Low-Power and Scalable BEOL-Compatible IGZO TFT eDRAM-Based Charge-Domain Computing

Wenjun Tang, *Graduate Student Member, IEEE*, Jialong Liu, *Graduate Student Member, IEEE*, Chen Sun, *Graduate Student Member, IEEE*, Zijie Zheng, *Graduate Student Member, IEEE*, Yongpan Liu, *Senior Member, IEEE*, Huazhong Yang, *Fellow, IEEE*, Chen Jiang, *Member, IEEE*, Kai Ni, *Member, IEEE*, Xiao Gong, *Member, IEEE*, and Xueqing Li, *Senior Member, IEEE*

***Abstract*— The rapid development of edge artificial intelligence (AI) raises high requirements for data-intensive neural network (NN) computing and storage of edge devices, under a limited chip footprint and energy supply source. As a promising approach for energy-efficient processing, computing-in-memory (CiM) has been widely explored in recent efforts to mitigate the data transmission bottleneck. However, CiM with small on-chip memory capacity results in expensive data reloads, limiting its deployment in large-scale NN applications. Moreover, the increased leakage under advanced CMOS scaling lowers the energy efficiency. In this work, device-circuit synergy based on the indium-gallium-zinc-oxide (IGZO) thin-film transistor (TFT) is adopted to address these challenges. First, 4-transistor-1-capacitor (4T1C) IGZO eDRAM CiM is proposed with higher density than SRAM-based CiM and enhanced data retention by both lower device leakage and a differential cell structure. Second, exploiting the back-end-of-line (BEOL) compatibility and vertical integration of emerging channel-all-around (CAA) IGZO devices, 3D eDRAM CiM is proposed, which paves the way for IGZO-based CiM with ultra-high density. Circuit techniques including time-interleaved computing and differential refresh are proposed to guarantee accuracy under large-capacity 3D CiM. As a proof of concept, a 128 × 32 CiM array is fabricated under a foundry low-temperature poly-crystalline and oxide (LTPO) technology, demonstrating high computing linearity and long data retention. Benchmarks on scaled 45nm IGZO technology show energy efficiency of 686 TOPS/W for array only, and 138 TOPS/W while considering peripheral overheads.**

***Index Terms*— Charge-domain computing, IGZO TFT, eDRAM, computing-in-memory, DNN accelerator.**

Manuscript received 18 June 2023; revised 27 August 2023; accepted 11 September 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706100; in part by NSFC under Grant U21B2030, Grant 92264204, and Grant 82151305; and in part by the Advanced Research and Technology Innovation Centre (ARTIC) Program, National University of Singapore, under Grant R-261-518-006-720. This article was recommended by Associate Editor R. Joshi. *(Corresponding author: Xueqing Li.)*

Wenjun Tang, Jialong Liu, Yongpan Liu, Huazhong Yang, Chen Jiang, and Xueqing Li are with BNRist, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: xueqingli@tsinghua.edu.cn).

Chen Sun, Zijie Zheng, and Xiao Gong are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117582.

Kai Ni is with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2023.3317170.

Digital Object Identifier 10.1109/TCSI.2023.3317170

I. Introduction

EDGE artificial intelligence (AI) has been developing rapidly with the exponential growth of edge devices. Deep neural network (DNN) is a key enabler of edge AI, which has shown great application prospects in a rainbow of intelligent scenarios such as image classification [1], speech recognition [2], and reinforcement learning [3]. However, the increasing data volume of DNN models, combined with the data-intensive computing paradigm, requires efficient hardware acceleration. In conventional von Neumann architecture, the separated memory unit and computing unit result in high costs of data transmission, thus limiting performance improvement when deploying large-scale DNN algorithms [4]. To mitigate this so-called "memory wall" issue, computing-in-memory (CiM) has emerged as a promising solution. By merging memory and computing, frequent data movement is reduced. And the array topology naturally enables high-parallelism multiply-accumulation (MAC), which is the basic operation in most DNN algorithms. Recent works based on SRAM [5], embedded DRAM (eDRAM) [6], and emerging nonvolatile memories (NVMs) [7] have shown its effectiveness.

Under the limited footprint, battery size, fabrication cost, and possible unstable energy supply sources and environments, CiM accelerators for edge AI pursue the improvement of density, energy efficiency, and immunity against large device variation. Additionally, CiM accelerators should have scalability for different edge AI application scenarios, which have varying network sizes and event frequencies. Therefore, this work considers the challenges and opportunities of CiM from several perspectives, as summarized in Fig. 1.

Firstly, memory density is a crucial aspect of edge AI devices, especially for deploying large DNN models. A high system-level memory density enables scalability for increased model size. For CiM macro design, a large memory capacity reduces the need for extra on-chip weight buffers or external DRAM, thereby decreasing the footprint, and reducing high-cost weight reloads from external memories in terms of dataflow [8]. eDRAM has a higher density than SRAM and more mature fabrication than emerging NVMs, making it a promising solution for CiM with high memory capacity.

Secondly, the frequent refresh of conventional Si-based eDRAM lowers energy efficiency, and also, limits the normal access of memory. Especially in advanced technologies, the high leakage current leads to a high refresh frequency, reducing the benefits of eDRAM-based CiM designs. Recently,

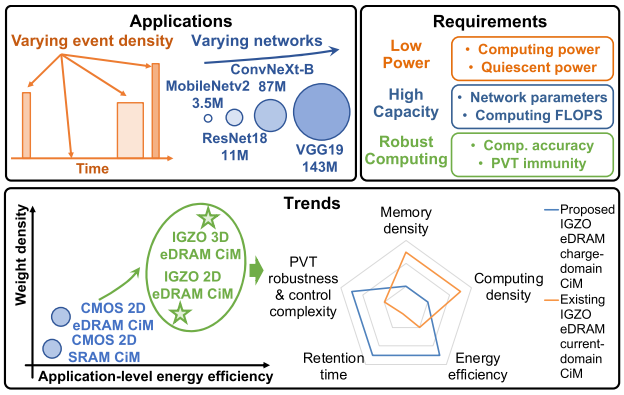


Fig. 1. Motivation and highlights of the proposed IGZO eDRAM-based charge-domain CiM.

indium-gallium-zinc-oxide (IGZO) thin-film transistor (TFT) has garnered much interest for its ultra-low off-state current, which makes it competitive for low-refresh eDRAM designs [9]. This feature not only reduces the charging energy for the internal storage node, but also lowers the energy consumption of frequently charging bitline capacitance.

Thirdly, the design approach of eDRAM-based CiM affects energy efficiency in terms of both quiescent and computing power. Recent IGZO TFT-based eDRAM CiM designs exploit current-domain computing that utilizes the on-state current of the transistor to perform the computation [10], [11], [12]. On the one hand, the quiescent power is determined by the refresh activities of eDRAM. Especially for eDRAM CiM relying on the absolute resistance controlled by the storage node, leakage-induced node charge loss can result in severe errors. Therefore, high-accuracy CiM computing with this scheme requires a significantly higher refresh frequency than normal memory mode, resulting in a high refresh energy overhead for long computing tasks. On the other hand, to reduce computing power, exploiting charge-domain computing can benefit from its DC-power-free operation [13]. This paper proposes a ratio-based charge-domain CiM design to reduce the high sensitivity to resistance variation. The proposed scheme achieves high computing accuracy while greatly reducing the need for frequent refreshes.

Fourthly, an even higher memory density of the eDRAM CiM remains to be explored. Recent advancements in 3D memory structures have paved the way for high-capacity CiM, which breaks the limit from device feature size in 2D layout. For IGZO TFT, the back-end-of-line (BEOL) compatibility enables multi-layer stacking over Si CMOS peripherals. More excitingly, the novel vertical device based on channel-all-around (CAA) structure demonstrates a more compact eDRAM cell design and the capability of 3D stacking of eDRAM layers [14], [15]. However, new challenges arise for the large-capacity 3D eDRAM CiM array, including data refresh, coupling interference, and sub-array interconnection. This paper explores the 3D eDRAM CiM design based on the CAA-IGZO device, and proposes multiple circuit-level techniques to address the new issues.

Different from prior works, this work explores both high weight density and high application-level energy efficiency, with robust high-linearity computing. In detail, this work makes the following contributions with highlighted features.

- High-robustness energy-efficient 4T1C eDRAM CiM design along with the cell-level test result. The 4T1C cell exploits a differential structure that is robust to both leakage-induced charge loss and PVT-induced degradation. The retention time, defined as the standby time without a refresh under the same computing accuracy, is >50x longer than the existing IGZO current-domain eDRAM CiM. The charge-domain computing scheme eliminates static power. The high-scalable operation without complex or precise timing controls makes it suitable for large-capacity 3D integration, and also, edge devices with limited resources.
- Taped-out chip and test result with 4 kb array-level integration of proposed eDRAM CiM based on low-temperature polycrystalline and oxide (LTPO) TFT, demonstrating the scalability of the proposed design. The merge of low-off-current IGZO and high-on-current low-temperature polycrystalline silicon (LTPS) enlarges the speed-retention-area design space of the 4T1C structure. The test result proves the long retention time of >3 hours, and the simulation based on an experimentally calibrated model shows ∼10 hours of retention.
- High-density BEOL 3D integration based on CAA-IGZO. The vertical device structure and further CiM layer stacking enable compact 4T1C cell designs lower than $8F^2$. Careful design of layer connection of sub-arrays shows the scalability of memory layer stacking with <10% of area cost. Additionally, the Si readout peripheral under CiM array breaks the ADC parallelism bottleneck. Techniques including time-interleaved control and differential refresh scheme guarantee high computing accuracy under large-capacity CiM array.
- System-level evaluations of typical edge sensing scenarios. The DC-power-free computing and enhanced retention time enable lower computing and quiescent power consumption, respectively.

Our previous papers [16], [17] have demonstrated the basic operation and designs of IGZO-based 4T1C eDRAM CiM and its BEOL 3D integration. This paper further illustrates the advantages and limits of the 4T1C structure, and gives the design guideline with BEOL circuit considerations emphasized. With the in-depth analysis, a prototype LTPO chip is fabricated and measured to verify the retention and accuracy. On the basis of the previous ideas, this paper contributes circuit techniques including layer-selection-based interconnection and differential refresh to explore the potential of proposed design under macro-level integration. In the rest of this paper, Section II introduces the recent progress in IGZO TFT and CiM. Section III presents the proposed 4T1C eDRAM CiM structure and cell-level measurement. Section IV reports the array-level fabrication based on LTPO technology. Section V illustrates the monolithic 3D integration of proposed 4T1C eDRAM CiM based on CAA-IGZO. Section VI evaluates the design at the circuit level and the application level. Section VII concludes this work.

II. Background

A. IGZO TFT Fundamentals

IGZO TFT has been widely investigated for display applications over the last few decades, benefitting from its ultra-low leakage and medium-to-high mobility. Recent works of IGZO TFT devices report off-state current as low as 10^{-17} A/μm – 10^{-20} A/μm [14], [18]. For a type of optimized IGZO device known as c-axis-aligned crystalline

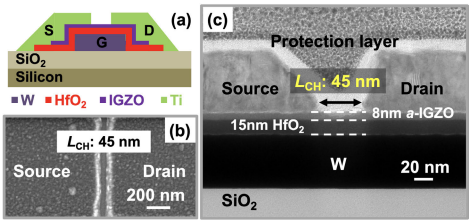


Fig. 2. IGZO TFT device [16]. (a) Structure illustration of IGZO device. (b) Top-view SEM image of the TFT with 45 nm L_{CH}. (c) Cross-section TEM image of the 45 nm TFT device.

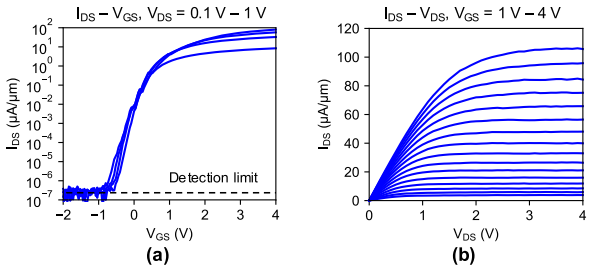


Fig. 3. Device characteristics of IGZO device with W/L = 1μm/0.045μm [16].

(CAAC) IGZO, low off-state current of 10^{-22} A/μm is measured [19]. These device fabrications and measurement results demonstrate the potential of IGZO for low-power display applications, and also, long-retention eDRAM or 4-bit multi-level-cell (MLC) memory [20]. Moreover, the low-temperature BEOL compatibility of IGZO TFT attracts much interest recently, which enables monolithic 3D integration with emerging near-computing high-density memory applications [14]. Besides, recent demonstrations of TFT-based processors [21], [22] and CiM designs [11], [16], [23] highlight opportunities for in-situ intelligence processing.

The scaling of IGZO TFT is advancing rapidly. On the one hand, planar IGZO devices have achieved extremely scaled channel length (L_{CH}) down to 12.3 nm, leading to high cut-off frequency and extrinsic transconductance [24]. On the other hand, the vertical structure of CAA-IGZO enables scaled 50 nm channel length and critical dimension (CD) down to 50 nm, which performs both device density improvement and L_{CH} scaling by exploiting the BEOL capability [15]. Fig. 2 and Fig. 3 show the IGZO device structure and characteristics [16].

However, issues from device instability and advanced scaling pose challenges to IGZO-based circuit designs. Threshold voltage (V_{TH}) shift of the TFT device caused by long-term bias stress or high junction temperature, i.e., bias-temperature instability (BTI), is among the main challenges different from Si CMOS [25], [26]. This effect results in both degraded memory storage capability and difficulty of analog circuit designs that are sensitive to V_{TH}. Furthermore, with advanced scaling of the IGZO device, the node capacitance will shrink, resulting in degraded retention for capacitor-less eDRAM application [14]. This paper proposes circuit-level techniques to address these issues and unlock the full potential of IGZO devices with enhanced reliability and retention time of eDRAM-based CiM.

The benefits from heterogeneous integration bring various opportunities for IGZO TFT. For high-performance display applications, LTPS TFT is still preferred because of its high on-state current. LTPO technology has been developed as one of the commercial TFT processes for displays, which

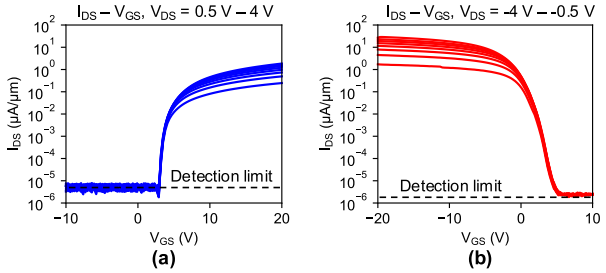


Fig. 4. Device characteristics of (a) IGZO device with W/L = 4μm/10μm, and (b) LTPS device with W/L = 9μm/4μm, from the foundry LTPO process [28].

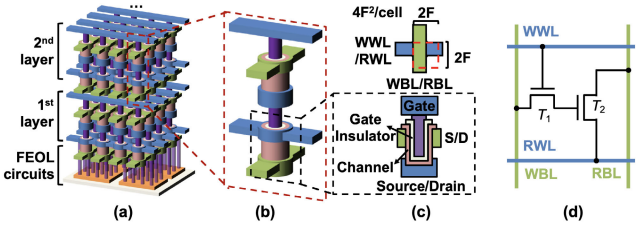


Fig. 5. CAA-IGZO device [14], [15]. (a) BEOL-compatible monolithic stacking towards high-density memory chip. (b) Cell structure. (c) Top-down view of a $4F^2$ cell and cross-section view of a single device. (d) 2T0C schematic. WWL: write wordline; RWL: read wordline; WBL: write bitline; RBL: read bitline.

integrates low-off-current IGZO TFT and high-on-current LTPS TFT. Successful display applications of LTPO have demonstrated its low-power characteristic with a refresh frequency down to 1 Hz [27]. Fig. 4 shows the device characteristics of LTPO from Tianma Microelectronics Co., Ltd. [28], where the real off-state current of IGZO is much lower than the detection limit. At the circuit level, LTPO technology is suitable for display and memory applications with the achievement of both long retention time and high-speed read capability.

For high-density 3D integration, recent reports introduce a novel vertical CAA structure for IGZO TFT [14], [15], as shown in Fig. 5. The vertical structure allows direct connection of gate and source/drain of two transistors, enabling a compact $4F^2$ footprint for a 2T0C eDRAM cell. BEOL memory stacking on the FEOL CMOS peripherals enables high-density eDRAM memory with an equivalent cell footprint lower than $4F^2$.

B. Existing CiM Designs

Recently, CiM has been actively researched with numerous implementations based on SRAM [5], eDRAM [6], [29], and emerging NVMs [7]. SRAM-based and eDRAM-based CiM benefits from device endurance, stability, and mature fabrication, while NVM-based CiM offers high memory density and low standby power. By circuit-level operating principles, CiM can be categorized as charge domain, current domain, time domain, and digital domain. The common intrinsic superiority of the CiM schemes over conventional computing is the merged memory and computing unit, enabling reduced data transfer cost and massive parallelism through array topology. Dealing with the increasing demands of large-scale NN models, CiM designs require efforts on the improvement of memory density. This is achieved by the exploitation of local computing cell [30], MLC [31], [32], ultra-dense 1T read-only memory (ROM) [33], etc. Furthermore, evaluations and demonstrations of 3D CiM

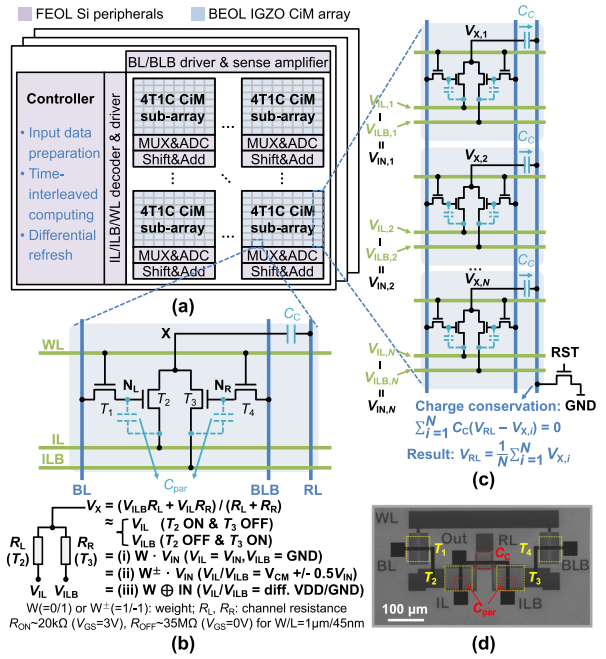


Fig. 6. Proposed 4T1C eDRAM CiM. (a) Architecture. (b) Cell structure with the principle of differential weight. (c) Array structure with the operation of charge-domain MAC. (d) SEM image of the fabricated 4T1C cell. BL/BLB: bitline pair; WL: wordline; IL/ILB: input line pair; RL: result line.

designs based on RRAM [34], [35] or 3D-NAND flash [36], [37] provide a new dimension for high-density CiM.

IGZO-based CiM usually exploits eDRAM structure for high memory density and low power [38]. Existing IGZO-based CiM works [10], [11], [12], [39] mostly utilize the non-destructive gain cell for current-domain computing, which is also capable of multi-level weight [11] [12], [39]. However, the low immunity to charge loss on internal nodes and transistor mismatch makes it challenging to achieve high computing accuracy. Also, the BEOL capability of IGZO TFT requires further exploration in eDRAM CiM designs. Among different analog CiM schemes, charge-domain computing exhibits high computing accuracy under high parallelism [13], benefitting from good capacitor matching. This paper utilizes the principle of charge-domain computing and proposes compact 4T1C eDRAM CiM with the unipolar IGZO device. Furthermore, the exploration of 3D eDRAM CiM extends the capability of IGZO-based CiM with several techniques for enhanced computing accuracy.

III. Proposed 4T1C eDRAM CiM and Cell Measurement

This section explains the proposed 4T1C eDRAM CiM cell and array design with detailed circuit considerations. Moreover, the cell measurement illustrates the operation principle and shows the long retention characteristic of IGZO TFT.

A. Structure and Operating Principles

The overall architecture of proposed IGZO eDRAM CiM is shown in Fig. 6(a), with the integration of FEOL peripherals and BEOL CiM array. The controllers include functions for improved computing accuracy, as illustrated in Sections V-C and V-D. The drivers include high-voltage WL/BL/BLB driving and low-voltage IL/ILB driving. The former can exploit Si PMOS thick-gate transistors and N-type IGZO TFTs [40], while the latter can be directly implemented by Si thin-gate transistors. The sense amplifiers (SAs) exploit a latch structure to perform efficient eDRAM data read and refresh [41], where the direct write to memory cell enables fast eDRAM refresh. The partition and interfacing details of the proposed BEOL-device-based CiM are introduced in Section V-B.

The circuit and operation are illustrated in Fig. 6(b)(c). Each CiM cell consists of four transistors T_1 – T_4 and one coupling capacitor C_C. Unlike existing eDRAM CiM designs, the weight is represented by the differential gate-source voltage on the gate-source parasitic capacitor C_{par} of T_2 and T_3, i.e., the on/off ratio of the channel resistance of T_2 (R_L) and T_3 (R_R). During write operation, a fixed low voltage $V_{IL,L}$ is kept on IL/ILB, the differential voltage $V_{BL,L}/V_{BL,H}$ or $V_{BL,H}/V_{BL,L}$ representing the weight is prepared on BL/BLB, and a pulse is applied on WL. During CiM operation, the input is applied to IL and ILB in multiple patterns including multi-bit analog input (i) and (ii), or 1-bit digital input (iii), as shown in Fig. 6(a). As will be shown in Section V, the differential input pattern (ii) and (iii) can have improved computing accuracy using several techniques. In-cell multiplication or XNOR logic is implemented by the voltage division of differential R_L and R_R pair. For array operation, C_C performs charge-domain accumulation across the rows in a column. First, the IL/ILB and RL are set to certain fixed voltages. Then float RL and apply inputs to IL and ILB. The in-cell calculation between input and cell weight generates output $V_{X,i}$ on the left plate of C_C. The average of $V_{X,i}$ is established on RL by C_C charge redistribution, as shown in Fig. 6(c).

The proposed 4T1C eDRAM CiM design has high immunity to the variation of transistor channel resistance introduced by node charge loss, mismatch, PVT variation, fluctuation of the internal storage node, etc. This immunity is owing to the large on-off ratio of R_L/R_R (or R_R/R_L), which attenuates the influence of R_L or R_R variation on $V_{X,k}$. Considering the random variation, the main contributor to the accuracy loss becomes the capacitor mismatch, which is typically low. Considering the leakage-induced accuracy degradation, the proposed design requests a low refresh frequency to maintain high-accuracy CiM operation, leading to low quiescent power. The remaining challenge is the variation of retention time caused by the mismatch of C_{par} and transistor off-current. A possible solution is to apply foreground calibration to determine the refresh frequency for a CiM sub-array.

Also, charge-domain computing allows for energy-efficient high-accuracy accumulation. The DC-current-free operation leads to low computing power. The static result generation on RL does not require accurate pulse control, and also, is robust to clock skew. These benefits of the proposed 4T1C eDRAM CiM scheme enable high scalability, which makes it feasible for large-capacity 3D integration, as explained in Section V.

B. Design Considerations

The design of the proposed 4T1C eDRAM CiM needs careful consideration in terms of layout parasitic and operating voltage.

For layout, high-accuracy computing should be guaranteed in both column and in-cell operations. For column accumulation, the floating RL should be shielded by node X to immunize the disturbance from other lines that cause data-dependent computation errors. For in-cell multiplication, the undesired parasitic coupling should be avoided. The internal storage node N_L or N_R has mainly three parasitic capacitors: C_{GD}, C_{GS} of T_2 or T_3, and C_{GD} of T_1 or T_4. Parasitic coupling during CiM operation can cause on-off resistance ratio degradation by unexpected increased or decreased V_{GS} of T_2 or T_3. A better design is to make C_{GS} of T_2 or T_3 larger than the other two. With C_{GS}, i.e., C_{par} in Fig. 6, as the storage capacitance, the 4T1C cell benefits from the bootstrapped gate voltage of T_2 or T_3 in CiM operation. Therefore, the V_{GS} can keep almost unchanged when applying voltages to IL and ILB. And also, the constant on-resistance enables fast settling of RL. A larger C_{par} can be implemented by increasing gate-source overlap of T_2 and T_3 or fully utilizing the layout space of the cell, which has negligible area overhead.

Designs of operating voltage must guarantee both accurate computing and efficient operation. For CiM operation, only IL/ILB is active. According to the principle of charge-domain computing, the voltage range of output RL remains the same as that of input IL/ILB. Therefore, the operating V_{IL} can be CMOS-compatible, e.g., 0 V – 0.8 V. Exploiting the separate power domains for write and CiM operations, energy-efficient CiM computing can be achieved. Also, voltage translators for IL, ILB and RL between IGZO array and Si CMOS input and readout peripherals can be eliminated. Actually, the IL/ILB voltage range should be constrained to avoid DC current along IL-T_2-T_3-ILB path and even incorrect in-cell computing results. Assume that the storage node parasitic other than C_{par} is negligible, the voltage constraint can be expressed as:

$$V_{IL,H} - V_{IL,L} < V_{TH} + V_{IL,L} - V_{BL,L} < V_{BL,H} - V_{BL,L}, \tag{1}$$

where $V_{IL,L}$ and $V_{IL,H}$ denote the lowest and highest voltage for IL, respectively. Further, to ensure high-accuracy computing, the on-off ratio of the differential pair should be focused. Assume that the idle voltage for BL/BLB is $V_{BL,L}$, and $\Delta V_N(t)$ represent the leakage-induced voltage drop at node N which initially stores $V_{BL,H}$. The worst-case ratio occurs when weight '0' is stored and $V_{IL,H}$ is applied as input, which is expressed as:

$$Ratio\,(t) = \frac{R_{DS}\left(V_{GS} = V_{BL,H} - V_{IL,L} - \Delta V_N\,(t)\right)}{R_{DS}\left(V_{GS} = V_{BL,L} + V_{IL,H} - 2V_{IL,L}\right)}. \tag{2}$$

By assigning proper operation voltages using this formula, the ratio can be guaranteed to be sufficiently large within the expected standby time.

For write and refresh operation, sufficiently turned-on IGZO TFT usually needs a higher voltage. On the one hand, these operations are much more infrequent compared with CiM operations, thanks to the low-leakage memory storage. For an application approaching the maximum computing throughput without weight reload, a refresh can cover $>10^8$ CiM operations, thus it has negligible overhead on overall performance. On the other hand, a lowered write V_{WL} and V_{BL} can exploit subthreshold operation of TFT to reduce power for write and refresh, with the cost of longer latency and worse reliability.

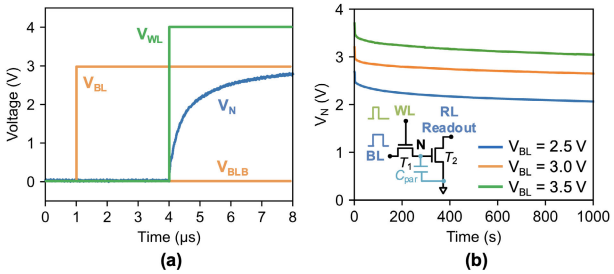


Fig. 7. Measurement of (a) write operation and (b) retention characteristic of a single IGZO eDRAM CiM cell [16].

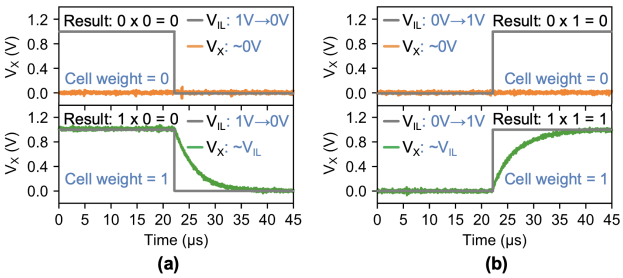


Fig. 8. Measurement of in-cell computing operation of a single IGZO eDRAM CiM cell with (a) input '0' and (b) input '1' [16].

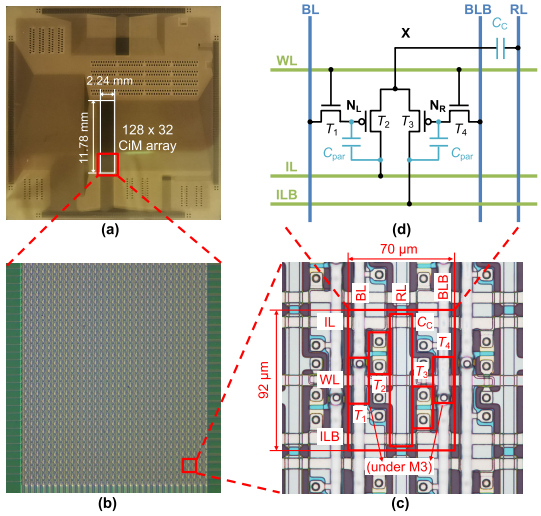


Fig. 9. Array-level integration of the proposed eDRAM CiM cell. (a) Chip photograph. (b) Array micrograph. (c) Cell layout. (d) Schematic for proposed LTPO-based 4T1C cell implementation.

C. Cell Measurement

To demonstrate the basic operation of the proposed eDRAM CiM, discrete cells are fabricated in the laboratory using IGZO TFT with 5 μm L_{CH}, as shown in Fig. 6(d). In cell measurement, the voltage range of BL and BLB is 3 V, the voltage range of WL is 4 V, and the voltage range of IL and ILB is 1 V.

First, half of the 4T1C cell is tested for write operation and data retention characteristics, with the waveform shown in Fig. 7. The result shows almost-nonvolatile memory characteristic with only 0.4 V charge loss on internal node N after 1000 s where V_{BL} = 3.5 V. Second, the in-cell multiplication of the 4T1C cell is performed in Fig. 8. In the demonstration, input pattern (i) is exploited where ILB is fixed to 0 V. Input on IL is applied continuously. When the cell weight is '0', the cell output V_X is always 0 V. When the cell weight is '1', the V_X almost equals the input V_{IL}. In the measurement of a single cell, the accumulation output V_{RL}

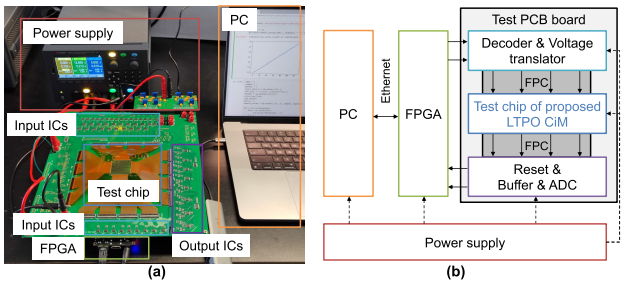


Fig. 10. Test platform of the proposed LTPO eDRAM CiM array chip. (a) Photograph. (b) Function diagram.

TABLE I
THE OPERATION VOLTAGE FOR DIFFERENT OPERATIONS

	Write	Read	Standby
WL	12 V	-12 V	-12 V
BL/BLB	5 V, -7 V	5 V	5 V
IL/ILB	0 V	0 V – 4 V	0 V
RL	0 V	Floating	0 V

should be the same as the V_X. The latency is limited by the 3.9 pF oscilloscope input capacitance.

IV. ARRAY INTEGRATION OF PROPOSED 4T1C EDRAM CIM

To further validate the array operation and show the effectiveness of the proposed TFT-based eDRAM CiM, a 4 kb CiM array with 128 rows and 32 columns is implemented in LTPO technology, demonstrating the large-scale capability.

A. Structure and Layout Design

The commercial TFT process utilizes LTPO technology for low-power displays, which combines the benefits of low-leakage IGZO and high-performance LTPS. The proposed eDRAM CiM exploits the LTPO technology, where T_1 and T_4 for memory access are N-type IGZO with $W/L = 4$ μm/10 μm, and T_2 and T_3 for computing are P-type LTPS with $W/L = 9$ μm/4 μm, as shown in Fig. 9. The LTPO-based eDRAM CiM cell can achieve long retention and high computing speed simultaneously under a compact cell area.

The layout design is shown in Fig. 9(c). The cell size is 92 μm × 70 μm, under 4 μm LTPO process with 4 metal layers utilized. The C_C has a metal-insulator-metal (MIM) structure. Considering the inert substate of the TFT process, the RL, i.e., the top plate of the C_C, is implemented by the bottom metal. Node X in the second metal layer provides shielding to RL. The remaining area other than the C_C is fully utilized to implement C_{par}, in order to increase cell retention as much as possible. The measurement shows a 0.41 pF C_C and a 0.77 pF C_{par}.

B. Array Operation and Measurement

The chip is fabricated by Tianma Microelectronics Co., Ltd. [28]. Then it is measured by the custom test platform as shown in Fig. 10. The fabricated LTPO eDRAM CiM chip is connected to the test PCB board by flexible printed circuit (FPC) connections. Read and write functions of the fabricated CiM array are controlled by peripheral chips on the custom PCB. The write control and input vector are sent from FPGA and converted to sufficient operation voltages for TFT. The

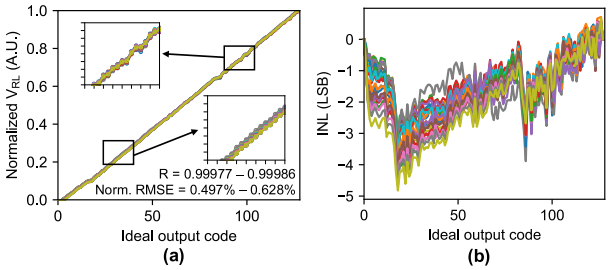


Fig. 11. Measurement results of the proposed LTPO eDRAM CiM array with varying inputs rows and weight storing all '1'. (a) V_{RL} for each column and (b) corresponding INL. Different colors represent different output columns.

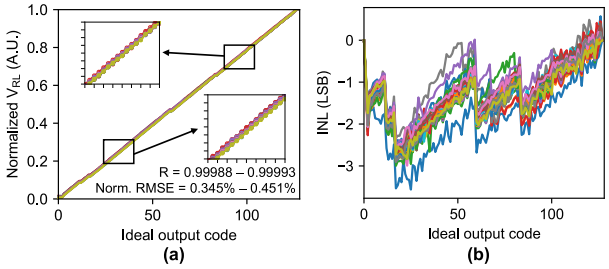


Fig. 12. Measurement results of the proposed LTPO eDRAM CiM array with varying rows applying '1' and weight storing all '1'. (a) V_{RL} for each column and (b) corresponding INL. Different colors represent different output columns.

reset for output RL is implemented by switches on the PCB. After a CiM computation, the analog output voltage is buffered and sensed on the PCB, and then sent to FPGA. PC generates the test vector and collects the computing results.

The operation voltage considering P-type computing transistors T_2 and T_3 is shown in Table I. High voltage levels are adopted to ensure switching with high on-off ratio. Firstly, the computing linearity is evaluated utilizing integral nonlinearity (INL) as the metric. INL is defined as the error of real output values to the ideal output values normalized to the ideal code width. Also, correlation coefficient R and root mean square error (RMSE) normalized to the output range are given. In this paper, two testbenches for linearity measurement are evaluated: (1) Store '1' in the whole CiM array, and sweep the number of rows applying 4V to ILs with the others remaining 0 V. (2) Apply 4 V to ILs of all rows of the array, and sweep the number of rows storing '1' with the others '0'. In the test, 29 out of 32 columns are used since the other 3 columns work abnormally. Fig. 11 and Fig. 12 show the measurement results of the two testbenches with V_{RL} normalized to the output range. Linearity of higher than 0.9995 correlation coefficient, less than 0.7% normalized RMSE, and lower than 5 LSB INL for all 29 columns is observed. The primary cause for the INL jumps is the non-writable rows, which may arise from electrostatic-discharge-induced failure, weak FPC bonding, voltage translator failures, etc. The variation of INL could originate from the C_C mismatch, the noise or disturbance from PCB, the variation of each reset operation by the switches on PCB, etc.

Secondly, the retention time of the fabricated LTPO chip is measured. Since the off-state current is much below the detection limit of the semiconductor parameter analyzer, a 2T0C cell, which is half of the proposed CiM cell, is utilized as shown in Fig. 13. Once the node N is charged, keep the WL with −12V and detect the channel current of T_2. As shown in Fig. 13(a), the measurement result of 5 different cells shows 0.10 V/h voltage drop of node N for the worst case. To further

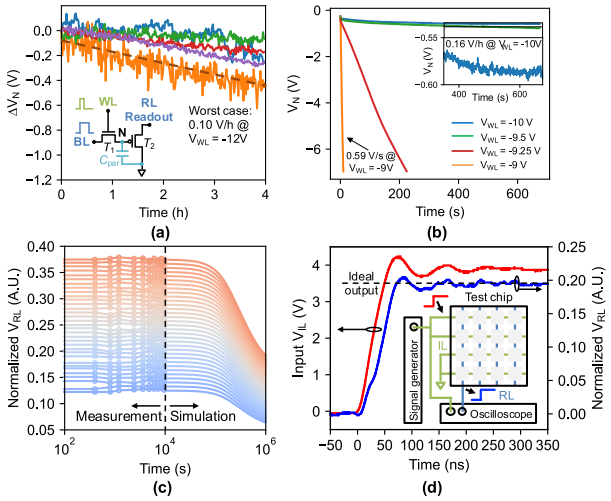


Fig. 13. The retention and latency measurement with weight storing all '1', initial write $V_{BL} = 0$V, and idle $V_{BL} = -7$V: (a) Voltage drop ΔV_N of 5 different cells with idle $V_{WL} = -12$ V. (b) Voltage drop of storage node 'X' with different idle VWL. (c) Retention measurement and prediction of CiM operation. (d) Latency measurement of CiM operation.

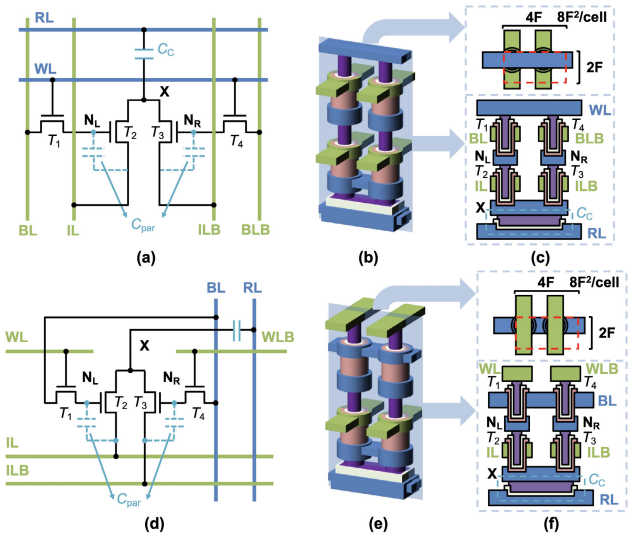


Fig. 14. The proposed 3D eDRAM CiM. Layout design (i) [17]: (a) Schematic. (b) 3D view. (c) Top-down view and cross-section view. Layout design (ii): (d) Schematic. (e) 3D view. (f) Top-down view and cross-section view.

investigate the effect of different idle V_{WL} on the leakage, sweep idle V_{WL} and obtain the corresponding voltage drop curves. As the idle V_{WL} decreases, the voltage drop slows down with the subthreshold characteristic. Accounting for the threshold voltage (V_{TH}) variation of the TFT, the array-level test of CiM operation adopts -12 V V_{WL} in idle state. Fig. 13(c) shows the retention test of the proposed 4T1C CiM array, with weight storing a matrix of staggered '0's and '1's, and ideal output code from 16 to 48. Measurement ensures >3 h no-refresh standby time for CiM operation. Further simulation with an experimentally calibrated model indicates ~10 h retention time with <0.5 LSB error of V_X, under idle V_{WL} of -12V. On the contrary, the retention of a current-domain CiM counterpart is ~10 min. However, V_{TH} shift after such a long-term bias at the storage node could be an issue to be further investigated.

Thirdly, the latency for the CiM operation is measured. The inputs on ILs are directly applied by a signal generator, and the output voltages on RLs are measured by an oscilloscope. The measured CiM latency is within 150 ns, primarily limited by the non-ideal rising edge of the signal generator.

V. 3D Structure Design of Proposed 4T1C eDRAM CiM

The proposed eDRAM-based CiM design has performed the potential for high memory density. Exploiting vertical CAA-IGZO device [14], [15], designs and optimizations of 3D eDRAM CiM are proposed in this work to fully exploit the BEOL capability for even higher density.

A. 3D Layout Based on CAA-IGZO TFT

Two layout options with compact $8F^2$/cell footprint are proposed in Fig. 14. The cell comprises of two vertical 2T0C cell, and C_C is implemented by the bottom MIM capacitor. Layout design (i) has a modified access direction of BL, BLB, and WL, while layout design (ii) has a split WL and WLB with shared BL for the write of node N_L and N_R. The former has a shorter write latency, while the latter enables differential refresh with shortened inaccessible time window as shown in Section V-C. Multiple rows and columns of cells form a 2D CiM layer, and CiM layer stacking improves memory density by times.

For clarity, the row and column directions are defined by the CiM operation in this paper, i.e., the inputs are applied by a row-wise scheme, and the results are sensed by a column-wise scheme.

B. 3D-Stacked eDRAM CiM Architecture

Aiming at high scalability and high-throughput output sensing, a 3D-stacked eDRAM CiM architecture is suggested in Fig. 15(a). Multiple CiM layers are stacked above the FEOL Si-based peripherals including input drivers, output ADCs, and subsequent digital logics. For interfacing between BEOL and FEOL, a straightforward approach is to directly connect lines of all CiM layers to FEOL, resulting in large area overhead. Noticed that the CiM computing throughput is determined by the FEOL readout circuits, a layer selection scheme with access IGZO transistors is designed, where only one CiM layer is activated at a time. To minimize the interconnection and routing cost, a CiM layer is partitioned into sub-arrays with local and global line selections. The RLs of each sub-array are connected to corresponding MUXs and ADCs for maximized throughput and minimized energy on interconnection parasitic. The other lines are connected to global lines.

Detailed designs of the selection transistors are shown in Fig. 15(b). Each line in the sub-array is connected to either GND for idle, or global/local lines for computing or write operations. For the local RL, multi-phase multiplexing for the time interleaving scheme in Section V-C is implemented, where a two-set example is demonstrated. With the layer selection design, each CiM layer has four selection signals: GSLB, GSL, RSLB, and RSL, which are the only extra area cost for each newly added CiM layer. For the circuit design, IGZO-based bootstrap switch can further reduce the need for high operating voltage [42]. For the layout design, all BEOL CiM layers have the same pattern except for the layer selection signals. Compact staircase design using minimal incremental layer cost (MiLC) process [43] can be adopted for minimized cost under massive layer stacking.

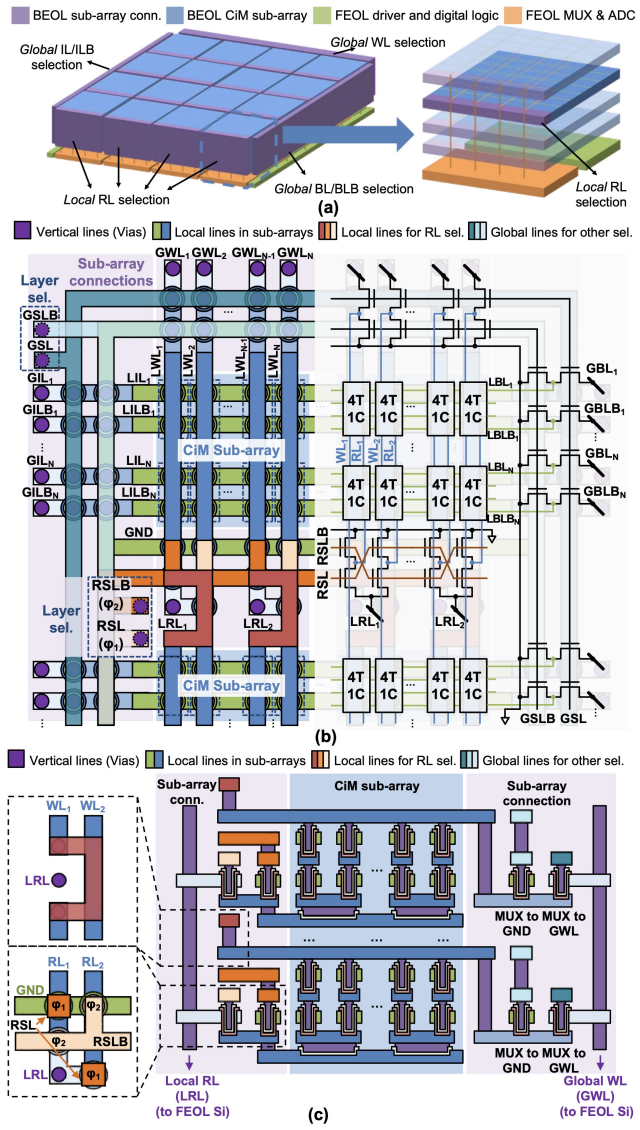


Fig. 15. 3D architecture based on the proposed vertical eDRAM CiM. (a) Overall layout with sub-array partition and routing connection. (b) Top-down view and (c) cross-section view of interconnection design. In the layout, distinct colors signify lines in different metal layers with different functions. Labels prefixed with 'L' and 'G' stand for local lines and global lines, respectively.

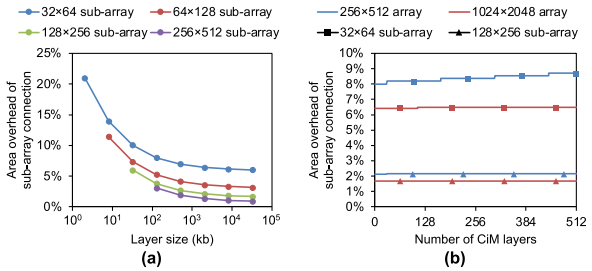


Fig. 16. Evaluation of interconnection area cost in the 3D architecture under (a) different layer sizes and (b) different numbers of CiM layers.

The 3D eDRAM CiM provides both high memory density and computing density. The stacked CiM layers provide high scalability for large NN models. The two-dimensional distributed output peripherals offer high computing density without extra area cost. Furthermore, short vertical wire interconnections retain high energy efficiency for charge-domain CiM operation. The area cost of the routing connection

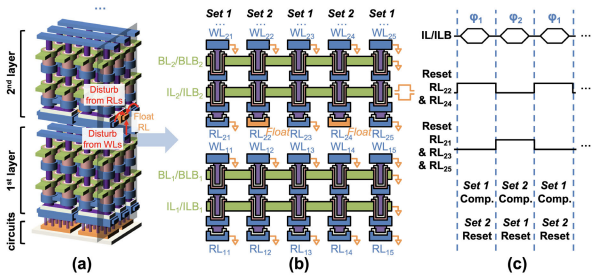


Fig. 17. Proposed time-interleaved computing scheme. (a) The compact 3D stacking of proposed CiM layers, where disturbances from adjacent lines to the floating RL are severe. (b) Proposed time-interleaved computing scheme and (c) corresponding timing diagram with a 2-set example [17].

scheme is evaluated in Fig. 16. Under multiple sub-array configurations, the area overheads reach different lower bounds of less than 10% with the increasing layer size. As the number of CiM layers increases, the area cost increment is slow, indicating the scalability for massive layer stacking.

C. Time-Interleaved Computing Scheme

Under the high-density 3D array, the disturbance to RLs can be severe. In the 3D layout, the floating RL for CiM operation cannot be adequately protected by node X, which can have significant parasitic capacitance to horizontally adjacent RLs and vertically adjacent WLs. Although the disturbance from WLs will only introduce gain errors, the disturbance from nearby RLs during CiM operation causes data-dependent errors, which can severely affect computing accuracy.

To tackle this issue, a time-interleaved CiM computing scheme is proposed as shown in Fig. 17. The columns are separated into alternating sets, and the operation is divided into corresponding phases. Thus, the computing for the whole sub-array is accomplished in multiple phases. In each phase, the RLs of one set are floating for CiM computing, with the others grounded. The inputs are applied by differential input pattern (ii) or (iii) illustrated in Section III-A. Fig. 17(b) gives an example of two sets using layout design (i). With the proposed scheme, the nearby ILs and ILBs have a differential voltage, which introduces negligible disturbance to the single-end RL. The other nearby lines have fixed voltages. Therefore, the nearby lines around a floating RL provide approximately shielding. The remaining lumped parasitic capacitance only contributes to gain errors, which can be reduced through foreground calibration.

D. Differential Refresh Scheme

Besides the leakage-induced data destruction, the V_{TH} shift (ΔV_{TH}) of TFT device under a long-term bias on node N_L and N_R can also severely degrade the storage capability. When T_2 and T_3 have a large discrepancy of ΔV_{TH}, the cell cannot readout correct weight value, as the cell on-off ratio corresponding to '0' and '1' can differ a lot. Recent CiM works tackle reliability issues of emerging devices by calibration methods [44].

Fully exploiting the benefit of the differential 4T1C structure, this work proposes a differential refresh scheme, as illustrated in Fig. 18. The stored voltages are refreshed with alternate normal and inverted values. The inputs are also inverted to ensure correct CiM results. Therefore, gate biases on T_2 and T_3 become alternating high or low voltages rather

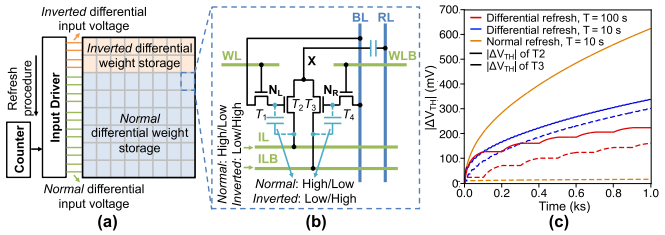


Fig. 18. Proposed differential refresh scheme with (a) array operation and (b) cell operation. (c) Simulation of V_{TH} shift illustrating the principle of the differential refresh. Temperature = 393K, V_{pp} of BL = 3.5 V, V_{pp} of WL = 4 V.

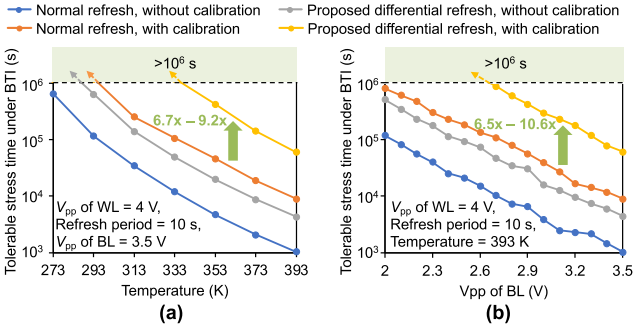


Fig. 19. Simulations of the proposed differential refresh scheme. (a) Tolerable stress time under different temperatures. (b) Tolerable stress time under different peak-to-peak voltage of BL.

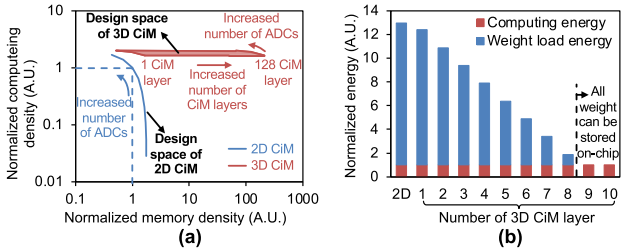


Fig. 20. Comparison of 3D CiM and its 2D counterpart, with the same cell area and peripherals. (a) Design space of 2D and 3D CiM considering 32 × 64 sub-array, 256 × 512 array, up to 256 ADC parallelism, and up to 128 stacked CiM layers. (b) Application-level average energy consumption of the 2D CiM and 3D CiM with different numbers of layers. The initial weight load is averaged and thus neglected. 20 pJ/bit for external DRAM read is estimated [48].

than long-term fixed voltages. Therefore, the ΔV_{TH} of the T_2 and T_3 is balanced, with charge loss also compensated. The differential refresh scheme can tolerate deterministic ΔV_{TH} with either positive or negative trends, thanks to the symmetrically applied gate biases on T_2 and T_3. The differential refresh can be supported by SA with additional MUXs to provide an inverted write path.

To showcase the principle of the proposed differential refresh, simulations of ΔV_{TH} are carried out in Fig. 18(c). CAA-IGZO model from [45] is adopted with L = 50 nm, CD = 130 nm, and BTI effect captured. The BTI for CAA-IGZO device under positive bias appears as negative ΔV_{TH} shift, explained as electron de-traps from deep states and hydrogen release [45]. With normal refresh scheme that never inverts weight data, the ΔV_{TH} difference between T_2 and T_3 increases rapidly under a high temperature. With proposed differential refresh, although the ΔV_{TH} still exists, the ΔV_{TH} difference between T_2 and T_3 can be sufficiently small to keep the accurate on-off relationship. For <0.1 V ΔV_{TH} difference, a refresh period of can be adopted.

Under the large 3D memory capacity, the access blockage caused by an improperly designed array-level differential refresh scheme can severely influence the read and CiM operation. Due to the mismatched inverted or normal weights and inputs, a partially inverted memory array may generate erroneous results. Consequently, a straightforward differential refresh procedure is uninterruptable, resulting in expensive worst-case stalls. To support an optimized array-level refresh, layout design (ii) with both horizontal WL/WLB and IL/ILB is preferred, as shown in Fig. 18(a). With the row-wise partially inverted stored weights, CiM operations can still be executed as long as inputs are inverted correspondingly. The counter records the inverted rows to control which inputs should be inverted. With the schedule, the uninterruptible memory blockage is reduced from $O(M \times S)$ to $O(1)$, where M is the number of rows and S is the number of layers in a CiM array.

Fig. 19 shows the simulation results that validate the effectiveness of the proposed differential refresh. The tolerable stress time is defined as the contiguous operating time that the 4T1C array can perform 32-row CiM computing within 0.5LSB error. For different temperatures and V_{pp} of BL, the differential refresh leads to over 4x improvement. Furthermore, the calibration of BL can improve the accuracy of both refresh schemes by recovering V_{ov} of T_2 and T_3 for the proper cell on-off ratio. A reference cell tracks the ΔV_{TH} with the same refresh period as the array and alternating write voltages for each refresh. The ΔV_{TH} of T_2 and T_3 of the reference cell is tracked and averaged, then applied as a calibration offset to write V_{BL}. The calibration improves the time by an order of magnitude, and proposed differential refresh outperforms the normal one by over 6x. The remaining asymmetry comes from the initial ΔV_{TH} difference, and leakage-induced storage degradation caused by T_1 and T_4 with shifted V_{TH}. Therefore, higher differential refresh frequency and additional calibration of V_{WL} can further improve the reliability. The ongoing advancement of IGZO device with high bias temperature stability [46], [47] will enable a more reliable eDRAM CiM computing.

E. Comparison Between 2D and 3D CiM

Fig. 20 showcases the superiority of 3D CiM integration over its 2D counterpart, particularly with regard to memory density, computing density, and energy efficiency. Fig. 20(a) illustrates the design space considering varied output ADC parallelism and stacked CiM layers. The results indicate the potential for high computing density through peripherals under array, and massively improved memory density via stacked layers. Macro-level energy efficiency benefits come from the short vertical interconnection of BEOL and FEOL circuitry, which is related to the process parameters. Application-level energy efficiency is enhanced by the high-capacity on-chip storage. Fig. 20(b) reveals that, under the same area of 2 mm^2, 3D CiM outperforms its 2D counterpart in a continuous NN acceleration scenario, e.g., the processing of image frames. 4-bit ResNet-50 is assessed with parameters fixed on-chip as much as possible and a weight-stationary computing strategy. With stacking, the 3D CiM can store most or the entire model, substantially reducing heavy weight reloads from external DRAM.

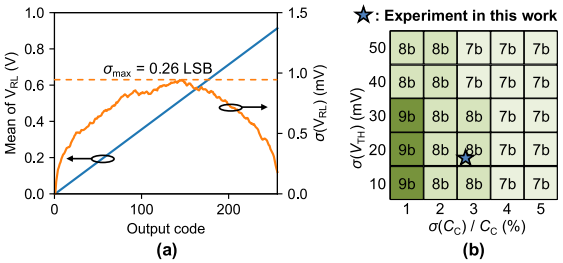


Fig. 21. Monte-Carlo simulation of (a) 8-bit MAC, and (b) output precision under different $\sigma(V_{TH})$ and $\sigma(C_C)$.

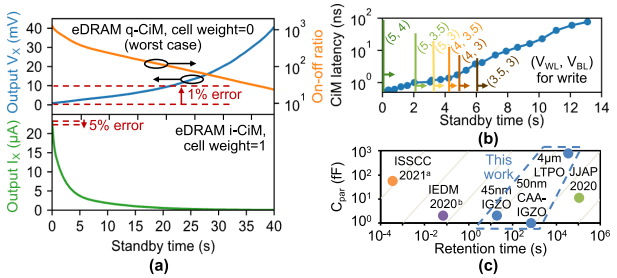


Fig. 22. Impact of leakage-induced charge loss. (a) Degraded output at node X of an eDRAM CiM cell. (b) Increased read delay. (c) Estimated retention time with different storage node C_{par}. a^{C}_{par} estimated by reported cell size; b^{C}_{par} and retention estimated by the same 45nm IGZO model, storage node capacitance and number of accumulation rows as those in this work.

VI. Performance Analysis and Application Benchmark

In this section, circuit-level and application-level evaluations are carried out to further explore the design space. Model extracted from scaled 45 nm IGZO devices [16] is adopted. Circuit parameters are set as C_{par} = 2 fF, C_C = 10 fF.

A. Circuit-Level Analysis Considering Non-Ideal Factors

Non-ideal factors, mainly including device mismatch and leakage, influence the metrics of eDRAM CiM. Fig. 21(a) shows Monte-Carlo simulations of a 256-row MAC considering the process mismatch of $\sigma(V_{TH})$ = 22 mV and $\sigma(C_C)/C_C$ = 3.1%. The achievable computing accuracy is defined as 3σ(INL) < 1 LSB, thus 8-bit accuracy is observed. Fig. 21(b) further investigates the accuracy of the proposed charge-domain eDRAM CiM under different device variations. For processes with well-matched passive capacitors, the proposed CiM can achieve high computing accuracy regardless of V_{TH} mismatch.

The impact of leakage-induced charge loss is shown in Fig. 22. The on-off ratio of R_L/R_R (or R_R/R_L) degrades with the standby time, which causes deviation of output V_X from the ideal $V_{IL,H}$ or $V_{IL,L}$ of each cell. However, the V_X degradation is considerably slower than that of current-domain CiM [10], [49], because the resistance ratio other than the absolute on-state current is much less sensitive to charge loss on the internal node. Also, decreased node voltage causes enlarged on-resistance, thus increasing CiM operation latency. Fig. 22(c) compares the retention time of CiM works with different C_{par}. Under the same technology node and circuit design, a larger C_{par} can lead to extended retention time with the cost of area. It is preferred to achieve longer retention under a smaller C_{par}, which can enable both lower eDRAM refresh frequency and higher memory density. It can be achieved by both device leakage optimizations and robustness-enhanced circuit designs.

With the continuous scaling of IGZO device, the severe parasitic-induced interference raises as a challenge, where fringe capacitance and adjacent coupling capacitance accounted for an increased portion. It introduces energy cost and affects the computing accuracy from both intra-cell weight storage and inter-cell accumulation. For intra-cell parasitic, large coupling capacitance redistributes the charge on C_{par} during a CiM operation, causing a reduced voltage and thus a degraded cell on-off ratio. Therefore, the V_{IL} range should be smaller than the write V_{BL} to guarantee sufficient V_{GS} for on-state T_2 or T_3. For inter-cell parasitic, the floating RL is sensitive to the crosstalk especially from adjacent RLs. For advanced planar IGZO process, an optimized 4T1C layout introduces <0.4% coupling capacitance from adjacent RLs for 10 fF C_C. However, the adjacent RLs in compact CAA-IGZO design are much closer which introduce ~15% coupling capacitance. Proposed time-interleaved computing scheme provides approximate shielding, where a two-set operation can reduce the interference by >10x.

Challenges also raised for massive BEOL CiM layer stacking. Thermal performance is a concern for monolithic 3D chips. In the proposed design, since CiM layers are sequentially activated, the power density is nearly constant for varied CiM layers. The power density is estimated to be 26 mW/mm^2, mainly contributed by FEOL peripherals. As the power density increases with advanced device scaling, efficient cooling architectures, e.g., dual-sided microfluidic, can be adopted [50]. Other challenges include increased coupling capacitance of vertical interconnections, inter-layer performance mismatch, etc. To this end, further calibration methods should be explored.

B. Application-Level Benchmark of NN Accelerator for Edge AI

To evaluate the advantages of the proposed IGZO eDRAM CiM in typical edge AI scenarios, a near-sensor NN accelerator architecture is adopted as shown in Fig. 23(a). Image classification task with either low or high event activity exploiting VGG-8 network on CIFAR-10 dataset is investigated. The input pattern of the CiM sub-arrays is configured as (i), with the inputs from either analog sensor data or feature map of the previous layer. Four CiM designs are benchmarked: proposed IGZO eDRAM charge-domain CiM (q-CiM), Si eDRAM q-CiM with the same cell structure in Fig. 6, eDRAM current-domain CiM (i-CiM) [10], and SRAM q-CiM [13]. The output dynamic range, the total bitline capacitance, and storage node capacitance are set the same for comparisons.

Fig. 23(b) evaluates the accuracy degradation with standby time of IGZO q-CiM and i-CiM. The framework performs bit-accurate evaluations with 4-bit post-training quantized weight mapped to the 128 × 128 CiM array and feature map quantized to 5-bit analog input. The leakage-induced variation on V_X is modeled as a Gaussian-distributed noise to the weight bits. The outputs are quantized by 8-bit. For i-CiM, although frequent refresh is not a must for memory mode, high refresh frequency is required for CiM computing to guarantee high accuracy. Under V_{TH} variation, q-CiM performs significant advantages of >50x standby time without a refresh over i-CiM. Fig. 23(c) further studies the effect of device's BTI. For normal refresh, CiM array has degraded weight storage capacity under long-

TABLE II
COMPARISON WITH EXISTING WORKS

	This work			IEDM 2020[10]	JJAP 2020[11]	ISSCC 2021[32]	ISSCC 2021[13]
Technology	**45nm IGZO**	**4μm LTPO**	**CAA-IGZO L=50nm, CD=130nm**	22nm IGZO	350nm IGZO + 110 nm Si	65nm Si	16nm Si
CiM scheme	**eDRAM q-CiM**			eDRAM i-CiM	eDRAM i-CiM	eDRAM i-CiM	SRAM q-CiM
Measured array size	**Single cell, 128 ×128 evaluated**	**128×32**	**No chip, 32×64 evaluated**	No chip	9×16	64×32	1152×256
Weights storage	**Binary**			Binary	Analog	Analog	Binary
Precision (Weights /Inputs/Outputs)	**4/Analog[l]/8**			Ternary/8/16	Analog/Analog/4	4/4/5	1-8/1-8/8
Peak computing energy efficiency (TOPS/W)	**686[af], 138[bef]**	**0.728[a]**	**3718[ag], 43.0[beg]**	1800[a], 15[b]	5[a]	102.2[bd]	121[bc]
Peak MAC computing density (TOPS/mm^2)	**9.76[aef], 1.87[bef]**	**>5.17×10^{-4} [a]**	**46.2[ae], 1.12[beg]**	37[a], 4[b]	1.02[a]	8.58[bd]	2.67[bc]
Memory density (Mb/mm^2)	**0.745[af], 0.321[bef]**	**1.48×10^{-4} [a]**	**7.05[a], 6.41[beg] (per CiM layer)**	3.67[a]	0.0193[a]	0.156[b]	1.02[b]
Retention time for CiM	**20 s[h]**	**~10 h[h]**	**720 s[h]**	72 ms[hj]	>30 h[i]	0.36 ms[h]	-
Main factors that affect computing robustness	**Capacitor mismatch**			PVT- and leakage-induced on-current variation	Deviation from quadratic device characteristic; PVT- and leakage-induced on-current variation	PVT- and leakage-induced on-current variation	Capacitor mismatch
NN accuracy	**88.24% (VGG-8 on CIFAR-10); 65.10% top-1, 87.00% top-5[k] (ResNet-18 on CIFAR-100)**			-	96.25% (3-layer fully connected network on MNIST)	91% (VGG-16 on CIFAR-10)	91.51%[c] (VGG-11 on CIFAT-10); 73.04%[ck] (ResNet-50 on ImageNet)

[a] CiM array only; [b] CiM array + CiM peripherals; [c] 4/4/8 config; [d] without sparsity saving; [e] 40 ns CiM operation latency, 8b 20MS/s output ADC under 16 nm Si, with 216 μm^2 area and 346 fJ per conversion [13]; [f] C_{par} = 2fF, C_C = 10fF, cell area estimated by layout under commercial Si process; [g] C_{par} = 0.6fF, C_C = 0.8fF, 10% interconnection area cost considered; [h] defined as the time at 0.5 LSB output error with the evaluated number of rows in a MAC; [i] defined as 5% degradation for analog multiplication; [j] estimated by the same 45nm IGZO model, C_{par} and number of accumulation rows as those in this work; [k] projection shortcuts in ResNet employs higher precision [13]. [l] the precision is determined by the application. In NN evaluation of this paper, the input is either from analog sensor or 5b DAC.

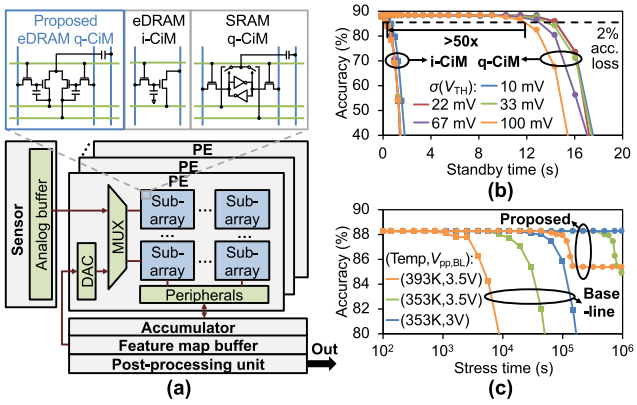


Fig. 23. NN application evaluation. (a) Adopted NN accelerator architecture for edge AI, with the cell configured as the proposed q-CiM or baseline i-CiM. (b) NN accuracy vs standby time under different $\sigma(V_{TH})$. (c) NN accuracy vs stress time under different temperature and V_{pp} of BL. Baseline: normal refresh without calibration; Proposed: differential refresh with V_{BL} calibration.

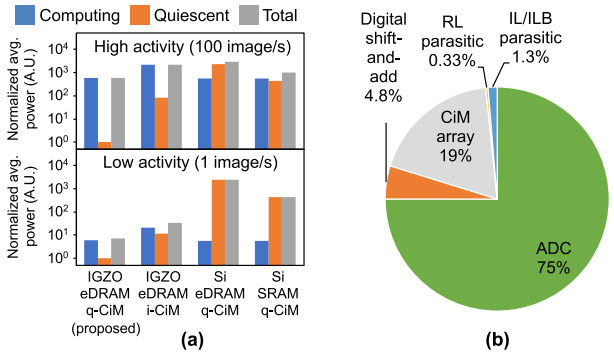


Fig. 24. (a) Comparison with existing CiM schemes in two typical event activities in edge AI scenarios. Normalized to the average quiescent power of IGZO eDRAM q-CiM. (b) Computing power breakdown of proposed design.

term stress. Proposed differential refresh with V_{BL} calibration improves the stability, where the symmetric stress process prevents rapid accuracy loss after a long stress time.

Fig. 24(a) compares the application-level average power of the four designs. The quiescent power includes the refresh cost of eDRAM, or the leakage power of SRAM. The computing power includes CiM core, output 8-bit ADCs, and digital shift-and-add for 4-bit weight, where the breakdown for this work is shown in Fig. 24(b). Under low event activity of 1 image/s, quiescent power dominates for Si-based CiMs, leading to a much higher power than low-leakage IGZO-based CiMs. And for IGZO-based designs, the proposed q-CiM has a lower refresh frequency and thus reduced power than i-CiM. Under high event activity of 100 image/s, IGZO-based designs still benefit from the low leakage. In this case when the computing power dominates, q-CiM exhibits superior power efficiency than i-CiM with the lower capacitor charging energy.

C. Comparison With Existing Works

Table II summarizes the comparison between the proposed 4T1C eDRAM CiM and existing IGZO-based and Si-based works. Configurations for the proposed 4T1C eDRAM CiM are unified across different processes, where 0.8 V dynamic range for IL/ILB is adopted except for LTPO, and 0.22 fF/μm wire parasitic is included. In this paper, an 'operation' is defined as either a 4-bit weight multiplying an analog input, or an addition of the multiplication result. Besides, our previous works [16], [17] focus on the metrics with the binary-cell operation defined with 1-bit weight and analog input.

Compared with Si-based designs, BEOL stacking enables both high memory and computing density. Compared with the

existing IGZO-based designs, high robustness and simplified control enable high scalability, which is verified by a taped-out 128 × 32 array. DC-power-free charge-domain CiM operation observes low computing energy, while the differential storage scheme enhances the retention time. The proposed IGZO eDRAM q-CiM under scaled 45nm IGZO technology achieves a high energy efficiency of 686 TOPS/W for array only and 138 TOPS/W with peripherals.

D. Future Work

The effectiveness of the proposed 4T1C eDRAM CiM has been validated by both simulations and taped-out measurements. In this paper, the proof-of-concept array-level integration exploits foundry 4μm LTPO technology with high-yield large-area fabrication, demonstrating long retention and high computing accuracy. With the active progress of advanced IGZO processes, future array-level tape-out chips with scaled devices will further verify the performance of proposed IGZO CiM, benefitting from lower operation voltages and capacitance.

Advanced IGZO device scaling for eDRAM CiM still faces challenges from variation issues, short-channel effects (SCEs), and retention degradation. To address the device variation issues, advanced lithography techniques should be employed. Furthermore, SCEs, including subthreshold swing (*SS*) degradation and drain-induced barrier lowering (DIBL), cause increased power consumption. Nevertheless, SCEs can be mitigated by device optimization, i.e., thickness reduction of IGZO layer. Degradation of eDRAM retention also occurs with scaling, due to scaled C_{par} and slightly increased gate leakage. However, better retention than i-CiM can still be achieved when performing comparisons in the same technology node.

Aside from the device scaling, there are other aspects for exploration. First, to address device-level non-idealities of TFT, circuit and algorithm techniques against device instability are to be enriched. Second, support for multi-level-cell (MLC) eDRAM charge-domain CiM remains to be explored. A straightforward method to implement multiple levels in 4T1C is to control the voltage level written to node N_L and N_R to achieve multi-level on-off ratios. However, it will not maintain the high immunity to variations, and also, introduce DC current in CiM operation. Third, calibration methods addressing the challenges in multi-layer 3D IGZO CiM are to be investigated. Moreover, the reconfigurable dataflow with the 3D CiM structure for NN operator implementation is to be optimized.

VII. Conclusion

This paper presents IGZO eDRAM-based charge-domain CiM, which provides a solution for robust and scalable CiM design. The proposed design performs high energy efficiency through reduced data refresh frequency and dynamic charge-domain computing. The differential cell structure enables robust high-accuracy computing. Compact vertical 4T1C cell design and 3D layer stacking with emerging CAA-IGZO technology facilitate high memory density. Additionally, circuit techniques including differential refresh and time-interleaved computing ensure high computing accuracy in the large-scale 3D CiM array. With the measurement of a taped-out 128 × 32 array, the scalability of the proposed CiM scheme is verified. Benchmarking results indicate the superiority of the proposed CiM design over existing works in terms of energy efficiency, memory density, and computing density.

References

[1] K. Bong, S. Choi, C. Kim, S. Kang, Y. Kim, and H.-J. Yoo, "14.6 A 0.62 mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on Haar-like face detector," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 248–249.

[2] R. Guo et al., "A 5.1 pJ/Neuron 127.3us/Inference RNN-based speech recognition processor using 16 computing-in-memory SRAM macros in 65 nm CMOS," in *Proc. Symp. VLSI Circuits*, Jun. 2019, pp. 120–121.

[3] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury, "A 55 nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 124–126.

[4] J. Backus, "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs," *Commun. ACM*, vol. 21, no. 8, pp. 613–641, Aug. 1978.

[5] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang, and M.-F. Chang, "Challenges and trends of SRAM-based computing-in-memory for AI edge devices," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 5, pp. 1773–1786, May 2021.

[6] S. Xie, C. Ni, P. Jain, F. Hamzaoglu, and J. P. Kulkarni, "Gain-cell CIM: Leakage and bitline swing aware 2T1C gain-cell eDRAM compute in memory design with bitline precharge DACs and compact Schmitt Trigger ADCs," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 112–113.

[7] J.-M. Hung, C.-J. Jhang, P.-C. Wu, Y.-C. Chiu, and M.-F. Chang, "Challenges and trends of nonvolatile in-memory-computation circuits for AI edge devices," *IEEE Open J. Solid-State Circuits Soc.*, vol. 1, pp. 171–183, 2021.

[8] J.-W. Su et al., "16.3 A 28 nm 384 kb 6T-SRAM computation-in-memory macro with 8 b precision for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 250–252.

[9] A. Belmonte et al., "Capacitor-less, long-retention (>400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM," in *IEDM Tech. Dig.*, Dec. 2020, pp. 28.2.1–28.2.4.

[10] P. Houshmand et al., "Opportunities and limitations of emerging analog in-memory compute DNN architectures," in *IEDM Tech. Dig.*, Dec. 2020, pp. 29.1.1–29.1.4.

[11] Y. Kurokawa et al., "CAAC-IGZO FET/Si-FET hybrid structured analog multiplier and vector-by-matrix multiplier for neural network," *Jpn. J. Appl. Phys.*, vol. 59, Feb. 2020, Art. no. SGGB03.

[12] J. Liu, W. Tang, Y. Liu, H. Yang, and X. Li, "Almost-nonvolatile IGZO-TFT-based near-sensor in-memory computing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.

[13] H. Jia et al., "Scalable and programmable neural network inference accelerator based on in-memory computing," *IEEE J. Solid-State Circuits*, vol. 57, no. 1, pp. 198–211, Jan. 2022.

[14] X. Duan et al., "Novel vertical channel-all-around (CAA) In-Ga-Zn-O FET for 2T0C-DRAM with high density beyond $4F^2$ by monolithic stacking," *IEEE Trans. Electron Devices*, vol. 69, no. 4, pp. 2196–2202, Apr. 2022.

[15] K. Huang et al., "Vertical channel-all-around (CAA) IGZO FET under 50 nm CD with high read current of 32.8 μA/μm (Vth + 1V), well-performed thermal stability up to 120 °C for low latency, high-density 2T0C 3D DRAM application," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 296–297.

[16] J. Liu et al., "Low-power and scalable retention-enhanced IGZO TFT eDRAM-based charge-domain computing," in *IEDM Tech. Dig.*, Dec. 2021, p. 21.1.1–21.1.4.

[17] W. Tang, J. Liu, H. Yang, C. Jiang, and X. Li, "High-density energy-efficient charge-domain computing based on CAA-IGZO TFT with BEOL-compatible 3D integration," in *Proc. IEEE Int. Flexible Electron. Technol. Conf. (IFETC)*, Aug. 2022, pp. 1–2.

[18] Q. Hu et al., "Optimized IGZO FETs for capacitorless DRAM with retention of 10 ks at RT and 7 ks at 85 °C at zero V hold with sub-10 ns speed and 3-bit operation," in *IEDM Tech. Dig.*, Dec. 2022, pp. 26.6.1–26.6.4.

[19] K. Kato et al., "Evaluation of off-state current characteristics of transistor using oxide semiconductor material, indium–gallium–zinc oxide," *Jpn. J. Appl. Phys.*, vol. 51, no. 2R, Jan. 2012, Art. no. 021201.

[20] T. Matsuzaki et al., "A 16-level-cell nonvolatile memory with crystalline In-Ga-Zn oxide FET," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2015, pp. 1–4.

[21] J. Biggs et al., "A natively flexible 32-bit arm microprocessor," *Nature*, vol. 595, no. 7868, pp. 532–536, Jul. 2021.

[22] H. Çeliker, A. Sou, B. Cobb, W. Dehaene, and K. Myny, "Flex6502: A flexible 8 b microprocessor in 0.8 μm metal-oxide thin-film transistor technology implemented with a complete digital design flow running complex assembly code," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 65, Feb. 2022, pp. 272–274.

[23] M.-C. Chen et al., "A >64 multiple states and >210 TOPS/W high efficient computing by monolithic Si/CAAC-IGZO + super-lattice ZrO_2/Al_2 O_3/ZrO_2 for ultra-low power edge AI application," in *IEDM Tech. Dig.*, Dec. 2022, pp. 18.2.1–18.2.4.

[24] C. Wang et al., "Extremely scaled bottom gate a-IGZO transistors using a novel patterning technique achieving record high gm of 479.5 μS/μm (VDS of 1V) and fT of 18.3 GHz (VDS of 3V)," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 294–295.

[25] Q. Kong et al., "New insights into the impact of hydrogen evolution on the reliability of IGZO FETs: Experiment and modeling," in *IEDM Tech. Dig.*, Dec. 2022, pp. 30.2.1–30.2.4.

[26] Q. Ma et al., "Robust gate driver on array based on amorphous IGZO thin-film transistor for large size high-resolution liquid crystal displays," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 717–721, 2019.

[27] T.-K. Chang, C.-W. Lin, and S. Chang, "39-3: $Invited Paper$: LTPO TFT Technology for AMOLEDs?" in *SID Symp. Dig. Tech. Papers*, vol. 50, no. 1, 2019, pp. 545–548.

[28] *Tianma Microelectronics*. Accessed: Jun. 3, 2023. [Online]. Available: https://en.tianma.com/

[29] J. Song et al., "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," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2023, pp. 1–2.

[30] X. Si et al., "A local computing cell and 6T SRAM-based computing-in-memory macro with 8-b MAC operation for edge AI chips," *IEEE J. Solid-State Circuits*, vol. 56, no. 9, pp. 2817–2831, Sep. 2021.

[31] W. Li, X. Sun, S. Huang, H. Jiang, and S. Yu, "A 40-nm MLC-RRAM compute-in-memory macro with sparsity control, on-chip write-verify, and temperature-independent ADC references," *IEEE J. Solid-State Circuits*, vol. 57, no. 9, pp. 2868–2877, Sep. 2022.

[32] Z. Chen, X. Chen, and J. Gu, "15.3 A 65 nm 3T dynamic analog RAM-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44TOPS/W system energy efficiency," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 240–242.

[33] Y. Chen et al., "YOLoC: Deploy large-scale neural network by ROM-based computing-in-memory using residual branch on a chip," in *Proc. 59th ACM/IEEE Design Autom. Conf.*, San Francisco, CA, USA, Jul. 2022, pp. 1093–1098.

[34] X. Peng et al., "Benchmarking monolithic 3D integration for compute-in-memory accelerators: Overcoming ADC bottlenecks and maintaining scalability to 7 nm or beyond," in *IEDM Tech. Dig.*, Dec. 2020, pp. 30.4.1–30.4.4.

[35] Q. Huo et al., "A computing-in-memory macro based on three-dimensional resistive random-access memory," *Nature Electron.*, vol. 5, no. 7, pp. 469–477, Jul. 2022.

[36] W. Shim and S. Yu, "Technological design of 3D NAND-based compute-in-memory architecture for GB-scale deep neural network," *IEEE Electron Device Lett.*, vol. 42, no. 2, pp. 160–163, Feb. 2021.

[37] H.-W. Hu et al., "A 512 Gb in-memory-computing 3D-NAND flash supporting similar-vector-matching operations on edge-AI devices," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 65, Feb. 2022, pp. 138–140.

[38] W. Tang et al., "Computing-in-memory with thin-film-transistors: Challenges and opportunities," *Flexible Printed Electron.*, vol. 7, no. 2, Feb. 2022, Art. no. 024001.

[39] S. R. Sundara Raman, S. Xie, and J. P. Kulkarni, "IGZO CIM: Enabling in-memory computations using multilevel capacitorless indium–gallium–zinc–oxide-based embedded DRAM technology," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 8, no. 1, pp. 35–43, Jun. 2022.

[40] Y. Luo et al., "A compute-in-memory hardware accelerator design with back-end-of-line (BEOL) transistor based reconfigurable interconnect," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 12, no. 2, pp. 445–457, Jun. 2022.

[41] S. Angizi and D. Fan, "ReDRAM: A reconfigurable processing-in-DRAM platform for accelerating bulk bit-wise operations," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD)*, Westminster, CO, USA, Nov. 2019, pp. 1–8.

[42] Y. Yakubo et al., "Crystalline oxide semiconductor-based 3D bank memory system for endpoint artificial intelligence with multiple neural networks facilitating context switching and power gating," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2023, pp. 12–14.

[43] Z. Jiang et al., "Next-generation ultrahigh-density 3-D vertical resistive switching memory (VRSM)—Part II: Design guidelines for device, array, and architecture," *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5147–5154, Dec. 2019.

[44] M. Lee et al., "Victor: A variation-resilient approach using cell-clustered charge-domain computing for high-density high-throughput MLC CiM," in *Proc. 60th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2023, pp. 1–6.

[45] J. Guo et al., "Compact modeling of IGZO-based CAA-FETs with time-zero-instability and BTI impact on device and capacitor-less DRAM retention reliability," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 300–301.

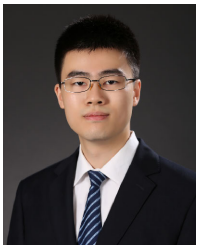
[46] W. Kim et al., "Demonstration of crystalline IGZO transistor with high thermal stability for memory applications," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2.

[47] J. Zhang et al., "First demonstration of BEOL-compatible atomic-layer-deposited InGaZnO TFTs with 1.5 nm channel thickness and 60 nm channel length achieving ON/OFF ratio exceeding 1011, SS of 68 mV/dec, normal-off operation and high positive gate bias stability," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2.

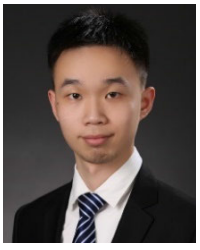
[48] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

[49] D. Saito et al., "IGZO-based compute cell for analog in-memory computing—DTCO analysis to enable ultralow-power AI at edge," *IEEE Trans. Electron Devices*, vol. 67, no. 11, pp. 4616–4620, Nov. 2020.

[50] A. Kaul, Y. Luo, X. Peng, S. Yu, and M. S. Bakir, "Thermal reliability considerations of resistive synaptic devices for 3D CIM system performance," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, Oct. 2021, pp. 1–5.



Wenjun Tang (Graduate Student Member, IEEE) received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His current research interests include TFT-based in-memory and in-sensor computing circuit designs for edge AI.



Jialong Liu (Graduate Student Member, IEEE) received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His current research interests include thin-film transistor (TFT) circuit design and in-memory/in-sensor computing architecture design based on large-area electronics.

Chen Sun (Graduate Student Member, IEEE) received the B.Eng. degree in electronic information engineering from the University of Electronic Science and Technology of China (UESTC) in 2018. He is currently pursuing the Ph.D. degree with the ECE Department, National University of Singapore (NUS). He has published top journals and conference papers, including IEEE ELECTRON DEVICE LETTERS, IEEE TRANSACTIONS ON ELECTRON DEVICES, Symposium on VLSI, and IEDM. His research interests include high-performance oxide-semiconductor-based thin-film transistors (TFTs), ferroelectric field-effect transistors (FeFETs), and low-power and high-retention DRAM using oxide semiconductor TFTs. He has received the Best Paper Award of ICICDT 2019.

Zijie Zheng (Graduate Student Member, IEEE) received the B.Eng. degree in electrical engineering and automation from Xian Jiaotong University, Xi'an, China, in 2020. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering (ECE), National University of Singapore (NUS). His current research interests include thin-film transistor based on oxide semiconductor materials and ferroelectric/anti-ferroelectric devices for data storage and neuromorphic computing.

Yongpan Liu (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Tsinghua University, China, in 1999, 2000, and 2007, respectively.

He is currently a Full Professor (Cheung Kong Scholar) with the Department of Electronic Engineering, Tsinghua University.

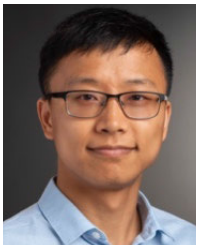
Prof. Liu is a Program Committee Member of ISSCC, A-SSCC, and DAC. He has received under 40 Young Innovators Award DAC 2017, Best Paper/Poster Award from ASPDAC 2021 and 2017, Micro Top Pick 2016, HPCA 2015, and Design Contest Awards of ISLPED 2012, 2013, and 2019. He served as the General Secretary for ASPDAC 2021 and the Technical Program Chair for NVMSA 2019. He was an Associate Editor of IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-II: EXPRESS BRIEFS, and the *IET Cyber-Physical Systems*. He served as an A-SSCC2020/AICAS2022 Tutorial Speaker and an IEEE CASS Distinguished Lecturer in 2021.

Huazhong Yang (Fellow, IEEE) received the B.S. degree in microelectronics and the M.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1989, 1993, and 1998, respectively.

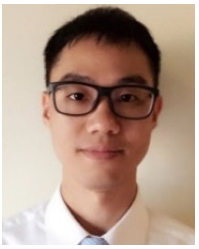
In 1993, he joined the Department of Electronic Engineering, Tsinghua University, where he has been a Full Professor since 1998. He has been in charge of several projects, including projects sponsored by the National Science and Technology Major Project, the 863 Program, NSFC, and several international research projects. He has authored or coauthored over 500 technical articles, seven books, and over 180 granted Chinese patents. His research interests include wireless sensor networks, data converters, energy-harvesting circuits, nonvolatile processors, and brain-inspired computing.

Prof. Yang received the Distinguished Young Researcher by NSFC in 2000, the Cheung Kong Scholar by the Chinese Ministry of Education (CME) in 2012, the Science and Technology Award First Prize by the China Highway and Transportation Society in 2016, the Technological Invention Award First Prize by CME in 2019, the Gold Prize of iNEA 2019, and several best paper awards, including ISVLSI 2012, FPGA 2017, NVMSA 2017, and ASP-DAC 2019. He has served as the Chair for the Northern China ACM SIGDA Chapter Science in 2014, the General Co-Chair for ASP-DAC 2020, a Navigating Committee Member for AsianHOST'18, and a TPC Member for ASP-DAC 2005, APCCAS 2006, ICCCAS 2007, ASQED 2009, and ICGCS 2010.

Chen Jiang (Member, IEEE) received the B.S. degree in engineering from the Department of Electronic Engineering, Shanghai Jiao Tong University, China, and the Ph.D. degree in engineering from the Department of Engineering, University of Cambridge, U.K. He is currently an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. After this, he carried out post-doctoral research as a Wellcome Trust Junior Interdisciplinary Fellow with the Department of Clinical Neurosciences, University of Cambridge. His current research interests are focused on novel electronic device architectures, large-area flexible transparent electronics, low-power circuits, and their applications to wearable/implantable sensing/stimulating interfaces and bioelectronic systems.



Kai Ni (Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2011, and the Ph.D. degree in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2016, by working on characterization, modeling, and reliability of III-V MOSFETs. Since then, he became a Post-Doctoral Associate with the University of Notre Dame, working on ferroelectric devices for nonvolatile memory and novel computing paradigms. He was an Assistant Professor in electrical and microelectronic engineering with the Rochester Institute of Technology. He joined University of Notre Dame in 2023. He has 80 publications in top journals and conference proceedings, including *Nature Electronics*, IEDM, VLSI Symposium, IRPS, and EDL. His current interests lie in nanoelectronic devices empowering unconventional computing, AI accelerator, and 3D memory technology.



Xiao Gong (Member, IEEE) received the Ph.D. degree from the National University of Singapore (NUS) in 2013. He was a Visiting Scientist with MIT in 2014. He is currently an Assistant Professor with the ECE Department, NUS. He has more than 250 publications in international journals and conferences, including more than 55 papers in IEDM and VLSI Symposium. His research interests include advanced transistors and emerging memories for in-memory computing, monolithic 3D integration, opto-electronic integrated circuits and their applications in quantum technology, and ultra-high frequency and ultra-wide bandgap device technology. He has won many awards, including the Bronze Medal at the Sixth TSMC Outstanding Student Researcher Award, the Best Student Paper Award at VLSI Symposium (2017 and 2021), the Best Demo Paper Award at VLSI Symposium (2022), the Best Paper Award at ICICDT (2019 and 2021), Emerging Leaders in *Journal of Physics D: Applied Physics* in 2021, and NUS Engineering Teaching Excellence Award. He is the Technical Program Chair of ICICDT (2019 and 2022) and the Sub-Committee Chair of ICICDT (2021) and EDTM (2022 and 2023) and in the technical committees of IEDM (2021 and 2022), VLSI-TSA (2022), ECS (2014, 2016, 2018, 2020, and 2022), ICMAT (2017), EDTM (2017–2021), and IWJT (2021). He is an Editor of IEEE ELECTRON DEVICE LETTERS.

Xueqing Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2007 and 2013, respectively. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. He has over 140 publications and 30 Chinese and U.S. patents. His research interests include mixed-signal circuit design, emerging memory, and memory-oriented computing exploration for AI acceleration. He received a few best paper awards (HPCA'15, ASP-DAC'17, and 2016 IEEE TMSCS), several teaching and thesis awards, and 2019 National Early-Career Award. He also served as a TPC Member in a few conferences (DAC, ICCAD, ASP-DAC, ISCAS, AICAS, GLSVLSI, ISVLSI, NVMSA, and COINS). He is an Associate Editor of IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS and IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING and a Guest Editor in a few journal issues (*International Journal of Circuit Theory and Applications* and *Flexible and Printed Electronics*). He is a Senior Member of CCF China.