Compute-in-Memory Array Design using Stacked Hybrid IGZO/Si eDRAM cells

Munhyeon Kim
Electrical and Computer Engineering
Seoul National University
Seoul, South Korea
munhyeon.kim@snu.ac.kr

Yulhwa Kim
Semiconductor Systems Engineering
Sungkyunkwan University
Suwon-Si, South Korea
yulhwakim@skku.edu

Jae-Joon Kim
Electrical and Computer Engineering
Seoul National University
Seoul, South Korea
kimjaejoon@snu.ac.kr

***Abstract*—To effectively accelerate neural networks in compute-in-memory (CIM) based systems, higher memory cell density is essential to handle the increasing computational workload and number of parameters. While CMOS embedded dynamic random access memory (eDRAM) is being explored as an alternative, improving the short retention time (t_{ret}) (<1 ms) remains a challenge that must be addressed for system applications. Recent studies show that InGaZnO (IGZO)-based eDRAM demonstrates an exceptionally long retention time t_{ret} (>100 s), but additional improvements are needed due to its substantial cell variability and slower operating speed compared to CMOS-based cells. This paper proposes a cell and array design for CIM using 3T-based stacked hybrid IGZO/Si eDRAM (Hybrid-3T) and performs a system-level deep neural network (DNN) evaluation. The Hybrid-3T cell, designed based on 7-nm FinFET technology, achieves t_{ret} that is 100 s longer compared to IGZO-based 3T eDRAM (IGZO-3T). The proposed Hybrid-3T offers a 3.4× higher bit cell density compared to 8T SRAM bit cells and a 2× higher density compared to CMOS-based 3T eDRAM (CMOS-3T), while demonstrating similar throughput and variability levels to CMOS eDRAM and SRAM-based systems. Furthermore, we evaluate DNN inference accuracy for vision and natural language processing (NLP) tasks using the proposed CIM design, examining the impact of improved cell variability and retention time on system-level characteristics. The retention time for ensuring CIM operation accuracy ($t_{ret,CIM}$) is 10^8 times longer in Hybrid-3T than CMOS-3T, and the $t_{ret,CIM}$ considering variability ($t_{ret,CIMv}$) is more than 3× longer than IGZO-3T eDRAM. As a result, the proposed Hybrid-3T eDRAM CIM leverages the advantages of both CMOS-3T and IGZO-3T CIM designs, enabling the development of high-performance, reliable systems.**

***Index Terms*—neural network accelerators, compute-in-memory, embedded DRAM, IGZO memory**

I. Introduction

Deep neural networks (DNNs) have garnered significant attention due to their widespread application across various tasks [1], [2], prompting extensive research into DNN accelerators [3]. In particular, there is a growing focus on compute-in-memory (CIM) approaches to alleviate the burden of the so-called "memory wall", often referred to as the von-Neumann bottleneck. Exploring cell structures suitable for CIM operations is one of the core aspects of CIM research. Initially, much research has been conducted based on static random access memory (SRAM). Various computing schemes ranging from 6T to 10T SRAM-based designs have been proposed [4]–[8]. However, as DNN tasks become more complex, more CIM arrays are needed for parameter storage and computation, making area-efficient CIM implementations increasingly important. While SRAM offers compatibility with CMOS technology, fast operation, and energy efficiency, it has limitations in enhancing area efficiency. Consequently, there has been a surge of research proposing embedded dynamic random access memory (eDRAM) as a suitable solution for memory operations and CIM computation [9], [10]. EDRAM cells, typically based on 2T to 4T configurations, can achieve performance levels comparable to SRAM while doubling area efficiency. Nonetheless, unlike SRAM, eDRAM introduces a disadvantage of charge loss-induced retention time (t_{ret}) due to the floating storage node (SN) during hold operations. Although various methods have been attempted to increase the t_{ret} of eDRAM, achieving compact cell structures of 2T and 3T with substantially higher t_{ret} still remains challenging.

Recently, a CIM design based on 3-dimensional (3D) vertically stacked InGaZnO (IGZO) transistor-based 3T eDRAM cells has been proposed [11], highlighting significantly longer retention time (t_{ret}) of IGZO eDRAM due to its low off-current characteristics [12], [13]. With a retention time exceeding 100 seconds, IGZO-3T eDRAM CIM has the potential to run neural network models without needing cell refresh. However, inherent properties of IGZO transistors, such as low mobility and large threshold voltage (V_{TH}) variations, present challenges in achieving high speed and accurate network performance.

To overcome these limitations, we propose a CIM architecture based on 3T stacked hybrid IGZO/Si eDRAM (Hybrid-3T) cells. This design is inspired by recently reported memory cells that combine CMOS with oxide semiconductors like IGZO [14], [15]. By utilizing IGZO transistors for the write path and silicon transistors for the read path, the Hybrid-3T achieves high performance, sufficient retention time, and high area efficiency, while offering V_{TH} variation comparable to Si CMOS technology, enabling high inference accuracy. The key contributions of this paper are as follows:

- We propose the Hybrid-3T eDRAM cell for CIM. We validate the process architecture based on industry-compatible design rules (DRs).
- We design a macro for Hybrid-3T eDRAM-based CIM operation and verify the computing operation scheme with the impact of cell variations.
- We conduct DNN evaluations for various tasks and

978-3-9826741-0-0/DATE25/© 2025 EDAA

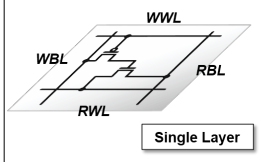
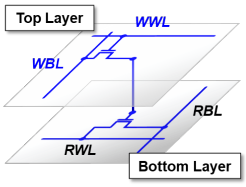
Category	CMOS-based eDRAM	IGZO-based eDRAM
2T Based Bit-Cell Configuration		
Channel Material	Si (Crystalline)	Oxide Semiconductor (e.g. IGZO, ITO)
Operation (Memory / CIM)	Memory [17], CIM [8]	Memory [13], CIM [11]
Process	Logic Compatible (CMOS Tech.)	Less Compatible (Scaled IGZO TFT Tech.)
Retention	< 1 ms for memory operation	> 100s for memory operation
Area efficiency	Conventional CMOS	Monolithic 3-D
Performance	High Mobility w/ Thin EOT	Relatively Low Mobility
Variability	Industrial process maturity	Material property limitations

Fig. 1. Comparison of characteristics of CMOS-based eDRAM and IGZO-based eDRAM.

demonstrate the superiority of Hybrid-3T. In particular, we analyze the retention time of CIM computation ($t_{ret,CIM}$) to demonstrate the applicability of Hybrid-3T for CIM.

II. Background and Motivation

A. Background: IGZO-based eDRAM

Recently, IGZO-based eDRAM with monolithic 3D technology has been actively researched to overcome the insufficient t_{ret} of eDRAM using CMOS technology. IGZO-based eDRAM demonstrates superior t_{ret} compared to CMOS-based eDRAM (see Fig. 1). The wide bandgap of IGZO effectively suppresses gate-induced drain leakage (GIDL), allowing for off-current (I_{OFF}) levels below 10^{-17} A/μm and the potential to achieve t_{ret} exceeding 100 seconds [16], [17]. The improvement of t_{ret} can be modeled using the following Eq. (1) and (2):

$$V_{SN} = V_{SN,init} \times e^{-\frac{t}{\tau}} \tag{1}$$

$$t_{ret} = ln(\frac{V_{SN,init}}{V_{SN,init} - \Delta V_{SN}}) \times \frac{V_{SN,init} \times C_{gg,Rtr}}{I_{OFF,Wtr} \times W_{Wtr}} \tag{2}$$

Eq. (1) represents the derivation of storage node voltage (V_{SN}) through the analysis of the transient response of the eDRAM cell after write (t = 0 s). Eq. (2) is the result of substituting t = 0 s and t = t_{ret} into Eq. (1), where t_{ret} is defined as the point at which V_{SN} becomes initial storage node voltage ($V_{SN,init}$) - ΔV_{SN}. Then, by separating τ into RC components, R is assigned as the value of $V_{SN,init}$ divided by $I_{OFF,Wtr} \times W_{Wtr}$, while C is assigned as $C_{gg,Rtr}$, which is the gate capacitance of the read transistor. Ultimately, it can be noted that to increase t_{ret}, low I_{OFF} and high $C_{gg,Rtr}$ are required. IGZO-based eDRAM cells can achieve a high t_{ret} of over 100 s due to low I_{OFF}.

B. Motivation: IGZO-3T based Compute-in-Memory

In CIM designs using eDRAM, 3T and 4T cells are typically used [9], [18], [19]. The reason for using additional transistors compared to the 2T gain cell commonly used in memory is to implement a computing scheme for multiply and accumulate (MAC) operations. There are two key considerations in

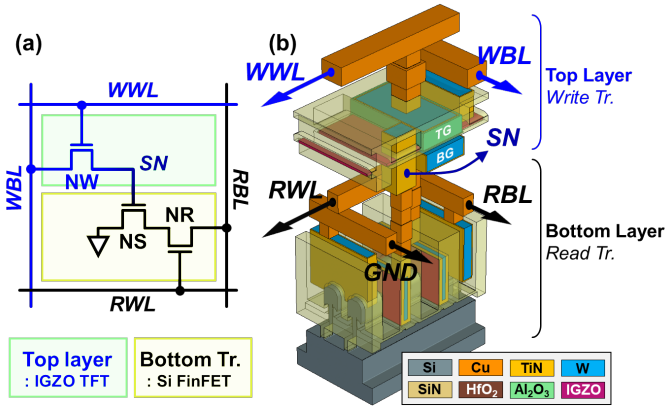


Fig. 2. (a) Schematic of a 3T-based stacked hybrid IGZO/Si eDRAM (Hybrid-3T) cell. (b) Three-dimensional (3D) structure of the Hybrid-3T cell.

eDRAM-based MAC operations: 1) t_{ret} of the cell and 2) variability. First, t_{ret} must be prioritized during MAC operations because the opportunity to reuse weights exists only as long as the data is retained in the cell. Moreover, a short refresh cycle leads to additional energy consumption due to insufficient t_{ret} [9]. Therefore, prior research has utilized IGZO-based 3T eDRAM cells (IGZO-3T) to ensure sufficient t_{ret} [11].

Second, cell variability is crucial for ensuring inference accuracy. During CIM operations, input activation occurs simultaneously on all wordlines (WLs) of the array, causing cell variability to accumulate in the partial sums. Although IGZO-3T demonstrates excellent t_{ret}, it is sensitive to process variations due to its material properties. The reported σV_{TH} of IGZO exceeds 70 mV [20], while it is less than 30 mV in 7-nm FinFET technology [21]. Therefore, there is a pressing need to develop a eDRAM cell structure that meets the requirements for long retention time and low variability as well as high-performance.

III. Proposed Hybrid-3T eDRAM CIM Cell

A. Concept of the Proposed Cell

The proposed Hybrid-3T is an eDRAM cell that efficiently arranges transistors composed of IGZO and single crystal silicon channels. As shown in Fig.2 (a), the top layer and bottom layer are connected to the storage node (SN). The top layer consists of an nMOS TFT (NW) based on IGZO channel for writing. The bottom layer is composed of an nMOS storage node transistor (NS) and read transistor (NR) based on a single crystal silicon channel. Fig. 2(b) illustrates the 3D structure of the unit cell. The unit cell consists of a total of four signal ports: write word-line (WWL), write bit-line (WBL), read word-line (RWL), read bit-line (RBL) and ground (GND) port as input/output nodes. Specifically, the NS and NR of the bottom layer are composed of fin field-effect transistor (FinFET), while the NW of the top layer is composed of a TFT with back gate (BG) and top gate (TG).

B. Cell Architecture

*a) **Cell Layouts**:* The Hybrid-3T cell is designed using a layout based on DRs corresponding to 7-nm FinFET technology [22]. The essential layers, the top layer and the bottom layer, are constructed as shown in Figs. 3(b) and (c),

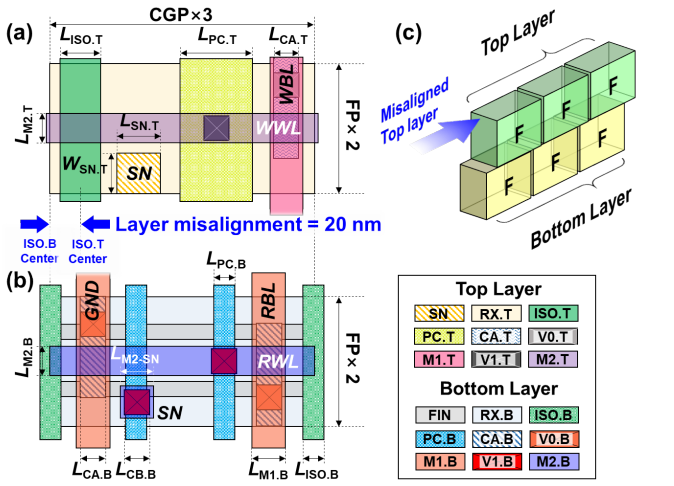


Fig. 3. Layout and layer information of the Hybrid-3T cell: (a) Bottom layer, (b) Top layer. (c) The concept of a misaligned layer.

TABLE I
LAYER INFORMATION AND DIMENSIONS OF HYBRID-3T

Bottom Layer: Read Transistor			**Top Layer: Write Transistor**		
Geometric Parameters	Unit	Value	Geometric Parameters	Unit	Value
Contact gate pitch (CGP)	nm	54	Storage node length ($L_{SN.T}$)	nm	30
Fin pitch (FP)	nm	32	Storage node width ($W_{SN.T}$)	nm	30
PC drawn length ($L_{PC.B}$)	nm	7	PC drawn length ($L_{PC.T}$)	nm	45
Physical gate length ($L_{G.B}$)	nm	20	Physical gate length ($L_{G.T}$)	nm	45
Contact CD ($L_{CA.B}$)	nm	20	Contact CD ($L_{CA.T}$)	nm	20
Contact width ($W_{CA.B}$)	nm	60	Contact width ($W_{CA.T}$)	nm	60
Isolation CD ($L_{ISO.B}$)	nm	20	Isolation CD ($L_{ISO.T}$)	nm	30
Via-0/1 length ($L_{V0/1.B}$)	nm	20	Via-0/1 length ($L_{V0/1.T}$)	nm	20
Metal-1/2 length ($L_{M1/2.B}$)	nm	22	Metal-1/2 length ($L_{M1/2.T}$)	nm	22

respectively. Detailed dimensions can be found in Table I. The two FinFET transistors located in the bottom layer (NS and NR) determine the cell size. Since the two bottom transistors share the source/drain (S/D), a CGP $\times$ 3 space is required to implement the structure with minimal dimensions. The unit cell layout includes isolation (ISO.B) positioned at the left and right ends. Adopting a 2-Fin structure secures large SN capacitance, maximizes cell performance, and reduces variation, resulting in a cell height (H_{cell}) of FP $\times$ 2. Implementing the layout of the bottom layer requires essential advanced process techniques such as contact over active gate (COAG) [23].

There are three distinctive features of the top layer: 1) SN aligns precisely with the NS gate of the bottom layer, 2) The gate length of the top layer transistor ($L_{G.T}$) is longer than the length of the bottom layer transistor ($L_{G.B}$), and 3) The isolation (ISO.T) of the top layer is located in the middle of the cell.

*b) **Misaligned Layer**:* To increase the t_{ret} of the Hybrid-3T cell without introducing additional materials or advanced technology, an efficient strategy is maximizing the gate length $L_{G.T}$ of the NW to minimize I_{OFF}. To increase $L_{G.T}$, a misalignment between the top and bottom layers is necessary, as shown in Fig. 3(a). This requirement arises because the position of the SN in the top layer is determined by the layout of the bottom layer. In other words, the position of the SN in the top layer must align with the x-coordinate of the NS gate. While the bottom layer of the Hybrid-3T cell includes two transistors, NS and NR, the top layer contains only one NW transistor. By configuring the top layer layout symmetrically as in the bottom layer, $L_{G.T}$ becomes 25 nm due to the SN position. It is clear that the x-direction centers of ISO.B and the isolation of the top layer (ISO.T) are misaligned by 20 nm. By intentionally shifting the center of the top layer by 20 nm to create an intentional top-bottom layer misalignment, $L_{G.T}$ can be increased from 25 to 45 nm, resulting in a significant reduction in I_{OFF} from 8×10^{-14} to 2×10^{-18} $A/\mu m$.

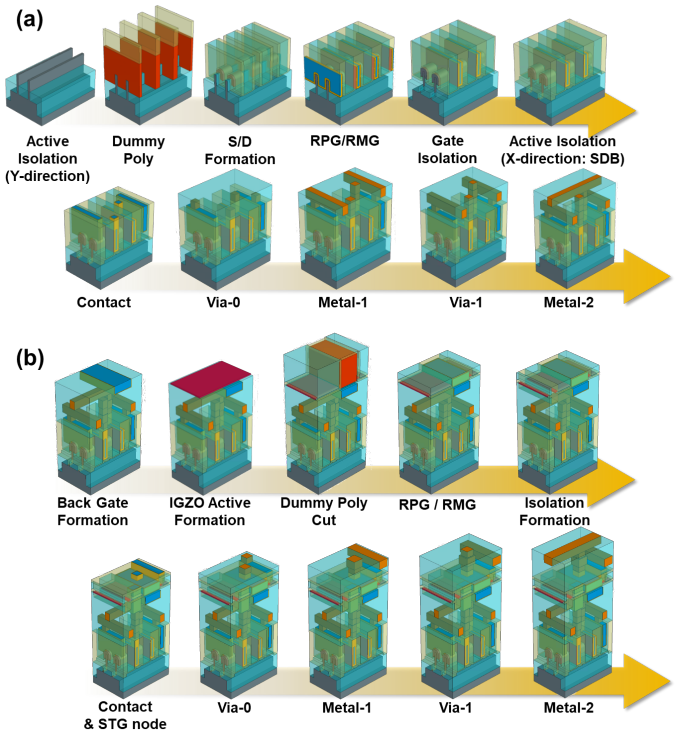


Fig. 4. Process flow for the fabrication of Hybrid-3T: (a) Bottom layer, (b) Top layer.

*c) **Process Flow**:* The unique structure of the Hybrid-3T cell can be implemented using CMOS technology adopted in the industry. To rigorously analyze manufacturability, a structure based on 7-nm technology is implemented using the commercial 3D technology-aided desig (TCAD) tool, $SentaurusProcess^{TM}$ (Figs. 4(a) and (b)).

The Hybrid-3T process flow has two major advantages. First, it enables monolithic 3D processing. Thanks to the stacking of layers through the monolithic 3D process, the bottom and top layers have an excellent alignment margin compared to wafer bonding. The misalignment overlay level of wafer bonding is greater than 50 nm [24], while the overlay in CMOS technology is less than 4 nm [25]. Second, it facilitates integration with advanced processes due to the adoption of existing CMOS and IGZO TFT processes. The single diffusion break (SDB) shown in Fig. 4(a) and the back gate formation shown in Fig. 4(b) can be utilized for cell process optimization while minimizing additional process development.

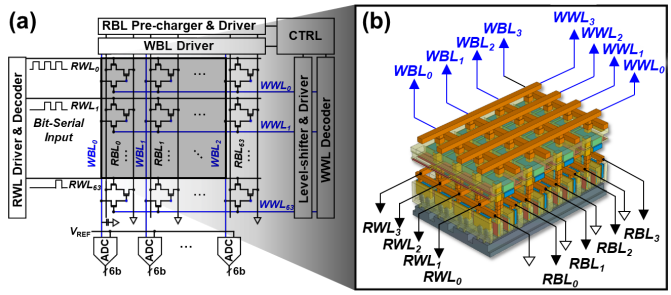


Fig. 5. (a) Schematic of Hybrid-3T based compute-in-memory (CIM) macro and (b) 3D bird's eye view of the 4×4 array of Hybrid-3T.

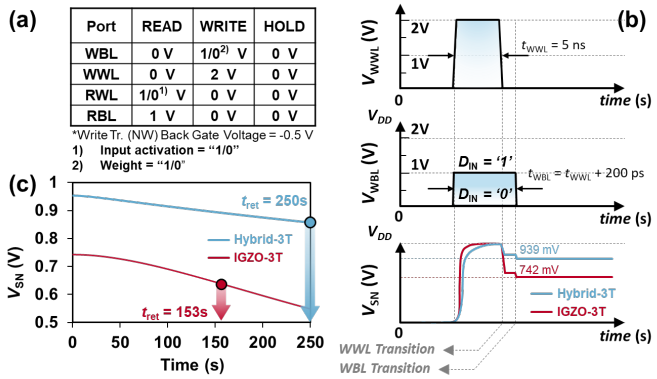


Fig. 6. (a) Operating conditions of Hybrid-3T. (b) Waveforms for the write operation of Hybrid-3T and 3T IGZO-based eDRAM (IGZO-3T) cells. (c) Initial storage node voltage ($V_{SN,init}$) and retention time (t_{ret}).

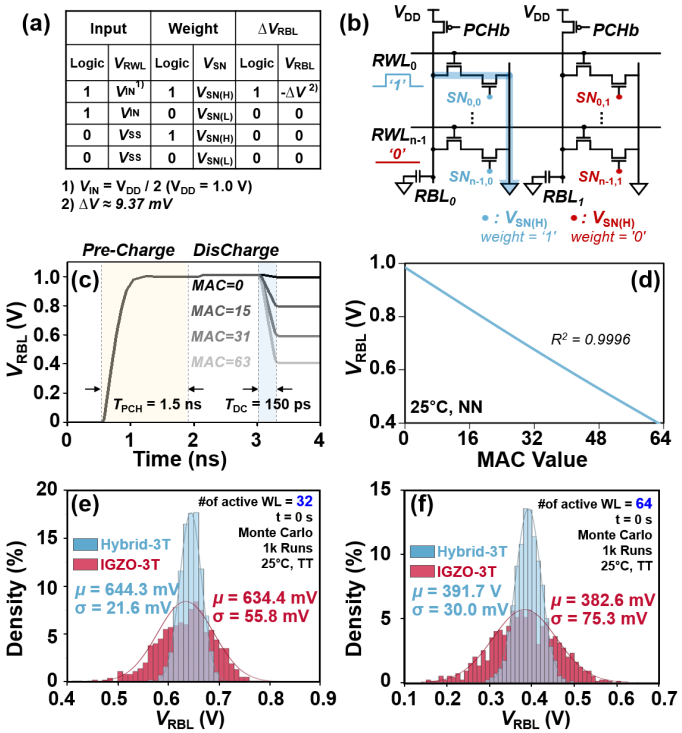


Fig. 7. (a) Truth table and voltage conditions for CIM operation. (b) Computing operations of the array. (c) Timing diagram of RBL voltage (V_{RBL}) for each multiply and accumulation (MAC) values. (d) Linearity of MAC operations.A 1k Monte-Carlo (MC) simulation of the V_{RBL} for both (a) 32 and (b) 64 active word-lines (WLs).

C. Compute-in-memory Macro with Hybrid-3T

The basic configuration and operation of the Hybrid-3T based CIM macro are shown in Fig. 5(a). The configuration includes a Hybrid-3T array, column/row drivers, a decoder, and a 6-bit resolution analog to digital converters (ADCs). A distinctive feature of the Hybrid-3T as an eDRAM is the separation of the RWL and WWL drivers. Among these components, a level shifter is required for the write driver to drive the monolithic 3D structure of the Hybrid-3T cell. This necessity arises because the NW, based on an IGZO transistor, requires a higher level of V_{WWL} (2 V) compared to other digital blocks that operate at a lower supply voltage ($V_{DD} = 1$ V). Particularly, Fig. 5(b) shows a 3D perspective view of the 4×4 array within the 64×64 cell array of Hybrid-3T in the CIM macro.

D. Basic Operation of Hybrid-3T Cell

a) ***Operation Conditions:*** The basic operating scheme of the Hybrid-3T cell is based on the operation of conventional eDRAM cells [26], [27]. The voltage conditions for each port of read, write, and hold operations from the CIM perspective are summarized in the table in Fig. 6(a). To perform a read operation, the RBL voltage (V_{RBL}) must be fixed at 1 V, and the RWL should be set to either 1 or 0 V. In CIM operations, the value applied to the RWL serves as the input activation value. During a write operation, both the RWL and RBL of the bottom layer are fixed at 0 V, with control exerted only by WWL and WBL. In CIM operations, the data stored in the cell represents the weight. Writing "1" maintains the V_{WBL} at 1 V while enabling the WWL, whereas writing "0" keeps the WBL voltage (V_{WBL}) at 0 V. During a hold operation, the WWL voltage (V_{WWL}) is set to 0 V, with V_{WBL}, V_{RWL}, and V_{RBL} all maintained at 0 V. To adjust the V_{TH} of IGZO-based transistors to a level similar to CMOS, -0.5 V is applied to the BG.

b) ***Initial Storage Node Voltage:*** Fig. 6(b) shows the write operation signals and the results of $V_{SN,init}$ generation, rigorously simulated using TCAD. The Hybrid-3T achieved 939 mV, which is about 200 mV higher than the 742 mV in the IGZO-3T. This improvement is due to the enhanced gate capacitance of the NS transistor ($C_{gg,NS}$) achieved through the 7-nm FinFET-based process. As a result, $V_{SN,init}$ increased, improving t_{ret} (Eq. (2)). However, the increase in $C_{gg,NS}$ leads to a longer write time. With a target voltage of 0.9 V (90% of 1 V V_{WBL}), IGZO-3T reaches this level within 300 ps, while Hybrid-3T takes 1.3 ns, resulting in a 1 ns difference. Nevertheless, as shown in Fig. 6(c), Hybrid-3T has a t_{ret} approximately 100 s longer than IGZO-3T.

IV. Evaluation Results

A. Computing Scheme Verification

a) ***Current-based Computing:*** The Hybrid-3T based CIM macro performs AND-based MAC operations. Fig. 7(a) shows the Boolean logic truth table and the voltage conditions corresponding to each port. The partial sum is represented by changes in V_{RBL}, with voltage changes occurring solely under the condition that both the weight and input of each cell are "1", resulting in discharge by ΔV. This truth table is visualized

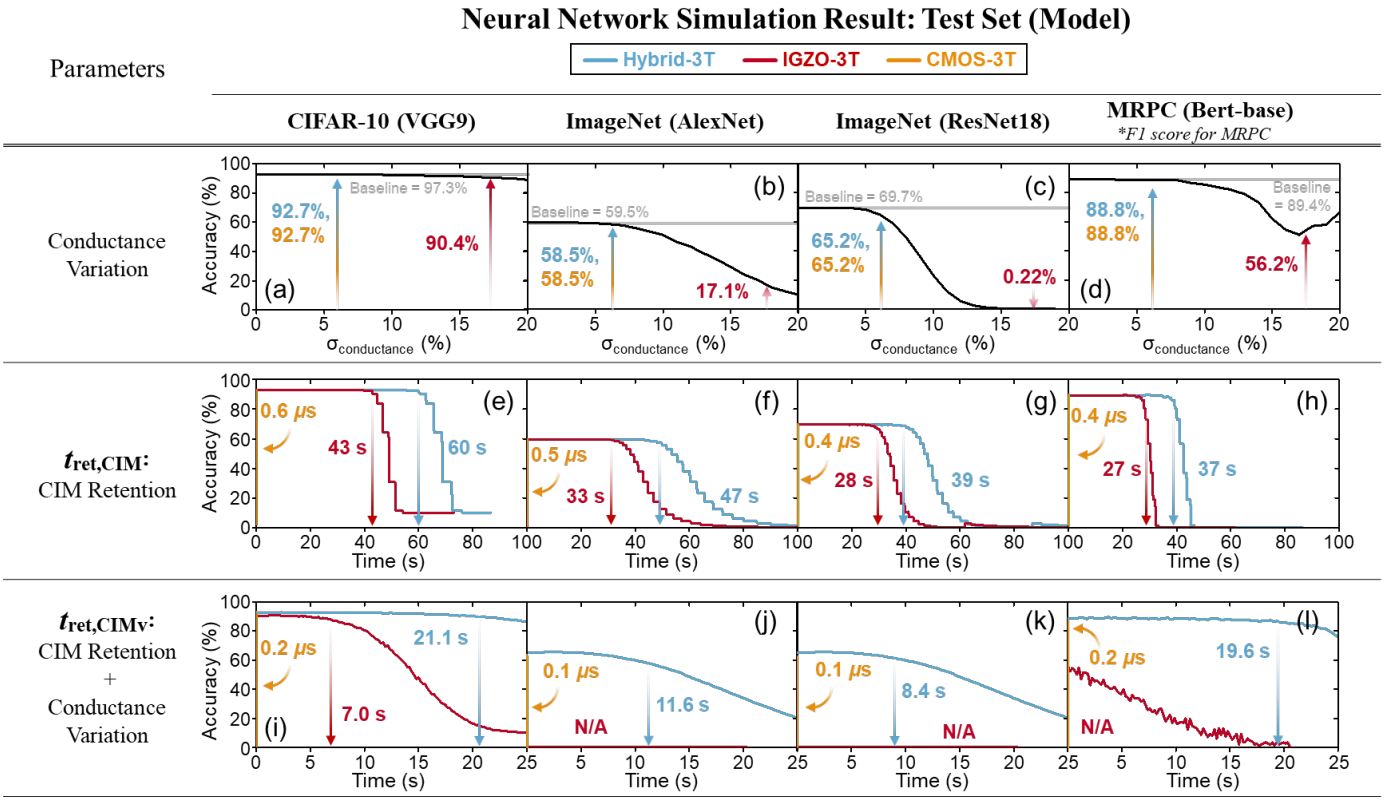


Fig. 8. DNN simulation results with CIM operation for Hybrid-3T, IGZO-3T and CMOS-3T. Impact of conductance variation ($\sigma_{\text{conductance}}$) on CIM inference accuracy for CIFAR-10: (a) VGG-9, ImageNet: (b) AlexNet and (c) ResNet-18 and MRPC: (d) BERT-base. CIM retention $t_{\text{ret,CIM}}$ for CIFAR-10: (e) VGG-9, ImageNet: (f) AlexNet and (g) ResNet-18 and MRPC: (h) BERT-base. The evaluation results of $t_{\text{ret,CIMv}}$, which comprehensively consider $\sigma_{\text{conductance}}$ and $t_{\text{ret,CIM}}$ for CIFAR-10: (i) VGG-9, ImageNet: (j) AlexNet and (k) ResNet-18 and MRPC: (l) BERT-base.

in Fig. 7(b). Each RBL includes one pre-charger and one capacitor, and the MAC operation occurs in two phases: precharge and discharge, allowing current-based accumulation. Fig. 7(c) shows the timing diagram for the RBL voltage V_{RBL}. The cycle time of the Hybrid-3T based CIM macro is 4.5 ns. The minimum MAC result "0" results in no discharge, maintaining the precharged V_{RBL} at V_{DD}. Conversely, the maximum MAC value "63" triggers maximum discharge of V_{RBL}, reducing the voltage to 0.4 V instead of V_{SS} to maintain the linearity of the MAC values. Consequently, a high correlation between MAC values and V_{RBL} (R^2 = 0.9996) is achieved (Fig. 7(d)).

*b) **Impact of Cell Variation on Partial Sum**:* In the proposed macro, currents from cells in the multiple rows are summed together at a bitline. Consequently, reducing the level of cell variation becomes more critical in CIM operations compared to memory operations. After applying cell conductance variation, Monte Carlo (MC) simulations are performed to assess the V_{BL} distribution following CIM operations. The results of the MC simulations, conducted 1,000 times under the typical-typical (TT) corner and at $25°C$, are shown in Fig. 7(e) and (f), illustrating scenarios with 32 and 64 activated WLs. The worst-case scenario for V_{BL} variation occurs in the situation where all weights are set to "1" and all WLs are activated to "1". In this scenario, the standard deviation value (σ) of V_{RBL} for Hybrid-3T is 30 mV, while for IGZO-3T, it reaches 75 mV, indicating a 2.5× difference. Since Hybrid-3T uses Si MOSFETs for read transistors, it exhibits the same level of cell variation as that of CMOS-3T, which is much lower than IGZO-3T, where IGZO transistors are used for the read transistors. As a result, considering the limited dynamic range of ADCs, the Hybrid-3T offers significant advantages in CIM operations over IGZO-3T due to its lower variability.

B. Deep Neural Network Simulation

A comprehensive analysis of DNN simulations utilizing CIM with the Hybrid-3T and IGZO-3T CIM designs is conducted. For vision-based DNNs, the classification accuracy of VGG-9 [28] and ResNet-18 [29] on the CIFAR-10 dataset [30], as well as AlexNet [31] and ResNet-18 on the ImageNet dataset [32], is analyzed. These vision-based DNNs use 4-bit weights and activations, with CIM processing applied to all convolutional layers and fully-connected (FC) layers, excluding the initial and final layers. Additionally, for NLP-focused DNNs, the F1 scores of BERT-base models [33] using the Microsoft Research Paraphrase Corpus (MRPC) dataset [34] are evaluated. These NLP-oriented DNNs utilize 8-bit weights and activations, and CIM processing is applied to the output FC layer of each transformer block.

*a) **Impact of Cell Variation on DNN Inference**:* Cell conductance variation ($\sigma_{\text{conductance}}$) significantly impacts computational accuracy in DNN operations using CIM, with effects varying across different networks. The $\sigma_{\text{conductance}}$ derived from integrating $\sigma_{V_{TH}}$ and cell architecture results in 6% for Hybrid-

3T/CMOS-based eDRAM CIM and 17.5% for IGZO-3T CIM. For simpler vision tasks such as CIFAR-10 inference (Fig. 8(a)), both Hybrid-3T/CMOS-based 3T eDRAM (CMOS-3T) and IGZO-3T CIM show relatively good accuracy, with Hybrid-3T/CMOS-3T maintaining approximately 2% higher accuracy than IGZO-3T CIM. However, in the context of more complex tasks like inference on the ImageNet dataset, the difference between Hybrid-3T/CMOS-3T and IGZO-3T CIM becomes more pronounced, exceeding 40% (Figs. 8(b) and (c)). Additionally, NLP task inference results (Figs. 12(c) and (d)) show that Hybrid-3T maintains an accuracy difference within 1% compared to digital operations, whereas IGZO-3T CIM experiences a decline of over 30%. In conclusion, while $\sigma_{\text{conductance}}$ does not significantly affect simpler tasks, Hybrid-3T/CMOS-3T CIM demonstrates much higher inference accuracy than IGZO-3T CIM in more complex tasks.

*b) **Impact of Cell Retention Time on DNN Inference**:* In the critical process of evaluating eDRAM-based CIM operations, considering the impact of data retention time is essential. This is because CIM accuracy is sensitive to the analog data levels within the cell over time. Therefore, we define $t_{\text{ret,CIM}}$ as the duration for maintaining DNN inference accuracy, indicating the point at which accuracy decreases by 3% from its accuracy at $t = 0$ s. For CIFAR-10 inference, both Hybrid-3T and IGZO-3T CIM achieve $t_{\text{ret,CIM}}$ of over 40 s, whereas CMOS-3T shows a much shorter $t_{\text{ret,CIM}}$ of 0.6 μs (Fig. 8(e)). In inference results on the ImageNet and MRPC datasets with AlexNet/ResNet-18 and BERT-base models, Hybrid-3T achieves $t_{\text{ret,CIM}}$ between 37 and 47 s, while IGZO-3T CIM shows a decline to around 10 s. CMOS-3T, however, is limited to a maximum of 0.5 μs (Figs. 8(f)-(h)). Furthermore, to comprehensively evaluate the performance of Hybrid-3T and IGZO-3T CIM macros, we define and assess a combined metric, $t_{\text{ret,CIMv}}$, which includes both $\sigma_{\text{conductance}}$ and $t_{\text{ret,CIM}}$ (with results noted as N/A should the initial accuracy drop more than 30% from baseline accuracy). For IGZO-3T CIM, $t_{\text{ret,CIMv}}$ shows only 7 s for the simpler CIFAR-10 dataset and is noted as N/A for ImageNet and MRPC. CMOS-based eDRAM CIM shows a maximum $t_{\text{ret,CIMv}}$ of 0.2 μs across all datasets. In conclusion, only Hybrid-3T achieves a meaningful $t_{\text{ret,CIMv}}$ even for complex datasets, with high levels of 11.6 and 19.6 s observed for ImageNet and BERT-base, respectively.

C. Benchmark and Comparisons

The CIM macro and DNN performance of the Hybrid-3T is compared to CMOS-3T, IGZO-3T, and 8T SRAM based on 7-nm technology DRs (Table II). One of the most notable advantages of Hybrid-3T over 8T SRAM and CMOS-3T is its higher bit-cell density; 3.4× and 2× higher than 8T SRAM and CMOS-3T, respectively. The comparison with 8T SRAM should take into account the comparison between push rule and logic rule, and the benefits are expected to be greater when compared on the same basis. For a fair comparison at a similar energy efficiency level, the same computing scheme is applied to each CIM. Simulation results show that Hybrid-3T CIM achieves similar throughput to that of 8T SRAM and CMOS-3T CIM, and 14x higher throughput than IGZO-3T CIM. Furthermore, in hardware-based DNN performance evaluations, Hybrid-3T demonstrates longer $t_{\text{ret,CIM}}$ and $t_{\text{ret,CIMv}}$ than CMOS-3T and IGZO-3T CIM.

TABLE II
FEATURE SUMMARY AND COMPARISON TO PREVIOUS WORK

Parameters		8T SRAM [8]	CMOS-3T [10]	IGZO-3T [11]	Hybrid-3T (This Work)
Physical Design*	Technology	⟷ 7-nm FinFET node ⟷			
	Array Size	⟷ 64 × 64 ⟷			
	Cell Structure	8T SRAM ⟷ 3T eDRAM ⟷			
	Push Rule	Yes ⟷ No ⟷			
	Bit-cell Area (μm^2)	0.053	0.026** ⟷		0.016 ⟷
Operation Conditions	Computing scheme	⟷ Current-mode Accumulation*** ⟷			
	Power Supply (V)	1	1	1.3	1
	Cycle time (ns)	4.5	4.5	65	4.5
Throughput* (GOPS)**	OPs (a/w:4b/4b)	113.7	97.3	7.9	113.7
	bOPs (a/w:1b/1b)	1819	1557	127	1819
Energy Efficiency (TOPS/W)	OPs (a/w:4b/4b)	87.8	87.1	79.5	88.6
	bOPs (a/w:1b/1b)	1404	1393	1271	1417
$t_{\text{ret,CIM}}$	CIFAR-10 (VGG9)	-	0.6 μs	43 s	60 s
	ImageNet (AlexNet)	-	0.5 μs	33 s	47 s
	ImageNet (ResNet18)	-	0.4 μs	28 s	39 s
	MRPC (Bert-base)	-	0.4 μs	27 s	37 s
$t_{\text{ret,CIMv}}$	CIFAR-10 (VGG9)	-	0.2 μs	7 s	21.1 s
	ImageNet (AlexNet)	-	0.1 μs	|	11.6 s
	ImageNet (ResNet18)	-	0.1 μs	N/A	8.4 s
	MRPC (Bert-base)	-	0.2 μs	|	19.6 s

*Based on [8], 7-nm technology / array size = 64 × 64 for a fair comparison.

**Reproduced based on 7-nm logic rules (for CIM & DNN evaluations)

***Computing scheme is restructured using current-mode accumulation for benchmark.

V. CONCLUSION

In this paper, we propose a Hybrid-3T-based array design for CIM operations. The design based on DRs with 7-nm technology is thoroughly validated through 3D TCAD and SPICE simulations. Our proposed cell achieves a 3.4× and 2× increase in cell density compared to CMOS-based 8T SRAM bit cells and CMOS-3T eDRAM cells for CIM, respectively, while maintaining comparable throughput—significantly higher than that of IGZO-3T CIM. Additionally, the Hybrid-3T-based CIM system demonstrates an 8 orders-of-magnitude improvement in $t_{\text{ret,CIM}}$ compared to CMOS-3T CIM. In summary, Hybrid-3T CIM designs combine the extended retention time of IGZO-3T with the low variation and high performance of CMOS-3T, offering the advantages of both technologies. This makes Hybrid-3T eDRAM CIM a compelling alternative to SRAM-based CIM.

ACKNOWLEDGEMENT

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00208606, NeuroHub+: Scheduler and Simulator for General In-Memory Neural Network Accelerators, No. 2022-0-00266, Development of Ultra-Low Power Low-Bit Precision Mixed-mode SRAM PIM, IITP-2023-RS-2023-00256081: artificial intelligence semiconductor support program to nurture the best talents), BK21 FOUR program at Seoul National University, and ISRC at Seoul National University. The EDA tool was supported by the IC Design Education Center(IDEC). (Corresponding Author: Jae-Joon Kim).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[4] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing cnn accelerator employing charge-domain compute," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, 2019.

[5] A. Biswas and A. P. Chandrakasan, "Conv-ram: An energy-efficient sram with embedded convolution computation for low-power cnn-based machine learning applications," in *2018 IEEE International Solid-State Circuits Conference - (ISSCC)*, pp. 488–490, 2018.

[6] J. Kim, J. Koo, T. Kim, Y. Kim, H. Kim, S. Yoo, and J.-J. Kim, "Area-efficient and variation-tolerant in-memory bnn computing using 6t sram array," in *2019 Symposium on VLSI Circuits*, pp. C118–C119, 2019.

[7] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, "C3sram: An in-memory-computing sram macro based on robust capacitive coupling computing mechanism," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, 2020.

[8] Q. Dong, M. E. Sinangil, B. Erbagci, D. Sun, W.-S. Khwa, H.-J. Liao, Y. Wang, and J. Chang, "15.3 a 351tops/w and 372.4gops compute-in-memory sram macro in 7nm finfet cmos for machine-learning applications," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, pp. 242–244, 2020.

[9] C. Yu, T. Yoo, H. Kim, T. T.-H. Kim, K. C. T. Chuan, and B. Kim, "A logic-compatible edram compute-in-memory with embedded adcs for processing neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 2, pp. 667–679, 2021.

[10] Z. Chen, X. Chen, and J. Gu, "15.3 a 65nm 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 *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, pp. 240–242, 2021.

[11] 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 Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 8, no. 1, pp. 35–43, 2022.

[12] K. Huang, X. Duan, J. Feng, Y. Sun, C. Lu, C. Chen, G. Jiao, X. Lin, J. Shao, S. Yin, J. Sheng, Z. Wang, W. Zhang, X. Chuai, J. Niu, W. Wang, Y. Wu, W. Jing, Z. Wang, J. Xu, G. Yang, D. Geng, L. Li, and M. Liu, "Vertical channel-all-around (caa) igzo fet under 50 nm cd with high read current of 32.8 a/m (vth + 1 v), well-performed thermal stability up to 120 °C for low latency, high-density 2t0c 3d dram application," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pp. 296–297, 2022.

[13] C. Chen, J. Xiang, X. Duan, C. Lu, J. Niu, K. Zhang, Y. Liu, N. Lu, Z. Jiao, Y. Shen, Q. Luan, G. Wang, C. Zhao, G. Yang, D. Geng, L. Li, and M. Liu, "First demonstration of stacked 2t0c-dram bit-cell constructed by two-layers of vertical channel-all-around igzo fets realizing 4f2 area cost," in *2023 International Electron Devices Meeting (IEDM)*, pp. 1–4, 2023.

[14] S. Liu, S. Qin, K. Jana, J. Chen, K. Toprasertpong, and H.-S. P. Wong, "First experimental demonstration of hybrid gain cell memory with si pmos and ito fet for high-speed on-chip memory," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pp. 1–2, 2024.

[15] M. Kim and J.-J. Kim, "4-transistor ternary content addressable memory cell design using stacked hybrid igzo/si transistors," in *2024 60th ACM/IEEE Design Automation Conference (DAC)*, 2024.

[16] A. Belmonte, H. Oh, S. Subhechha, N. Rassoul, H. Hody, H. Dekkers, R. Delhougne, L. Ricotti, K. Banerjee, A. Chasin, M. J. van Setten, H. Puliyalil, M. Pak, L. Teugels, D. Tsvetanova, K. Vandersmissen, S. Kundu, J. Heijlen, D. Batuk, J. Geypen, L. Goux, and G. S. Kar, "Tailoring igzo-tft architecture for capacitorless dram, demonstrating $>10^3$s retention, $>10^{11}$ cycles endurance and lg scalability down to 14nm," in *2021 IEEE International Electron Devices Meeting (IEDM)*, pp. 10.6.1–10.6.4, 2021.

[17] M. Oota, Y. Ando, K. Tsuda, T. Koshida, S. Oshita, A. Suzuki, K. Fukushima, S. Nagatsuka, T. Onuki, R. Hodo, T. Ikeda, and S. Yamazaki, "3d-stacked caac-in-ga-zn oxide fets with gate length of 72nm," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 3.2.1–3.2.4, 2019.

[18] I. Lee, E. Kim, N. Kang, H. Oh, and J.-J. Kim, "In-memory neural network accelerator based on edram cell with enhanced retention time," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2023.

[19] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing cnn accelerator employing charge-domain compute," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, 2019.

[20] H. Kunitake, K. Ohshima, K. Tsuda, N. Matsumoto, T. Koshida, S. Ohshita, H. Sawai, Y. Yanagisawa, S. Saga, R. Arasawa, T. Seki, R. Honda, H. Baba, D. Shimada, H. Kimura, R. Tokumaru, T. Atsumi, K. Kato, and S. Yamazaki, "A c -axis-aligned crystalline in-ga-zn oxide fet with a gate length of 21 nm suitable for memory applications," *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 495–502, 2019.

[21] M. S. Bhoir *et al.*, "Variability sources in nanoscale bulk finfets and titan-a promising low variability wfm for 7/5nm cmos nodes," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 36.2.1–36.2.4, 2019.

[22] "International roadmap for devices and systems ($IRDS^{TM}$) 2018 edition.," 2018. Accessed: Oct. 2019. [Online]. Available: https://irds.ieee.

[23] K. Cheng, C. Park, H. Wu, J. Li, S. Nguyen, J. Zhang, M. Wang, S. Mehta, Z. Liu, R. Conti, *et al.*, "Improved air spacer co-integrated with self-aligned contact (sac) and contact over active gate (coag) for highly scaled cmos technology," in *2020 IEEE Symposium on VLSI Technology*, pp. 1–2, IEEE, 2020.

[24] S.-A. Chew, B. Zhang, K. Vanstreels, E. Chery, J. De Messemaeker, L. Witters, K. Van Sever, S. Iacovo, S. Dewilde, M. Stucchi, J. De Vos, G. Beyer, A. Miller, and E. Beyne, "The challenges and solutions of cu/sicn wafer-to-wafer hybrid bonding scaling down to 400nm pitch," in *2023 International Electron Devices Meeting (IEDM)*, pp. 1–4, 2023.

[25] J. Mulkens, M. Hanna, H. Wei, V. Vaenkatesan, H. Megens, and D. Slotboom, "Overlay and edge placement control strategies for the 7nm node using EUV and ArF lithography," in *Extreme Ultraviolet (EUV) Lithography VI* (O. R. W. II and E. M. Panning, eds.), vol. 9422, p. 94221Q, International Society for Optics and Photonics, SPIE, 2015.

[26] K. C. Chun, P. Jain, T.-H. Kim, and C. H. Kim, "A 667 mhz logic-compatible embedded dram featuring an asymmetric 2t gain cell for high speed on-die caches," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 2, pp. 547–559, 2012.

[27] K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "A 3t gain cell embedded dram utilizing preferential boosting for high density and low power on-die caches," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, 2011.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR'15)*, 2015.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'16)*, pp. 770–778, 2016.

[30] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," 2009.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'12)*, 2012.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[34] B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third international workshop on paraphrasing (IWP2005)*, 2005.