# A Calibration-Free 15-level/Cell eDRAM Computing-in-Memory Macro with 3T1C Current-Programmed Dynamic-Cascoded MLC achieving 233-to-304-TOPS/W 4b MAC

Jiahao Song[1], Xiyuan Tang[1], Haoyang Luo[1], Haoyi Zhang[1], Xin Qiao[1], Zixuan Sun[1], Xiangxing Yang[2], Yuan Wang[1], Runsheng Wang[1], Ru Huang[1]

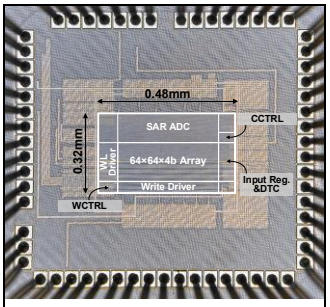
[1]Peking University, Beijing, China; [2]pSemi Corporation, Austin, TX

2023 IEEE Custom Integrated Circuits Conference (CICC) | 979-8-3503-9948-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/CICC57935.2023.10121207

The smart edge nodes require efficient matrix-vector multiplications for local deep neural network (DNN) inference. Benefiting from its high density and CMOS compatibility, the eDRAM-based computing-in-memory (CIM) [1-4], especially with multi-level cells (MLCs) [4], attracts rising attention. However, the performance of prior MLC-eDRAM CIM is severely limited by the inconsistency of weight representations during the programming and computing: weights are programmed as fixed voltages while transistor currents are used for computation. Thus, the programming of MLCs requires calibration due to the nonlinear transistor I-V, which can be extremely complicated in the presence of $V_{TH}$ variations. Furthermore, the computing precision is severely degraded by $V_{TH}$ variations when small computing currents are used for high parallelism. To fundamentally surmount this dilemma, we propose *the first current-programming eDRAM CIM* that unifies the weight programming and computing in the current domain. The enabling technique is a novel 3T1C eDRAM cell (Fig. 1, top right). It confers several key merits: 1) the cell is programmed by the weight current directly with the self-calibrated voltage generated on the storage capacitor; it essentially stores the weight current instead of a fixed voltage, thus mitigating $V_{TH}$ variation and nonlinear transistor I-V impacts; 2) a dynamic-cascoded read structure is proposed to significantly reduce the $V_{BL}$ sensitivity while not requiring any bias voltage; 3) thanks to the accurately programmed cell, it supports MLC operation (8 current levels) without any calibration, largely increasing density; 4) a voltage-current two-step programming scheme significantly boosts the sub-μA current-weight writing speed. Combining these merits, the proposed eDRAM cell is naturally immune to transistor-level nonidealities, thus allowing a small LSB weight current of only 100nA. A 4b CIM cell composed of 2 MLCs is developed to support 4b-signed weights. It contains 15 current levels ranging from -700nA to 700nA. Fabricated in a 65nm CMOS, the prototype achieves the highest macro-level 4b-MAC energy efficiency of 233-305TOPS/W among eDRAM CIMs.

Fig. 1 depicts the basic concept of the proposed current-programming eDRAM cell. The conventional memory cell operates with a voltage-programmed and current-read style. However, due to various transistor nonidealities, it is unsuitable for CIMs targeting both high energy efficiency and high precision. This issue can be addressed by the proposed current-programmed cell. In the programming phase, the read transistor is diode-connected. Once the programming current $I_{write}$ is applied, a self-calibrated $V_{GS}$ will be generated and stored in the cell. In contrast to the fixed voltage stored in conventional cells, this self-calibrated $V_{GS}$ naturally tackles eDRAM cells' nonidealities. Therefore, this proposed cell equivalently stores an accurate current regardless of I-V nonlinearity or $V_{TH}$ mismatches. It presents a >10x computing current variation reduction compared to the conventional cell design (Fig. 1 bottom left), hereby enabling an 8-level cell with a small LSB current of 100nA without calibration. Another issue in current-based CIM is the $I_{comp}$ sensitivity to $V_{RBL}$ caused by the short-channel effect (i.e., the limited output impedance), resulting in reduced accumulation accuracy. One may use a cascode stage to improve the output impedance. However, the conventional cascode transistor requires dedicated biasing, which is impractical in the memory array. This issue is addressed by the proposed dynamic-cascoded read structure, which consists of one LVT cascode transistor and one HVT main transistor with gates connected together. Once enabled, the $V_{DS}$ of the main transistor is set by the HVT and LVT threshold voltage difference (~200mV), which is sufficient for saturation region biasing. With the dynamic-cascoded stage, $I_{comp}$ sensitivity to $V_{RBL}$ is reduced by 4x without requiring any additional biasing branch. Incorporated with the voltage-current two-step write driver, the 3T1C MLC can is operated in sub-μA current, realizing high energy efficiency.

Fig. 2 shows the overall CIM architecture, including a 64x64 4b-cell array, the voltage-current two-step write driver, the 5b SAR ADCs and 4b DTCs for CIM operations, and control blocks for writing and computation (WCTRL, CCTRL). Inside the 4b-cell are two 3T1C MLCs, pseudo-differentially combined for 4b-signed weights (-700nA~700nA). To enhance the cells' retention, a MOM capacitor is placed on top of the transistors. The 4b activations are encoded to RWL pulse width by the DTC. All DTCs generate RWL pulses from the shared global timing signal TD0~TD15 sourced from CCTRL for good matching [5]. In the ADC, a current-source-assisted dynamic comparator is adopted to reduce offset sensitivity to input common-mode voltage [6]. During one computing cycle (180ns), the macro operations are divided into 3 phases. First, the BL capacitance and CDAC in SAR ADC are connected and precharged. Then, the RWL pulses activate CIM cells to discharge the RBL, performing current-based MACs. Last, the sampled analog-MAC value is quantized.



Die micrograph.

Fig. 3 describes the schematic and operation of the voltage-current two-step write driver that enables fast programming with small currents. Note that although the small operating current brings higher energy efficiency, driving the BL parasitics (50fF) with a small current (100nA) results in a long settling time (~400ns). To speed it up, the driver uses a two-step process: voltage-mode coarse writing followed by current-mode fine writing. In the first step, the voltage-write block drives the BL/BLb to $V_W$ or $V_{SS}$ (depending on the weight value) within 5ns. $V_W$ (0.52V) is the nominal $V_{GS}$ of read transistors when 100nA $I_{write}$ is applied. Following this, current-mode fine writing is used to accurately program the cell current to the target value. The simulation reveals that all current levels can be settled within 60ns, presenting a >6x speed up.

The prototype chip is fabricated in a 65nm CMOS. To verify the effectiveness of the proposed programming scheme, CIM transfer functions with voltage ($V_{write}$=0.52V) and current ($I_{write}$=100nA) programming are measured, respectively (Fig. 4 top). As can be seen, a 2.2x macro-level variation reduction is achieved with the proposed current-programming technique. Fig. 4 (bottom) shows the transfer functions of macro with the proposed voltage-current two-step driver for 15-level weights programming; the measured programming speed matches well with the simulation (65ns in total).

Fig. 5 shows the measurement results of retention time, refresh overhead, DNN performance, and macro energy. In the 2ms retention test, the "+7/-7" weights are written to the macro, and the drift values of cells are observed and normalized by ADC output. Within 0.4ms, 99.7% of cells realize a ≤1LSB drift, ensuring a >90% classification accuracy. With the refresh interval of 0.4ms, the refresh overhead of throughput and energy is only 1.1% and 0.07fJ/operation (normalized by 2.2k computing cycles in each refresh cycle). With the input sparsity ranging from 25%-to-75%, the measured 4b-MAC energy efficiency is 233-to-304TOPS/W.

Fig. 6 compares this work with state-of-the-art eDRAM CIM designs. For the first time, this work unifies the programming and computing in the current domain, thus mitigating various issues faced by conventional current-based CIMs, including the nonlinear I-V, $V_{TH}$ variations, and short-channel effect. In addition, with the sub-μA current MLC, the prototype achieves the highest macro-level 4b-MAC energy efficiency of 233-305TOPS/W among eDRAM CIMs.

**Acknowledgment:**

This work was supported by NSFC (U20A20204) and 111 Project (B18001). The authors thank Lin Bao and Yingming Lu for the technical discussions. The corresponding authors are Xiyuan Tang and Yuan Wang.

**References:**

[1] S. Xie *et al*., VLSI 2022. [2] S. Xie *et al*., ISSCC 2021.
[3] C. Yu *et al*., TCAS I, 2021. [4] Z. Chen *et al*., ISSCC 2021.
[5] A. Biswas *et al*., ISSCC 2018. [6] C.-C. Liu *et al*., JSSC 2010.
[7] N. R. Shanbhag *et al*., CICC 2022.

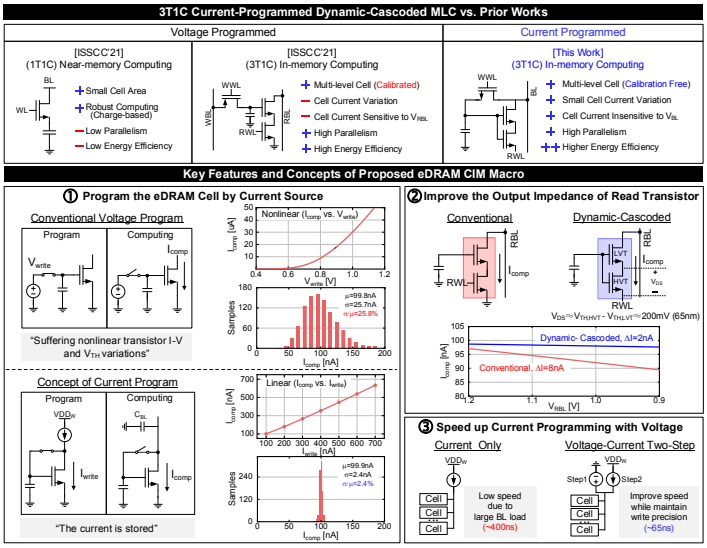



Fig. 1. Conventional eDRAM CIM cells and the proposed 3T1C current-programmed dynamic-cascoded MLC; key features and concepts of proposed eDRAM CIM macro.

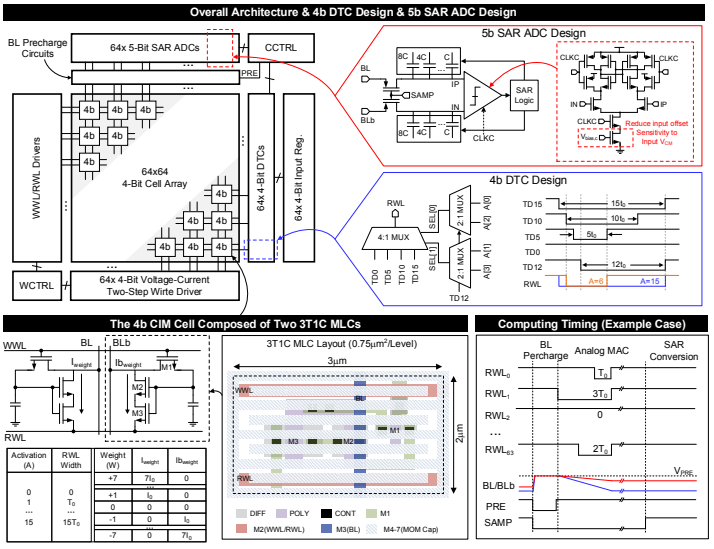



Fig. 2. Overall architecture of the proposed eDRAM CIM macro, DTC, and ADC schematic; 4b CIM cell schematic and 3T1C MLC layout, and computing timing diagram.

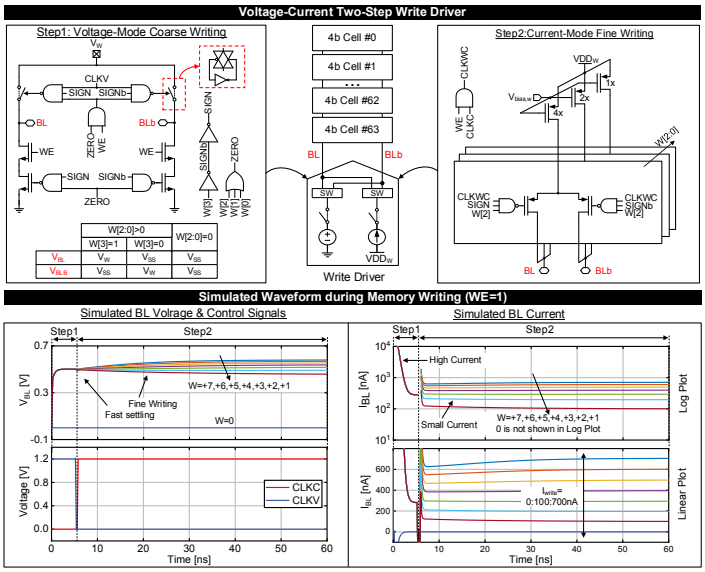



Fig. 3. Schematic and simulated waveforms of the voltage-current two-step write driver.

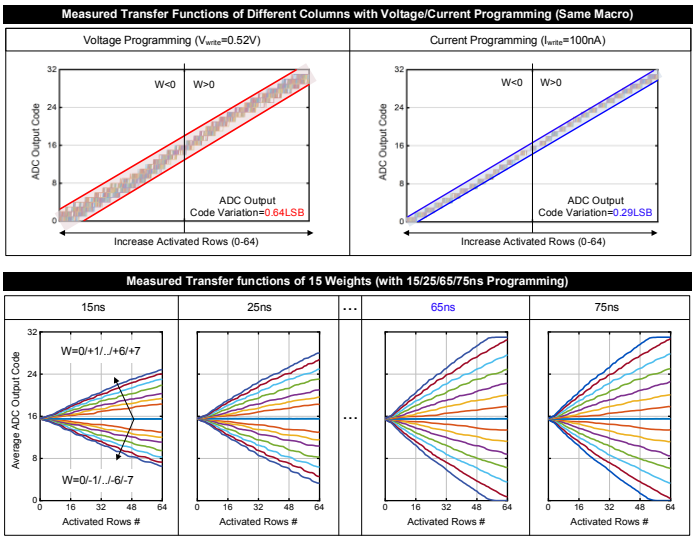



Fig. 4. Transfer functions of different CIM columns with voltage and current programming performed on the same macro (top). Transfer functions of 15 weight levels with 15/25/65/75ns programming time.

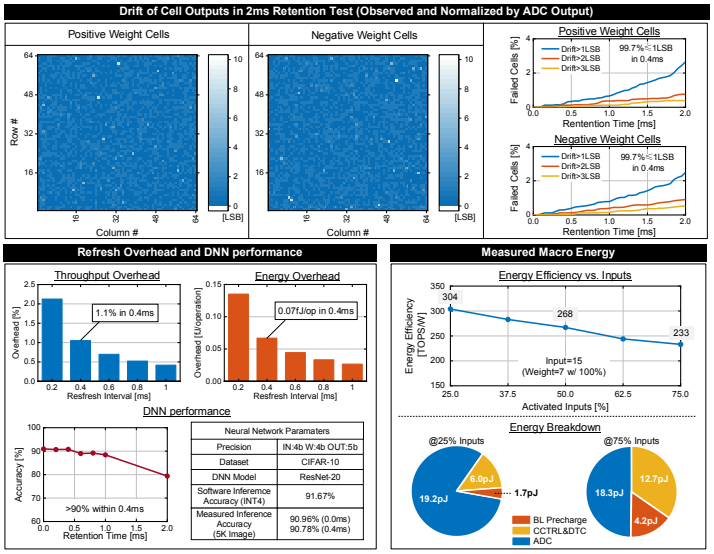



Fig. 5. The measurement results of retention time, refresh overhead, DNN performance, and macro energy.

**Comparison with Other eDRAM CIMs**

| | This Work | VLSI'22 S. Xie | ISSCC'21 S. Xie | ISSCC'21 Z. Chen | TCAS I '21 C. Yu |
|---|---|---|---|---|---|
| Technology | 65 nm | 65 nm | 65 nm | 65 nm | 65 nm |
| Programming | Current | Voltage | | | |
| Computing | Current | Charge | | Current | |
| eDRAM Cell | 3T1C Cell | 2T1C Gain Cell | 1T1C Cell | 3T1C Gain Cell | 2T1C Gain Cell |
| Macro Size | 16Kb | 32Kb | 16Kb | 8Kb | 16Kb |
| Fully-Parallel MAC | ✓ | ✗ | ✗ | ✓ | ✓ |
| Multi-level Cell (MLC) | ✓ | ✗ | ✗ | ✓ | ✗ |
| Calibration Free | ✓ | ✓ | ✓ | ✗ | ✗ |
| OTA Free | ✓ | ✗ | ✗ | ✓ | ✓ |
| Parallelism | 64 | 27 | 1 | 64 | 64 |
| Levels/CIM Cell | 15 | 2 | 2 | 16 | 2 |
| CIM Cell Area | 2×6μm² | N/A | 22.08μm² | N/A | 1.08μm² |
| [1]CIM Column Input Precision | 4b | 2b | 8b | 4b | 1b |
| [1]CIM Column Weight Precision | 4b | 1b | 8b | 4b | 1.5b |
| Dataset | CIFAR10 | CIFAR10 | CIFAR10 | CIFAR10 | CIFAR10 |
| Classification Accuracy | [3]90.78-90.96% | 92.02% | 80.1% | 90.6% | 82.8% |
| Computing Density (TOPS/mm²) | [3]2.963 (I:4b W:4b O:5b) | 0.4113 (I:2b W:1b O:2b) | 0.00826 (I:8b W:8b O:8b) | N/A | 309 (simulated) (I:1b W:1.5b O:1b) |
| [4]Normalized Computing Density (TOPS/mm²) | 237.04 | 1.65 | 4.23 | N/A | 463.5 (simulated) |
| Energy Efficiency (TOPS/W) | [5]233-304 (I:4b W:4b O:5b) | 236 (I:2b W:1b O:2b) | 4.76 (I:8b W:8b O:8b) | 102.2 (I:4b W:4b O:5b) | 552.5 (simulated) (I:1b W:1.5b O:1b) |
| [6]Normalized Energy Efficiency (TOPS/W) | 24320 | 944 | 2437.12 | 8176 | 828.75 (simulated) |

[1] An CIM column includes the circuitry and computations that precede the input to a single ADC [7].
[2] Within 0.4ms retention time.
[3] Assuming 1OP=1 addition or 1 multiplication.
[4] Normalized computing density= input precision × weight precision × output precision × computing density.
[5] Measured with 25%-to-75% input ratio and assuming 1OP=1 addition or 1 multiplication.
[6] Normalized energy efficiency= input precision × weight precision × output precision × energy efficiency.

Fig. 6. Comparison with other state-of-the-art eDRAM CIMs.