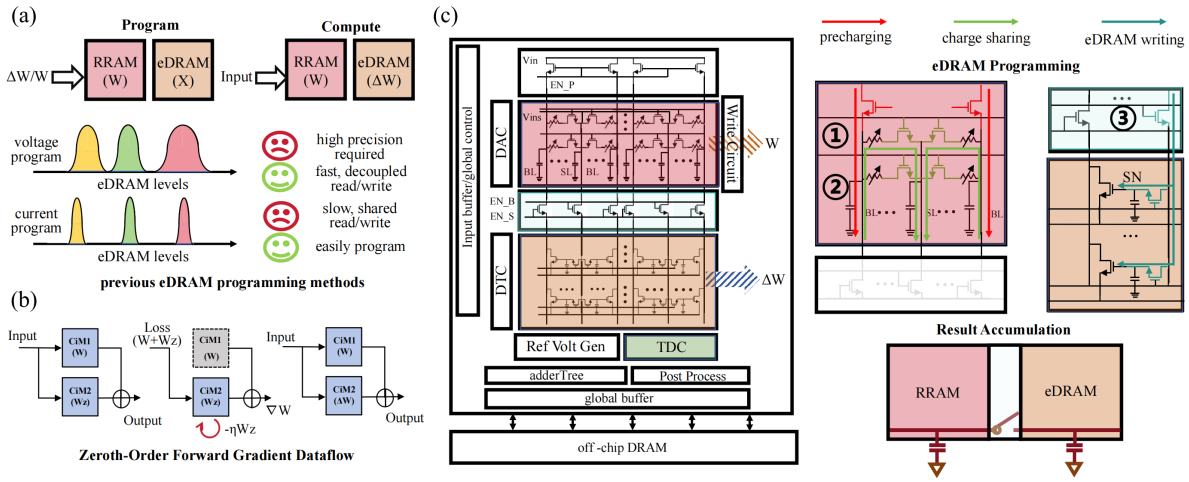

Fig. 2. Overview of the analog MLC eDRAM-RRAM CIM architecture co-designed with the zeroth-order fine-tuning scheme. (a) Guidelines for model mapping on eDRAM/RRAM technologies, and challenges associated with traditional eDRAM MLC programming schemes. (b) CIM-based zeroth-order forward gradient computation utilizes RRAM for static operations and eDRAM for dynamic updates, addressing both RRAM endurance and eDRAM retention limitations. (c) Circuit architecture and operation principles (MLC programming and result accumulation) for on-chip zeroth-order fine-tuning. Our RRAM-assisted eDRAM MLC programming approach mitigates voltage sensitivity and enables efficient fine-tuning.

Stanford Sentiment Treebank 2 (SST-2) dataset. As shown in Fig. 1(b), up to 3× memory savings with only ∼0.5% accuracy loss are obtained. Drawing upon these results, our CIM design implemented in 40 nm process integrates multi-level-cell (MLC) eDRAMs and RRAMs within a single CIM macro, eliminating redundant peripherals while enabling transposed-weight-free computation, as illustrated in Fig. 1(c). An RRAM-assisted eDRAM MLC programming scheme is developed, enabling robust programming of 16 levels for eDRAMs handling dynamic operations. BEOL In_2O_3 FETs are further exploited for bit density and retention enhancement.

II. Analog MLC eDRAM-RRAM CIM Design

A. CIM Architecture for Reliability-Aware Fine-tuning

In our reliability-aware eDRAM-RRAM CIM architecture, RRAM cells handle static weights while endurance-unlimited eDRAM cells are tailored for dynamic operations that do not require long retention times, as illustrated in Fig. 2 (a). Previous MLC eDRAM designs typically rely on current programming mode with shared read/write ports to mitigate PVT variation in naive voltage-based programming, often at the cost of increased write latency [6]. To address this, we develop an RRAM-assisted eDRAM programming scheme within the heterogeneous analog CIM macro. After programming the perturbations, the added weights are computed alongside the pre-trained weights stored in MLC RRAMs. The dataflow for forward gradient computation is illustrated in Fig. 2(b). The forward pass weight and the perturbation, which is generated using a fixed random seed, are multiplied with the input and summed to produce both the output and Jacobian-vector product. The loss function then multiplies the regenerated perturbation to compute the local weight gradient, which is subsequently used for weight updates. In the testing phase, the weight and weight updates are combined to compute the forward pass. The heterogeneous CIM macro mainly consists of a shared precharger, integrated 2T-2R RRAMs and gain-cell eDRAMs in a dense array architecture, a digital-to-time converter, a reference voltage generator for PVT compensation, and a time-to-digital converter (TDC), as shown in Fig. 2(c).

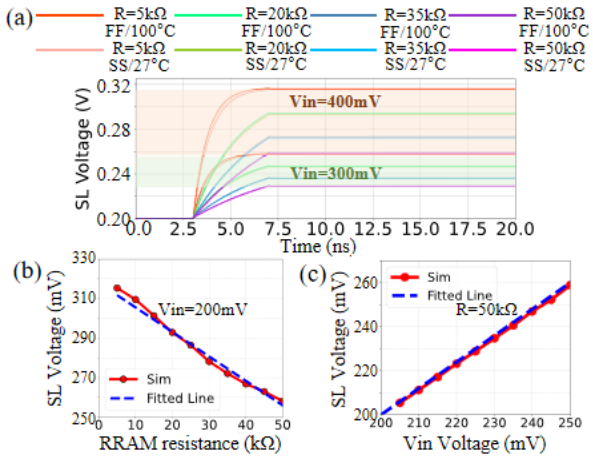

Fig. 3. RRAM-assisted eDRAM MLC programming waveforms based on full circuit simulations in 40 nm. (a) DC-current-free eDRAM MLC programming waveforms for Vin = 300 mV and 400 mV, where Vin is the precharge voltage. (b) Linearity analysis showing the correlation between SL output voltages and RRAM resistance levels. (c) Linearity analysis showing SL output voltages under varying Vin biases.

B. RRAM-Assisted eDRAM MLC Operations

During the RRAM programming phase, two eDRAM access transistors are deactivated while RRAM undergoes a standard write-verify process [8]. Next, the bitline (BL) and select line (SL) access transistors are turned off, and the RRAM BL and SL are precharged to V_{in} and V_{ins}, respectively. Once precharging is complete, the precharge transistor turns off, and

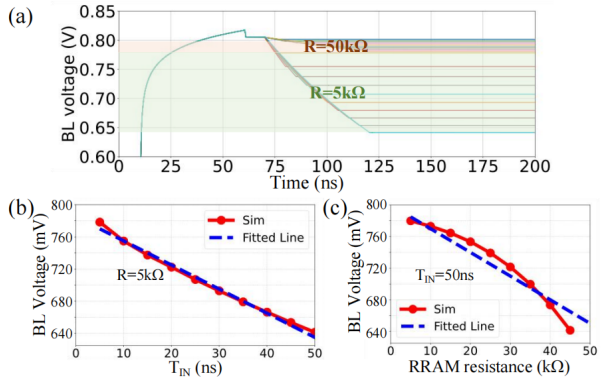

Fig. 4. Time-domain dot-product computation with MLC eDRAM-RRAM. (a) eDRAM BL waveforms during computation, with RRAM resistance = 5 kΩ and 50 kΩ. (b) Linearity of eDRAM BL outputs against input pulse durations. (c) Correlation of eDRAM BL outputs and RRAM resistances under the worst case scenario.

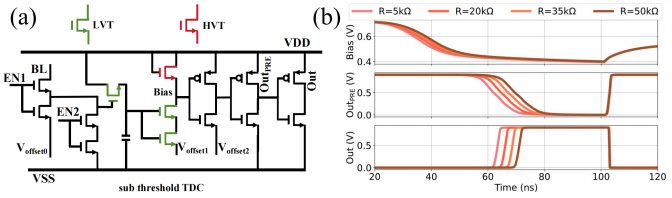

Fig. 5. PVT-robust subthreshold TDC design and simulations. (a) The subthreshold TDC design compensates for PVT variations using an offset voltage group while maintaining a large sensing window. (b) Operating waveforms from simulations. After BL input biasing and subthreshold amplification, the signal is then biased at the 'Bias' point. The inverter chain enhances the sensing window, sharpens the rising edge, and generates an output at the 'Out' point for 4 levels of RRAM resistances from $5k\Omega$ to $50k\Omega$.

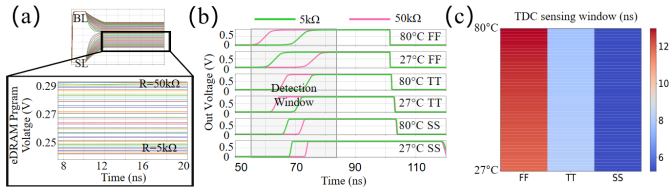

Fig. 6. (a) PVT-robust eDRAM MLC programming using MLC RRAMs with resistance ranging from $5k\Omega$ to $50k\Omega$ at $3k\Omega$ intervals, resulting in 16-level (4-bit) eDRAMs. (b) Output waveform at the Out point, where the rising edge between the lowest ($5k\Omega$) and highest ($50k\Omega$) RRAM resistances defines the sensing window. (c) Heatmap of the TDC sensing windows (ns) across different PVT corners from $27°C$ to $80°C$.

the RRAM word line (WL) is activated for charge transfer. The SL transistor then opens to precharge the entire SL. Following this, the eDRAM access transistor activates to program eDRAM cells on a row-by-row basis. This scheme essentially capitalizes a DC-current-free charge-sharing process between heterogeneous memory cells in an integrated array. During result accumulation, the lower section of the eDRAM SL is precharged, while the upper RRAM SL and BL are reset to VSS. With BL and SL switch transistors turned off, the heterogeneous CIM operates as two independent CIM macros. After separate vector-multiply computations, the SL switch is activated, and the final result is obtained via charge-sharing accumulation and the eDRAM BL capacitance decides the update magnitude. The operations are summarized in Fig. 2(c). All the follwing simulations and results are based on 40 nm technology.

Ensuring linearity in eDRAM MLC programming and the dot-product operations is essential to reduce calibration requirements and enable scalable, accurate multi-bit computations. Fig. 3(a) shows the simulated eDRAM program voltages across different V_{in} values, with robustness against PVT variations. Fig. 3(b) and (c) further provides the linearity analysis, showing the correlation between SL output voltages against different RRAM resistance levels and Vin biases. After the MLC programming, the dot-product computation in memory is analyzed with time-domain simulations as shown in Fig. 4. The computation waveform with 4-bit eDRAM weights and 4-bit time-domain inputs are captured in Fig. 4(a). RRAM resistances of 50 kΩ and 5 kΩ correspond to weight data of "1" and "15", respectively. From the full circuit simulations from programming to computation phases, high linearity is obtained consistently from the proposed design and RRAM-assisted MLC scheme. These dynamic operations do not pose endurance challenges for either RRAMs or eDRAM.

C. PVT-Robust MLC Programming and Sensing

To maintain high bit density, we develop a PVT-robust, compact TDC with a wide sensing window (Fig. 5). The subthreshold TDC (Fig. 5(a)) integrates a biasing stage, a subthreshold charger, and a wave-shaping inverter chain. The proposed TDC takes advantage of the low RRAM read voltage. Together with the high sensitivity and ultra-low current from the subthreshold circuit, the sensing window is large and the charging capacitance is low (∼3fF). A DAC-based offset voltage stabilizes the detection window at 25 ns immune to PVT variations. The waveform in Fig. 5(b) shows how BL voltage switches the subthreshold transistor, while the inverter chain sharpens ''Bias" voltages to generate the output signal which can be further converted to digital signal with standard time-sampling module.

Fig. 6(a) presents the circuit simulation results of DC-current-free, high-precision eDRAM MLC programming, assisted with MLC RRAM, spanning from the $27°C$ SS to the $80°C$ FF corners. To enhance TDC stability, a pair of calibration voltages (0 and 200 mV) are applied. Further, Fig. 6(b) confirms that proper offset adjustment achieves a 25 ns detection window and a 20 ns sensing window, with the PVT dependency shown in Fig. 6(c).

Finally, the layouts for the eDRAM and RRAM cells used in the hetero-CIM design are shown in Fig. 7. Integrating a two-finger MOM in parallel with gain cell's internal capacitance enhances the capacitance density without increasing cell area. For 2T-2R RRAMs, a shared-SL layout is adopted.

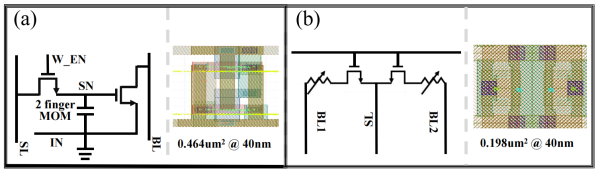

Fig. 7. (a) Layout of the gain-cell eDRAM with a two-finger MOM for enhanced retention. (b) Layout of shared-SL 2T-2R RRAM.

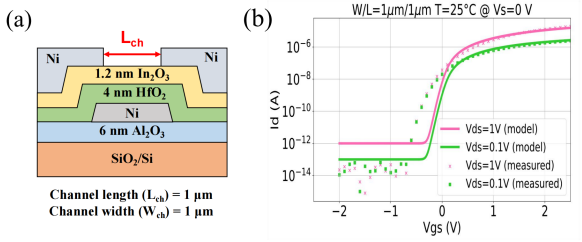

Fig. 8. BEOL In_2O_3 FET is utilized for the Si + In_2O_3 eDRAM design. (a) Device stack for the ALD In_2O_3 transistor. (b) Measured I-V characteristics used for compact modeling in circuit simulations. Devices with 1 μm channel width/length were used for leakage characterization purpose.

III. BEOL-In_2O_3-Augmented MLC eDRAM

Our baseline silicon eDRAM design exploits both the capacitance enhancement from MEOL (two-finger MOM), as well as a new MLC programming scheme from BEOL (RRAM-assisted). Furthermore, we explore In_2O_3 FET-Augmented MLC eDRAM design by replacing the read transistor in the gain cell, while leveraging the high ON/OFF, low leakage, and CMOS compatibility of ultra-thin-channel ALD In_2O_3 FETs [9]. The oxide-semiconductor transistor stack is shown in Fig. 8 (a). We build a device model of the fabricated In_2O_3 FET samples [9] using BSIM4 to conduct circuit analysis, with measured and modeled I-V characteristics shown in Fig. 8 (b). The root mean square errors (RMS) for the I-V data at V_{ds} voltages of 0.1 V and 1 V are 5.57% and 5.71%, respectively.

The Si+In_2O_3 eDRAM exhibits an additional $\sim 2\times$ improvement in MLC retention time (Fig. 9 (a)), and another $\sim 80\%$ savings in cell area due to the BEOL-stacked In_2O_3 FET serving as the read transistor. The retention time spectrum of full MLC levels for Si+In_2O_3 eDRAM and Si CMOS eDRAM are shown in Fig. 9 (b), with the box plot in Fig. 9 (c) capturing the statistics. The retention enhancement comes from the increased capacitance density and the lower leakage owing to the BEOL-stacked In_2O_3 FET with the MEOL-located two-finger MOM (M3-M5 metal layers leveraged). Table 1 summarizes two MLC-eDRAM designs in this work. The bit density is $\sim 12\times$ higher compared to that of a state-of-the-art MLC design [6]. The density and retention are further improved by $\sim 5\times$ and $\sim 2\times$ with the BEOL In_2O_3 FET.

IV. Conclusion

This work presents a novel analog eDRAM-RRAM CIM solution co-designed with zeroth-order optimization for efficient LLM fine-tuning. The heterogeneous CIM architecture features DC-current-free eDRAM MLC with a robust RRAM-assisted programming scheme, charge-sharing eDRAM-RRAM accumulation, and a PVT-robust TDC design in 40 nm. Our work shows how MEOL-BEOL integration (two-finger MOM, BEOL In_2O_3 FETs) and FEOL-BEOL interaction (RRAM-assisted eDRAM MLC) can be jointly exploited for efficient memory-centric computing.

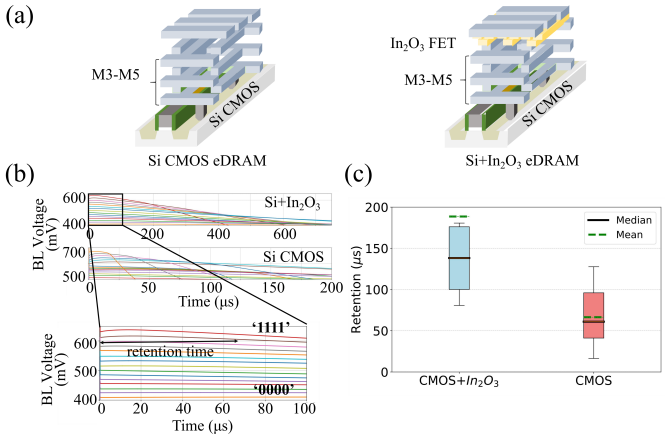

Fig. 9. (a) Metal stack illustrations of Si CMOS eDRAM and Si+In_2O_3 eDRAM designs. (b) Retention time spectrum from MLC eDRAMs comparing Si+In_2O_3 stack and Si CMOS stack, with the zoomed-in plot showing retention behaviors of the 16 levels. $\sim 2\times$ improvement in retention is gained with BEOL In_2O_3 integration. (c) Box plot of MLC retention time spectrum.

TABLE I
Benchmarking MLC-eDRAM Designs for CIM Applications.

	This work		CICC'23 Song *et al.* [6]	JSSCC'24 Zhan *et al.* [10]	ISSCC'24 He *et al.* [11]
	Si CMOS	Si + In_2O_3 FETs			
cell type	TF MOM	TF MOM	MOM		
Technology	40 nm	40 nm + In_2O_3	65 nm	28 nm	28 nm
Bit/cell	4	4	4	1	1
Retention (μs)	16–128	80–614	400	35.5	-
Cell density* ($bit/\mu m^2$)	8.475	40.322	0.667	5.618	2.353

* Storage density defined as (bit per cell)/ (cell area) : $\frac{\#bits}{A}$

Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Award No. 2425498 with industry partners as specified in the Future of Semiconductors (FuSe2) program, and in part by the UPWARDS for the Future Network program with funding from the NSF (Award No. 2329784), Micron Technology, and TEL.

References

[1] Y. Zhang *et al.*, ICML, 2024, pp. 59173–59190.
[2] Hinton, Geoffrey E. ArXiv abs/2212.13345, 2022.
[3] A. Kosta *et al.*, DATE, 2022, pp. 88-91.
[4] A. S. Lele *et al.*, JSSC, vol. 59, no. 1, pp. 52-64, 2024.
[5] L. Bagheriye, *et al.*, TCAS-II, vol. 65, no. 11, pp. 1708-1712, 2018.
[6] J. Song *et al.*, CICC, 2023, pp. 1-2.
[7] J. Song *et al.*, ISSCC, 2024, pp. 490-492.
[8] J. -H. Yoon *et al.*, ISSCC, 2021, pp. 404-406.
[9] J. -Y. Lin *et al.*, IEDM, pp. 1851-1854, 2024.
[10] Y. Zhan *et al.*, JSSC, vol. 59, pp. 3866-3876, 2024.
[11] Y. He *et al.*, ISSCC, 2024, pp. 578-580.