

# 3D Integration of Hybrid IGZO/Si and IGZO eDRAMs for High-Density/High-Performance On-Chip Memory

Munhyeon Kim<sup>1</sup>, Sukhyun Choi<sup>2</sup>, Yulhwa Kim<sup>3</sup>, and Jae-Joon Kim<sup>2</sup>

<sup>1</sup>Department of Electrical and Information Engineering, Seoul National University of Science and Technology

<sup>2</sup>Department of Electrical and Computer Engineering, Seoul National University

<sup>3</sup>Department of Semiconductor Systems Engineering, Sungkyunkwan University

Emails: <sup>1</sup>munhyeon.kim@seoultech.ac.kr, <sup>2</sup>sukhyunchoi@snu.ac.kr, <sup>3</sup>yulhwakim@skku.edu, <sup>2</sup>kimjaejuon@snu.ac.kr

**Abstract**—The growing need for advanced memory architectures leveraging 3D integration has become increasingly critical in modern computing systems. In particular, memory architectures that match the performance of static random access memory (SRAM) while significantly increasing density are highly impactful. In this paper, we propose a 3D integration-based hybrid InGaZnO(IGZO)/Si embedded dynamic random access memory architecture (Hybrid-3D) and circuit design, which markedly increases on-chip memory density and enhances system performance. The superiority of Hybrid-3D is demonstrated through rigorous validation involving process integration verification, transistor-level modeling, and circuit-level memory design. Detailed evaluations of the vertically stacked memory operation confirm stable operations, enabling a  $22\times$  increase in on-chip memory density compared to SRAM. Integrating Hybrid-3D on-chip memory into neural processing unit (NPU) architectures results in substantial improvements in energy efficiency and processing speed. System-level evaluations across vision and natural language processing (NLP) tasks reveal a maximum energy efficiency improvement of  $3.2\times$  and a throughput increase of  $2.6\times$ .

**Index Terms**—IGZO memory, embedded DRAM, monolithic 3D integration, neural network accelerators, on-chip memory

## I. INTRODUCTION

The neural processing unit (NPU) accelerates deep learning computations and have been advancing in tandem with deep learning technology [1]–[3]. With the growing complexity of deep learning models, the demand to efficiently process these large models within an NPU has made enhancing the density and performance of on-chip memory a top hardware priority [4]–[6].

However, the scaling of CMOS technology has been slowing down. In particular, scaling of static random access memory (SRAM) has become particularly challenging compared to logic scaling [7], making implementation of large on-chip memory more difficult. To achieve higher integration density for on-chip memory, 2T embedded dynamic random access memory (eDRAM) has garnered attention for its higher density than 6T SRAM [8], [9] as shown in Fig. 1(a). However, 2T eDRAM based on Si channels faces a limitation of insufficient retention time ( $t_{\text{ret}}$ ) due to data loss caused by high subthreshold leakage. To ensure sufficient  $t_{\text{ret}}$ , IGZO-based eDRAM (IGZO-2T), which exhibits low-leakage characteristics, has recently gained attention [10]–[13] as shown in Fig. 1(b). Furthermore, unlike Si-based devices, IGZO-2T allows for 3D stacking, which enables even higher cell density. Despite such advantages, IGZO-2T suffers from low operating speed due to small driving current caused by the low carrier mobility, failing to meet on-chip memory performance requirements.

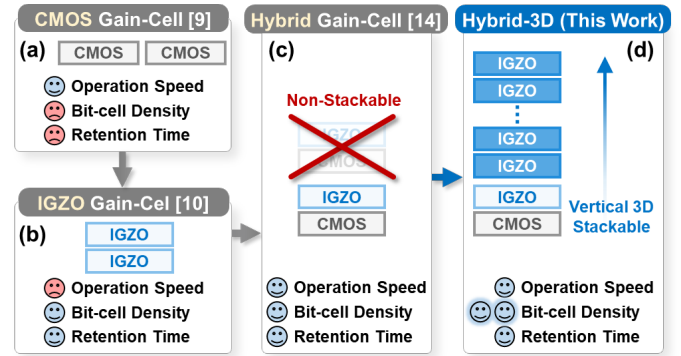


Fig. 1. (a) A CMOS-based gain cell [9], an IGZO-based 3D-stacked gain cell, a gain cell combining IGZO and Si, and a 3D-stacked IGZO/Si eDRAM-based hybrid memory architecture (Hybrid-3D).

Recently, as shown in Fig. 1(c), a hybrid Si/IGZO eDRAM (Hybrid-2T) has been introduced as an alternative for on-chip memory [14], [15], combining the long retention characteristics of IGZO-2T and the high performance of Si eDRAM. By using IGZO for the write transistor channel and Si for the read transistor channel, Hybrid-2T achieves both long retention and high read performance. However, Hybrid-2T faces limitations in benefiting from further cell density improvements through multi-layer monolithic 3D integration because the Si read transistors are composed of transistors with a single-crystal silicon lattice.

To overcome this limitation, we propose a 3D stacked IGZO/Si eDRAM based hybrid memory architecture (Hybrid-3D), in which a Hybrid-2T is fabricated on the bottom layer and IGZO 2T devices are stacked above, as shown in Fig. 1(d). Hybrid-3D allows monolithic 3D integration because Si transistors are placed at the bottom layer and memory cells at the upper layers use IGZO transistors only. From the perspective of 3D memory architecture, the bottom silicon-based layer responsible for read operations in the Hybrid-2T memory, located at the lowest tier, is co-fabricated on the same layer as the silicon channel-based transistors that constitute the peripheral circuits. Along with a single IGZO layer vertically stacked above, the memory array forms the Hybrid-2T structure. Additionally stacked IGZO layers are used to vertically implement the read and write transistors of the IGZO-2T, respectively. In conclusion, this approach results in an on-chip memory structure that simultaneously achieves both high density and performance.

The key contributions are summarized as follows:

TABLE I  
LAYER INFORMATIONS OF THE HYBRID-2T/IGZO-2T

Geometric Parameters	Bottom Layer: Read Transistor		Top Layer: Write Transistor (Hybrid-2T/IGZO-2T)	
	Hybrid-2T (nm)	IGZO-2T (nm)	Geometric Parameters	Value (nm)
Contact gate pitch (CGP)	120	240	Storage node width ( $W_{SN,T}$ )	40
Active width ( $W_{RX,B}$ )	180	220	Active width ( $W_{RX,T}$ )	200
PC length ( $L_{PC,B}$ )	30	40	PC length ( $L_{PC,T}$ )	40
CNT CD/width ( $L_{CA,B}/W_{CA,B}$ )	40/60	60	CNT CD/width ( $L_{CA,T}/W_{CA,T}$ )	60/70
Isolation CD ( $L_{ISO,B}$ )	30	40	Isolation CD ( $L_{ISO,T}$ )	30
Metal-1/2 length ( $L_{M1/2,B}$ )	50	50	Metal-1/2 length ( $L_{M1/2,T}$ )	50

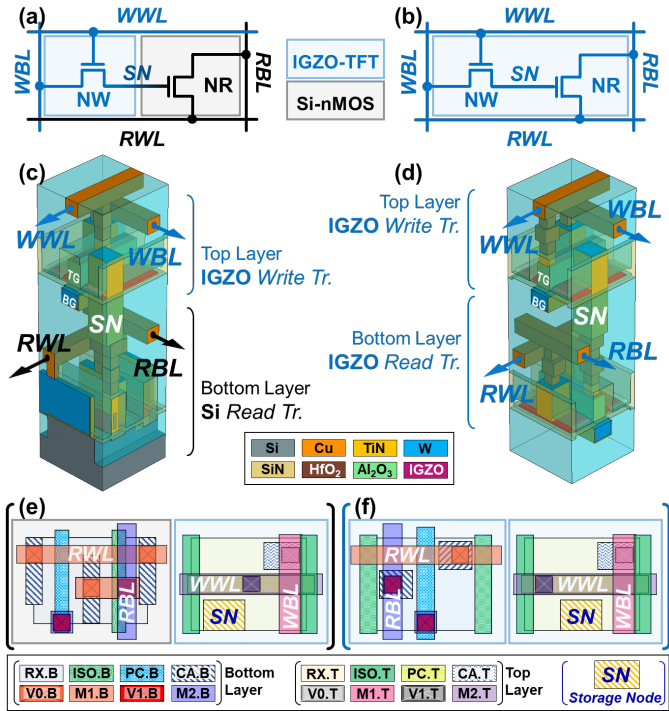


Fig. 2. Schematics of (a) Hybrid IGZO/Si eDRAM (Hybrid-2T) and (b) IGZO-based eDRAM (IGZO-2T). 3D architectures of (c) Hybrid-2T and (d) IGZO-2T. Layout and layer details of the (e) read transistor (NR) and write transistor (NW) in Hybrid-2T, and (f) NR and NW in IGZO-2T.

- We propose the Hybrid-3D memory structure and process architecture to achieve high memory density while maintaining performance similar to SRAM. Hybrid-3D offers  $22\times$  the on-chip memory capacity compared to Si CMOS SRAM.
- We validate the operation of Hybrid-3D by conducting a macro-level memory design using 3D technology computer-aided design (TCAD) process and SPICE circuit simulations. These simulations confirmed that reliable memory operations were maintained across all vertically stacked layers.
- We demonstrate the advantages of Hybrid-3D by evaluating the system-level operation of NPU equipped with a Hybrid-3D-based on-chip buffer. The results showed up to a  $3.2\times$  increase in energy efficiency and a  $2.6\times$  increase in throughput compared with NPU with SRAM buffer.

## II. BACKGROUND AND MOTIVATION

### A. Background: Limitations of SRAM scaling

SRAM is an essential component for CPUs, serving as cache memory, and for NPUs, acting as buffer memory due to its high-speed operation [1], [16]. However, the requirement of six transistors per cell limits area efficiency, making on-chip memory capacity a critical factor in system design. Push-rules have been applied in each technology generation to improve density and energy efficiency [17]. Nevertheless, as lithography scaling slows, SRAM now scales less effectively than logic circuits [7]. Thus, expanding on-chip memory requires circuit–system co-design with careful consideration of device technology.

### B. Motivation: IGZO-based eDRAM Technologies

Research aimed at increasing density while maintaining the high performance of SRAM is advancing rapidly. CMOS-based eDRAM [8], [9] and IGZO-based eDRAM [10], [11], [13] are two popular memory technologies that rely solely on conventional materials and processes commonly used in the industry. A clear trade-off exists between these two technologies. CMOS-based eDRAM can achieve performance levels comparable to SRAM; however, it suffers from refresh overhead due to its data retention time at the  $\mu\text{s}$  scale. This limitation results in additional power and timing overhead. In contrast, IGZO-based eDRAM can avoid refresh overhead due to its retention time exceeding 10 s. However, its low mobility makes it challenging to replace SRAM due to slow read/write time. Recently, Hybrid-2T, which uses single-crystal silicon for read transistors and IGZO for write transistors, has been proposed as an alternative that meets both the need for a long refresh time and high performance [14], [15]. However, the Si-based read transistor in Hybrid-2T restricts its potential for monolithic 3D integration. Thus, demonstrating a memory architecture that combines the high cell density, performance, long retention time of Hybrid-2T, and compatibility with 3D integration technology becomes crucial.

## III. PROPOSED HYBRID-3D MEMORY ARCHITECTURE

### A. Bit Cell Configurations

*a) Cell Structure:* Each unit cell in Hybrid-3D functions as an eDRAM-based memory, providing a compact area with high performance. Figs. 2(a) and (b) show the cell schematics for Hybrid-2T and IGZO-2T, respectively, showing both read transistors (NR) and write transistors (NW). The structural characteristics of each cell center on two aspects: 1) the channel materials used for NR and NW, and 2) architectural design. Hybrid-2T incorporates different channel materials, with Si nMOS for NR and IGZO-TFT for NW (Fig. 2(c)). In contrast, IGZO-2T uses IGZO-TFT transistors for both NR and NW (Fig. 2(d)). Another key structural feature in both Hybrid-2T and IGZO-2T is the  $90^\circ$  rotation between NW and NR, which enables a compact cell layout and flexible design options. Both types of cells employ M1/M2 layers on the bottom layer for the read word-line (RWL) and read bit-line (RBL), while the top layer positions M1/M2 layers for the write word-line (WWL) and write bit-line (WBL), forming the in/out ports.

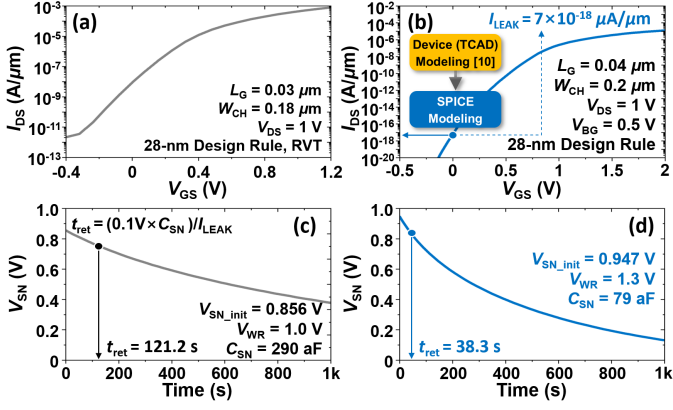


Fig. 3. (a) Gate voltage ( $V_{GS}$ ) - drain current ( $I_{DS}$ ) curve of NR in Hybrid-2T and (b) IGZO thin-film transistor (TFT) modeling [10] based on 28-nm technology. Retention characteristics of (c) Hybrid-2T and (d) IGZO-2T.

**b) Cell Layouts:** The cell layout of Hybrid-2T and IGZO-2T is designed based on the design rules of 28-nm technology. The most notable point is that the identical feature size of Hybrid-2T and IGZO-2T results in a cell density  $5.5\times$  higher than that of SRAM. Fig. 2(e) shows the layouts of NR and NW of Hybrid-2T, while Fig. 2(f) presents the cell layouts of IGZO-2T. The detailed dimensions of each layout component are available in Table I. A distinctive layout feature is that the storage node (SN) in NW aligns with the gate contact of NR. With NW (PC.T) and NR (PC.B) gates arranged orthogonally, the gate of NR coincides with the drain of NW, providing additional cell density gains beyond simple NR–NW stacking.

### B. Modeling and Electrical Characteristics

**a) Transistor-Level Modeling:** To predict and validate the seamless operation of Hybrid-2T and IGZO-2T within Hybrid-3D, accurate modeling of the bit cell characteristics is crucial. We performed sequential TCAD-level and SPICE-level modeling based on the characteristics of fabricated devices [18] and the process design kit (PDK) to ensure consistency. Hybrid-2T has characteristics based on the CMOS process, so we used the SPICE model of a 28-nm PDK (Fig. 3(a)). For IGZO-2T, we conducted TCAD modeling based on fabricated devices [10], then created a Verilog-A model and performed rigorous calibration (Fig. 3(b)). IGZO, with a wide bandgap of 3.0 eV, strongly suppresses gate-induced drain leakage (GIDL), resulting in an extremely low subthreshold leakage ( $I_{LEAK}$ ) of  $7 \times 10^{-18} \mu A/\mu m$ .

**b) Retention Time Evaluation:** Figs. 3(c) and (d) show the  $t_{ret}$ s of Hybrid-2T and IGZO-2T, measured at 121.2 and 38.3 s, respectively. After writing data “1” (D1) in the write condition, the SN voltage ( $V_{SN}$ ) is monitored while maintaining the hold condition. Due to the difference in equivalent oxide thickness (EOT), the gate oxide capacitance ( $C_{OX}$ ) of the NR in Hybrid-2T is higher than that of IGZO-2T (1.5 and 6 nm). Therefore, the storage node capacitance  $C_{SN}$  differs significantly, at 290 and 79 aF, respectively. Even with the same  $I_{LEAK}$  flowing through the NW in each cell, this capacitance difference causes a variance in  $t_{ret}$ . Nonetheless, both Hybrid-2T and IGZO-2T achieve excellent retention times  $t_{ret}$  of

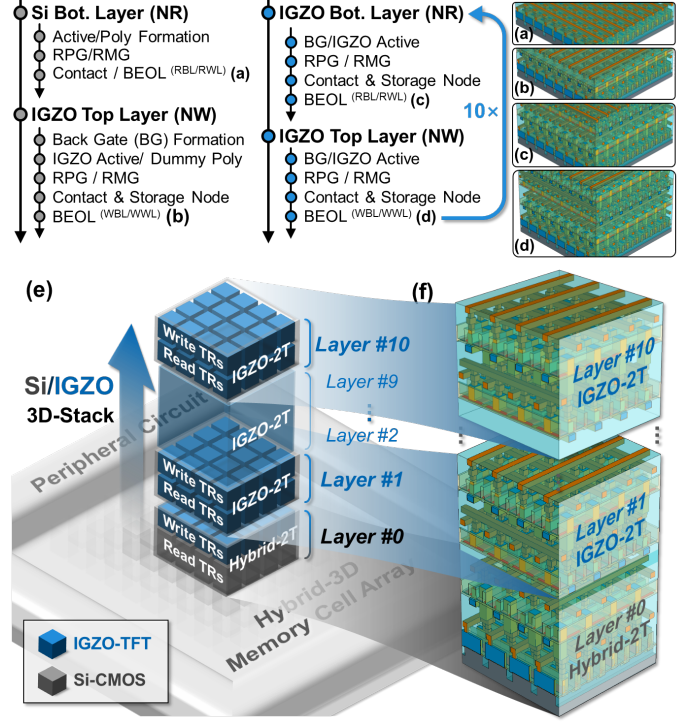


Fig. 4. 3D process integration of Hybrid-3D: (a) bottom and (b) top layers of Hybrid-2T, (c) bottom and (d) top layers of IGZO-2T, and (e) conceptual 3D view of the Hybrid-3D memory and (f) detailed cell array structure.

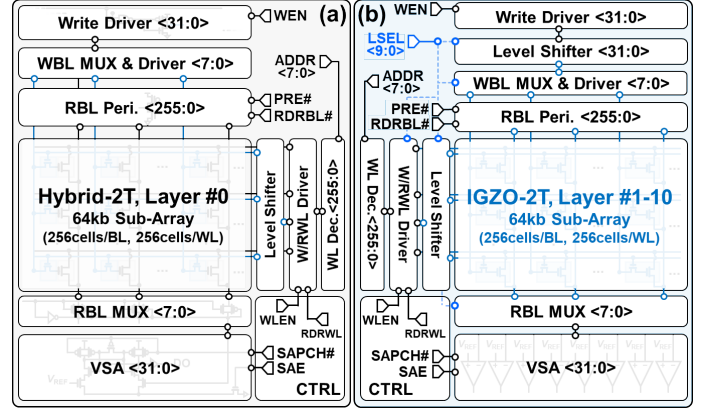


Fig. 5. Design of the memory macro for (a) Hybrid-2T and (b) IGZO-2T. Concept diagram including the placement of each block and schematic.

over 30 s, providing long enough time for use in NPU on-chip memory to effectively disregard refresh cycles during operations.

### C. Process Integration Scheme

Hybrid-3D integrates Hybrid-2T and IGZO-2T cells, vertically stacked via monolithic 3D integration. We performed 3D-TCAD process simulations to predict cell/circuit performance from the 3D structure and verify port-connection design rules. The process flow maps as follows: Hybrid-2T NR and NW correspond to Figs. 4(a) and (b), and IGZO-2T NR and NW to (c) and (d); in all layers, NW and NR are fabricated using a high- $\kappa$  metal-gate process. Hybrid-3D comprises 11 layers: layer #0 uses Hybrid-2T with a single-crystal Si channel, while layers #1–#10 use IGZO-2T. Starting from layer #0, IGZO

TABLE III  
FEATURE SUMMARY

Hardware Configuration		Unit	SRAM	Hybrid-3D
On-Chip Memory Design	Technology	-	28-nm	
	Cell Configuration	-	6T SRAM	Hybrid-2T/IGZO-2T
	Norm. Cell Density*	-	1×	5.5×
NPU Design	PE Array	ea		32×32
	DRAM Bandwidth	GB/s		2.5
	Operation Frequency	MHz		500
	On-Chip Memory Capacity	MB	1 (1×)	22 (22×)

\*Reproduced by the logic design rules of the 28-nm PDK

The LSEL<9:0> signals denote active layers from #1 to #10, while non-active layers remain in hold state to reduce power consumption. Two dedicated level shifters are used: one between the WWL/RWL driver and the sub-array, and another between the write driver and the sub-array. During write operations, the high supply voltage ( $V_{DDH} = 1.8$  V) is applied to the WWL of Hybrid-2T to drive its NW transistor and to the WBL of IGZO-2T to drive its NR transistor. In addition, the WWL level shifter generates a boosted voltage ( $V_{DDB} = 2.15$  V) to drive the WWL of IGZO-2T, thereby activating its NW transistor and enabling sufficient charging of  $V_{SN}$  during D1 writes. Since the nominal drive voltage for the 28-nm PDK is 1 V, the read operation voltage (RBL/RWL voltages for both Hybrid-2T and IGZO-2T) is constrained by PDK specifications, which strictly define cell operation and peripheral integration as shown in Table II.

### B. RC Components for Vertical Stacking

In the Hybrid-3D structure, IGZO-2T layers share common peripheral circuits at the bottom, so resistance and capacitance ( $R_v, C_v$ ) from vertical stacking affect signal transfer (Fig. 6(a)). As stack height increases, the sub-array–peripheral connection length grows, raising  $R_v$  and  $C_v$  proportionally in upper layers. These components influence read/write signals, with IGZO-2T showing slower RBL development than Hybrid-2T due to the limited drivability of IGZO-TFTs. Upper layers suffer further delay, so the RBL development time is set by layer #10 to guarantee reliable sensing. The vertical RBL pitch is  $2.4 \mu\text{m}$ , while the horizontal matches the cell pitch.  $R_v$  includes metal, via, and vertical contact resistance ( $30 \Omega/\text{layer}$ ), and  $C_v$  arises mainly from wire coupling ( $3 \text{ fF}/\text{layer}$ ).

Fig. 6(b) shows read behavior from layer #1–#10. Compared to 2D structures, 3D stacking requires a longer read time, here 10 ns. Layer #10 exhibits a data-out delay ( $t_{\text{SAD}}$ ) of 196 ps, degraded by 38 ps from layer #1, but still reliable with extended cycles. Fig. 6(c) highlights SAE# activation, showing a 20 mV reduction in sensing margin between top and bottom layers. The worst-case margin occurs between D0 at layer #1 and D1 at layer #10, so  $V_{\text{REF}} = 950$  mV is chosen for safe separation. Thus, full-stack read/write characteristics must be carefully evaluated to ensure reliable operation under  $R_v, C_v$ , and reduced sensing margin.

### C. Read/Write Characteristics

The read/write operation of Hybrid-3D is verified at 500 MHz. Figs. 7(a) and (b) show timing for Hybrid-2T and IGZO-2T, where the key difference is observed in cycle time ( $t_{\text{read}}$ ,

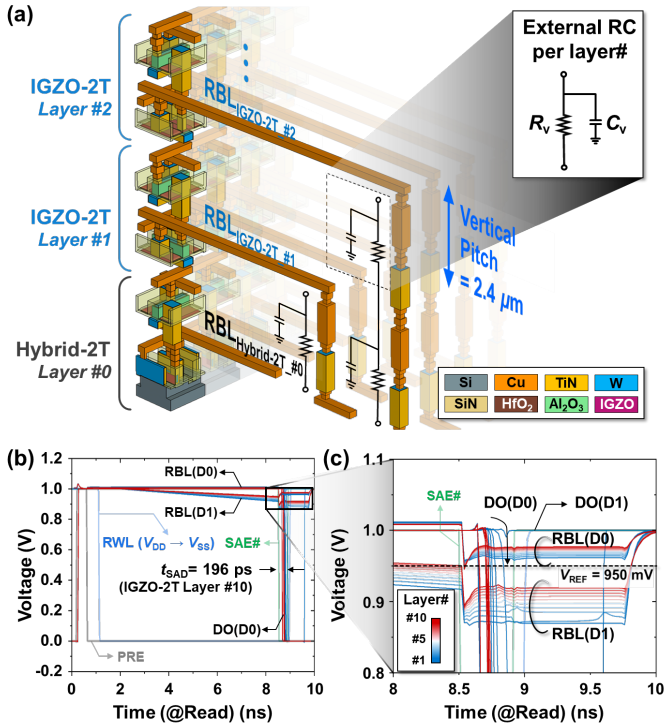


Fig. 6. (a) External resistance and capacitance in the vertical direction ( $R_v$  and  $C_v$ ) of the stacked IGZO-2T. (b) D1/D0 read timing diagrams for layers from #1 to 10 and (c) rigorous verification of the sensing margin.

TABLE II  
MEMORY OPERATION CONDITIONS OF HYBRID-2T AND IGZO-2T

	Hybrid-2T				IGZO-2T			
	WWL	WBL	RWL	RBL	WWL	WBL	RWL	RBL
<b>Hold</b>	$V_{SS}$	$V_{SS}$	$V_{DD}$	$V_{DD}$	$V_{SS}$	$V_{SS}$	$V_{DD}$	$V_{DD}$
<b>Write</b>	$V_{DDH}^{(1)}$	$V_{SS}/V_{DD}^{(3)}$	$V_{DD}$	$V_{DD}$	$V_{DDB}^{(2)}$	$V_{SS}/V_{DDH}$	$V_{DD}$	$V_{DD}$
<b>Read</b>	$V_{SS}$	$V_{SS}$	$V_{DD} \rightarrow V_{SS}$	$V_{DD}(\text{PCH})^{(4)}$	$V_{SS}$	$V_{SS}$	$V_{DD} \rightarrow V_{SS}$	$V_{DD}(\text{PCH})$

<sup>1)</sup>  $V_{DDH} = 1.8$  V / <sup>2)</sup>  $V_{DDB} = 2.15$  V for Hybrid/IGZO-2T NW/NR ( $V_{DDH}$ ) and IGZO-2T NW ( $V_{DDB}$ )

<sup>3)</sup> D0/D1 write condition of Hybrid-2T

<sup>4)</sup> Pre-charge condition for selected cells

channels are vertically stacked, with NR and then NW added iteratively to complete all 10 IGZO-2T layers. Feature sizes remain constant across layers for precise alignment. Fig. 4(e) provides the 3D overview consolidating the process integration, while (f) presents the concrete 3D structure rigorously validated by the process simulation.

## IV. MACRO ARRAY CIRCUIT DESIGN AND ANALYSIS

### A. Circuit-Level Memory Macro Designs

The Hybrid-3D memory macro integrates Hybrid-2T for layer #0 and IGZO-2T for layers #1–#10. To minimize circuit overhead and maximize capacity, each type of macro (Hybrid-2T and IGZO-2T) shares one memory control and a set of peripheral circuits (Figs. 5(a) and (b)). Each IGZO-2T sub-array provides 64 kb, configured as a  $256 \times 256$  cell array. Compared to an SRAM macro, the proposed design introduces two major features: 1) a layer select signal (LSEL) for efficient control of stacked IGZO-2T layers, and 2) level shifters for proper voltage management of IGZO-TFT cells.

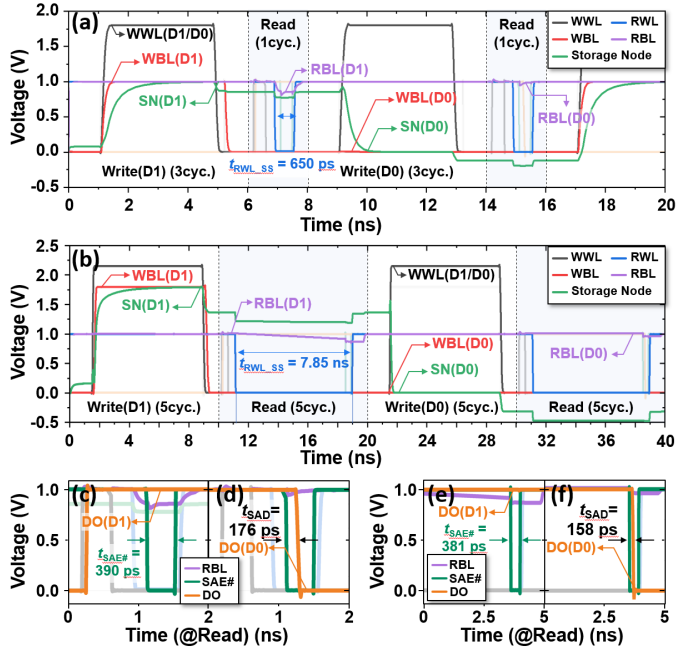


Fig. 7. Timing diagram for the write/read of data "1" (D1) and data "0" (D0) in (a) Hybrid-2T and (b) IGZO-2T. Analysis of the data out (DO) signal for (c) D1 and (d) D0 in Hybrid-2T and the corresponding analysis for (e) D1 and (f) D0 in IGZO-2T.

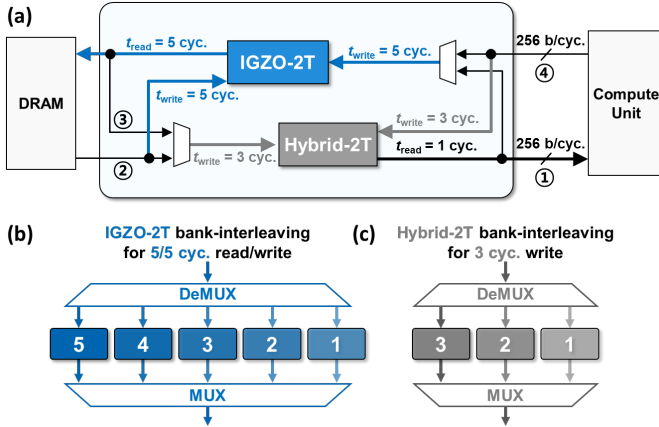


Fig. 8. (a) Supported communication paths of Hybrid-3D-based on-chip memory. (b) Interleaving strategy of Hybrid-2T/IGZO-2T enabling memory operations equivalent to SRAM.

$t_{write}$ ) due to the lower current drivability of IGZO transistors compared to Si. The design is defined in three aspects: 1) Hybrid-2T  $t_{read}$  is 2 ns (1 cycle), equivalent to SRAM speed, indicating that Hybrid-3D operates at the same level as SRAM in terms of data read. 2) Hybrid-2T  $t_{write}$  is 6 ns (3 cycles), extended to compensate for weak NW drivability. 3) IGZO-2T  $t_{read}/t_{write}$  are 10 ns (5 cycles), where longer  $t_{write}$  is required to charge RBL to  $V_{DDB}$ , and extended  $t_{read}$  ensures robustness of the top IGZO-2T layer.

A notable difference is observed in  $t_{RWL\_SS}$ , the interval during which RWL remains active ( $V_{RWL} = V_{SS}$ ): 650 ps for Hybrid-2T and 7.85 ns for IGZO-2T. This extended time secures the sensing margin and represents the main cause of the read latency gap.  $t_{read}$  must be regulated according to the

time designated for RWL to remain at  $V_{SS}$ , since during this period RBL discharges and the sense amplifier (SA) must be enabled to detect against  $V_{REF}$ . Figs. 7(c)–(f) show DO behavior, with  $t_{SAD}$  after SA enable measured as 176 ps in Hybrid-2T and 158 ps in IGZO-2T. As can be confirmed in Table II, these differences in operating conditions highlight the distinct characteristics of each structure. In summary, Hybrid-3D integrates Hybrid-2T and IGZO-2T, each showing distinct  $t_{read}/t_{write}$ , so an on-chip buffer design equivalent to SRAM is required for practical use. This highlights that although Hybrid-3D can significantly enhance memory density, careful timing regulation and buffer design are indispensable to ensure stable and efficient system-level operation.

#### D. On-Chip Buffer Design with SRAM Equivalence

Fig. 8(a) shows the proposed memory buffer structure designed to ensure that the Hybrid-3D-based memory exhibits the same read/write characteristics as an SRAM buffer. This structure highlights communication paths with the DRAM, the compute unit, and newly added internal paths between Hybrid-2T and IGZO-2T. In this configuration, ① the compute unit receives data from the Hybrid-2T in Hybrid-3D memory. Since the  $t_{read}$  of Hybrid-2T is 1 cycle which is identical to that of SRAM, Hybrid-3D-to-compute unit data transfers achieve the same throughput as SRAM-based buffers without requiring any modifications. ② DRAM data can be transferred to both IGZO-2T and Hybrid-2T. Despite the  $t_{write}$  of IGZO-2T being 5 cycles and that of Hybrid-2T being 3 cycles, the DRAM-to-Hybrid-3D data transfer rate remains consistent with DRAM-to-SRAM transfers because the data transfer rate from DRAM is inherently lower than the on-chip data rates, obviating the need for circuit modifications. ③ For internal data transfers from IGZO-2T to Hybrid-2T, the longer  $t_{read}$  of IGZO-2T and the  $t_{write}$  of Hybrid-2T compared to those of SRAM are effectively hidden using a bank-interleaving scheme, as illustrated in Figs. 8(b) and (c). This interleaved bank structure ensures that the read/write throughput aligns with that of an SRAM buffer as shown in Table III (256 bit/cycle). Similarly, ④ for data transfers from the compute unit to IGZO-2T or Hybrid-2T, the longer  $t_{write}$  can also be masked using the interleaving scheme. Consequently, the proposed Hybrid-3D buffer delivers macro-level read/write characteristics equivalent to those of an SRAM buffer, while offering significantly greater memory capacity.

#### E. On-Chip Memory Configurations

The sub-array level expands to a  $256 \times 256$  configuration, placing x/y-direction decoders and memory controllers among four sub-arrays to form a single sub-bank. Four sub-banks combine to create one bank. Hybrid-2T employs four banks, while IGZO-2T utilizes eight banks through interleaving. The  $5.5 \times$  higher cell density of Hybrid-2T enables a twofold increase in on-chip memory capacity compared to 1 MB SRAM within the same area, achieving 2 MB using Hybrid-2T alone. The smaller relative increase in memory capacity at the array level, compared to the improvement in cell density, arises from the limited scalability of peripheral circuits. Encouragingly, with

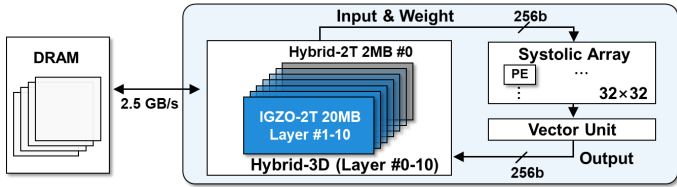


Fig. 9. Neural processing unit (NPU) utilizing Hybrid-3D as on-chip memory. System-level.

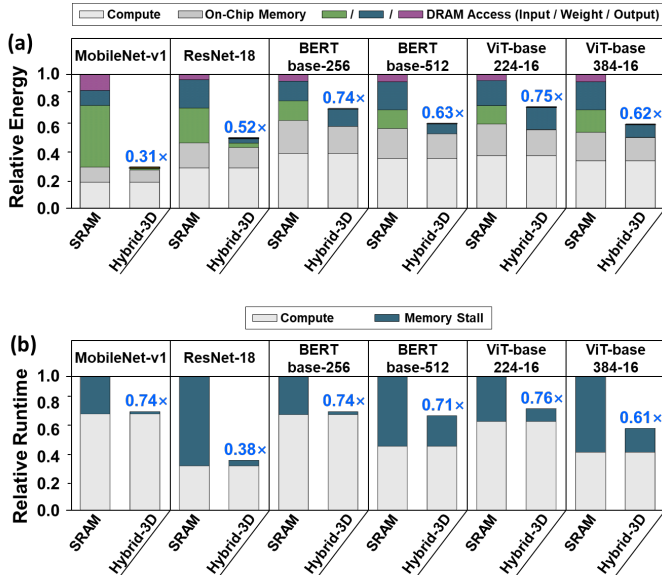


Fig. 10. (a) Neural processing unit (NPU) utilizing Hybrid-3D as on-chip memory. System-level (b) energy consumption and (c) runtime analysis for vision and language models with Hybrid-3D and SRAM used as on-chip memory.

the 3D integration of IGZO-2T securing an additional capacity of 20 MB, Hybrid-3D achieves a 22 $\times$  on-chip memory capacity advantage over SRAM. Table III summarizes the features of Hybrid-3D compared to SRAM.

## V. SYSTEM-LEVEL EVALUATION

### A. Experimental Setup

*a) NPU with Hybrid-3D On-Chip Buffer:* We implemented the edge NPU architecture as shown in Fig. 9 to evaluate the system-level impact of Hybrid-3D.

This design utilizes 22 MB Hybrid-3D as the on-chip memory, whereas the baseline architecture employs 1 MB SRAM.

The NPU architecture employs a systolic array with a weight-stationary dataflow for matrix multiplication [1]. Following the hardware configuration of a previous edge NPU [3], the systolic array consists of 32 $\times$ 32 processing elements (PEs). Consistent with DRAM settings in edge devices, the DRAM bandwidth is set to 2.5 GB/s [3], assuming the use of LPDDR DRAM [19]. A vector unit is responsible for executing vector operations, such as partial sum accumulation and activation functions. Energy consumption and overall runtime of the NPUs are evaluated by using the SCALE-Sim [20], a simulator for systolic array based accelerators.

*b) Benchmark:* To compare the system-level performance of NPUs incorporating the proposed Hybrid-3D memory versus conventional SRAM, we evaluate their energy consumption and runtime across various deep learning models. For vision tasks, we include convolutional neural networks (CNNs) such as MobileNet-v1 [21] and ResNet-18 [22], as well as Vision Transformers (ViTs) like ViT-base-224-16 and ViT-base-384-16 [23], which use image sizes of 224 $\times$ 224 and 384 $\times$ 384, respectively, with a patch size of 16 $\times$ 16. For natural language processing (NLP) tasks, we evaluate BERT-base models, with input token lengths indicated in their names, such as BERT-base-256 and BERT-base-512 [24].

### B. Evaluation Results

The evaluation results for energy consumption and overall runtime are presented in Figs. 10(a) and (b), respectively. By increasing on-chip memory capacity by 22 $\times$  compared to SRAM, Hybrid-3D significantly reduces data movement between DRAM and the NPU. Thus, Hybrid-3D delivers improvements in both energy efficiency and speedup across a range of deep learning models. The performance improvements with Hybrid-3D are most pronounced where DRAM access is the primary performance bottleneck. For example, in MobileNet-v1, which suffers from significant energy overhead due to frequent DRAM access, adopting Hybrid-3D reduces overall energy consumption by 69% compared to the baseline and improves energy efficiency by 3.2 $\times$ . Similarly, for ResNet-18, which experiences significant memory stalls, adopting Hybrid-3D reduces overall runtime by 62% compared to the baseline and improves throughput by 2.6 $\times$  compared to the baseline.

## VI. CONCLUSION

In this paper, Hybrid-3D is proposed to increase memory capacity and achieve high performance. The Hybrid-2T/IGZO-2T unit cell inside the memory architecture achieves a 5.5 $\times$  improvement in cell density compared to SRAM. Thanks to monolithic 3D integration, Hybrid-3D improves on-chip memory capacity by 22 $\times$ . Transistor-level modeling and circuit-level design of Hybrid-3D memory operations are finalized, followed by rigorous verification. The evaluation includes a thorough assessment of external components due to the stacking structure, ensuring reliable operation. Furthermore, the introduction of Hybrid-3D as on-chip memory in the NPU architecture demonstrates superior performance. The system-level evaluation covers tasks from vision to NLP, achieving up to 3.2 $\times$  energy efficiency improvement and 2.6 $\times$  increase in throughput.

### 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, RS-2024-00395134, DPU-Centric Datacenter Architecture for Next-Generation AI Devices, No. 2021-0-01343: IITP-2023-RS-2023-00256081: artificial intelligence semiconductor support program to nurture the best talents), BK21 FOUR program and ISRC at Seoul National University. The EDA tool was supported by the IC Design Education Center(IDECE). (Corresponding Author: Jae-Joon Kim).

## REFERENCES

- [1] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-I. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snellman, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA '17*, (New York, NY, USA), Association for Computing Machinery, 2017.
- [2] N. P. Jouppi, D. Hyun Yoon, M. Ashcraft, M. Gottscho, T. B. Jablin, G. Kurian, J. Laudon, S. Li, P. Ma, X. Ma, T. Norrie, N. Patil, S. Prasad, C. Young, Z. Zhou, and D. Patterson, "Ten lessons from three generations shaped google's tpuv4i : Industrial product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1–14, 2021.
- [3] C.-H. Lin, C.-C. Cheng, Y.-M. Tsai, S.-J. Hung, Y.-T. Kuo, P. H. Wang, P.-K. Tsung, J.-Y. Hsu, W.-C. Lai, C.-H. Liu, *et al.*, "7.1 a 3.4-to-13.3 tops/w 3.6 tops dual-core deep-learning accelerator for versatile ai applications in 7nm 5g smartphone soc," in *2020 IEEE International Solid-State Circuits Conference-ISSCC*, pp. 134–136, IEEE, 2020.
- [4] S. Lie, "Cerebras architecture deep dive: First look inside the hw/sw co-design for deep learning: Cerebras systems," in *2022 IEEE Hot Chips 34 Symposium (HCS)*, pp. 1–34, IEEE Computer Society, 2022.
- [5] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, *et al.*, "Dadianna: A machine-learning supercomputer," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 609–622, IEEE, 2014.
- [6] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 27–39, 2016.
- [7] T. Song, H. Kim, W. Rim, Y. Kim, S. Park, C. Park, M. Hong, G. Yang, J. Do, J. Lim, S. Lee, I. Kim, S. Baek, J. Jung, D. Ha, H. Jang, T. Lee, C.-H. Park, B. Kwon, H. Jung, S. Cho, Y. Choo, and J. Choi, "12.2 a 7nm finfet sram macro using euv lithography for peripheral repair analysis," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 208–209, 2017.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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  $I_g$  scalability down to 14 nm," in *2021 IEEE International Electron Devices Meeting (IEDM)*, pp. 10.6.1–10.6.4, 2021.
- [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] M. Kim and J.-J. Kim, "4-transistor ternary content addressable memory cell design using stacked hybrid igzo/si transistors," in *Proceedings of the 61st ACM/IEEE Design Automation Conference, DAC '24*, (New York, NY, USA), Association for Computing Machinery, 2024.
- [15] 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.
- [16] J. Chang, M. Huang, J. Shoemaker, J. Benoit, S.-L. Chen, W. Chen, S. Chiu, R. Ganesan, G. Leong, V. Lukka, S. Rusu, and D. Srivastava, "The 65-nm 16-mb shared on-die l3 cache for the dual-core intel xeon processor 7100 series," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 4, pp. 846–852, 2007.
- [17] J. Wang, X. Wang, C. Eckert, A. Subramanian, R. Das, D. Blaauw, and D. Sylvester, "A 28-nm compute sram with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE Journal of Solid-State Circuits*, vol. 55, pp. 76–86, Jan 2020.
- [18] S. He, H. Li, G. Xu, X. Tang, Y. Li, J. Kim, T. Gu, X. Xue, Z. Li, H. Xu, H. Dong, K. Zhou, X. Hu, and S. Long, "Modeling the thermal characteristics of stacked 2t0c memory array based on ingazno4 thin-film transistors," *IEEE Transactions on Electron Devices*, vol. 70, no. 12, pp. 6369–6374, 2023.
- [19] K. T. Malladi, B. C. Lee, F. A. Nothaft, C. Kozyrakis, K. Periyathambi, and M. Horowitz, "Towards energy-proportional datacenter memory with mobile dram," *ACM SIGARCH Computer Architecture News*, vol. 40, no. 3, pp. 37–48, 2012.
- [20] A. Samajdar, J. M. Joseph, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna, "A systematic methodology for characterizing scalability of dnn accelerators using scale-sim," in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 58–68, 2020.
- [21] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.
- [22] 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*, pp. 770–778, 2016.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
- [24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.