Received 15 May 2026; revised 17 September 2026; accepted 30 September 2026. Date of publication 6 October 2026; date of current version 15 October 2026.

Digital Object Identifier 10.1109/JXCDC.2026.3740041

Device–Circuit–System Co-Design of Reconfigurable Oxide–Semiconductor 2T Nonvolatile Gain-Cell Memory and Sandwich-3-D LLM Accelerator

LUQI ZHENG (Graduate Student Member, IEEE), CHANG NIU, SHUTING DU (Graduate Student Member, IEEE), SUMI LEE (Graduate Student Member, IEEE), AMIR MASSAH BAVANI (Member, IEEE), ZEHAO LIN, PEIDE D. YE (Fellow, IEEE), and HAITONG LI (Member, IEEE)

Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA

CORRESPONDING AUTHOR: H. LI (haitongli@purdue.edu)

This work was supported in part by the U.S. NSF under Award 2425498 and in part by NSF CAREER Award 2543547.

ABSTRACT In this work, we present an oxide–semiconductor (OS)-based 2T nonvolatile gain-cell (NVGC) memory and a "sandwich 3-D" stacking scheme, co-designed using atomic layer deposition (ALD) InOx oxide–semiconductor FETs (OSFETs) and OS-ferroelectric FETs (FeFETs), for efficient large language model (LLM) acceleration. The new 2T-NVGC enables native, in-memory reconfiguration of static/dynamic kernels for transformers and context-switching workloads, while realizing efficient floating point (FP)/integer (INT) compute-in-memory (CIM). With full-stack, cross-layer co-designs, we present the first complete mixed-signal physical design implementation of an OS-on-CMOS 3D-integrated circuit (IC) with 32.9 TOPS/W energy efficiency and 102.4 TOPS/mm^2 compute density. The post-3D-layout simulations calibrated with device experiments demonstrate that the 3D-NVGC LLM accelerator delivers a 52.3× gain in holistic system efficiency over planar CMOS-based and a 3.2× gain over 3-D dynamic random-access memory (DRAM)-based solutions, respectively.

INDEX TERMS 3D-integrated circuit (IC), compute-in-memory (CIM), device-circuit co-design, ferroelectric FET (FeFET), floating-point (FP) acceleration, large language model (LLM) accelerator, nonvolatile gain-cell (NVGC) memory, oxide–semiconductor (OS), reconfigurable memory.

I. INTRODUCTION

THE large language models (LLMs) driving remarkable advances across diverse areas of computing and intelligence demand efficient hardware solutions without sacrificing precision. The state-of-the-arts that exploit promising "architectural recipes," such as compute-in-memory (CIM) and 3-D [1], [2], [3], [4], [5], [6], [7], [8], and "technological ingredients," including back-end-of-line (BEOL) logic (e.g., amorphous oxide–semiconductor (OS) transistors [9]) and memories [e.g., FeRAM/ferroelectric FETs (FeFETs)] [10], [11], [12], [13], [14], [15], still face several major challenges, as summarized in Fig. 1. First, transformers attention mechanism requires both static and dynamic matrix operations with evolving model- and task-dependency. Such fundamental fine-grained reconfigurability is not yet attainable with any incumbent or emerging memories. Second, floating-point (FP) formats, crucial for reliable LLM workloads, lack efficient in-memory realization even in advanced CIM designs due to weight prealignment padding overheads [1], [2]. Last, the 3-D memory architectures with 3-D V-NAND/Fe-NAND [7], [16] and 3-D dynamic random-access memory (DRAM) [8] restrict all logic and peripheral circuits to the CMOS-under-array (CuA) tier, ultimately limiting compute efficiency due to confined silicon footprint with intensive, intermediate data movement down to CuA.

To tackle all three challenges simultaneously, we adopt a device–circuit–system co-design approach experimentally harnessing atomic layer deposition (ALD) InOx-based oxide–semiconductor FETs (OSFETs)

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

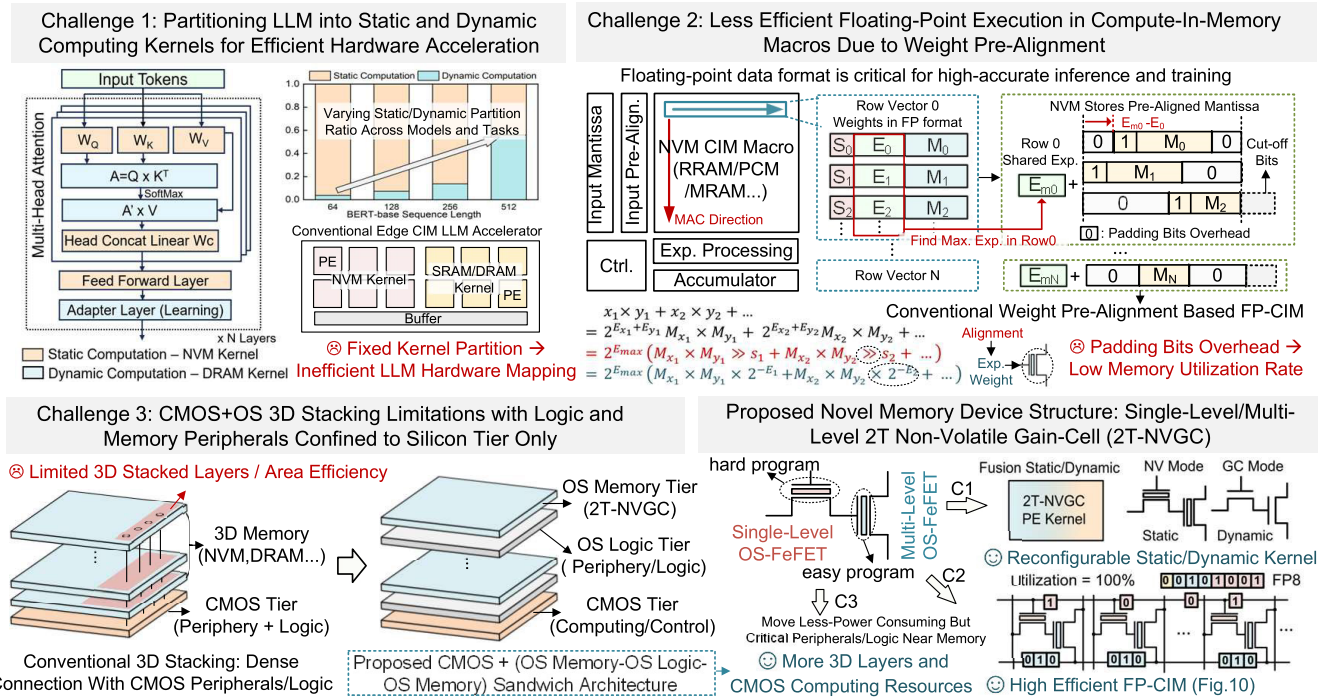


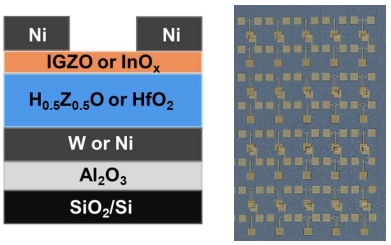
FIGURE 1. **Key challenges in LLM accelerators and our full-stack co-design solution harnessing OS technologies.**

and oxide–semiconductor FeFETs (OS-FeFETs). A new two-transistor nonvolatile (NV) gain-cell (2T-NVGC) memory is developed to enable reconfigurable static/dynamic kernels natively in memory for context-switching LLM adaptability, while realizing high-utilization FP data storage and high-efficiency FP computation. Co-optimizing NV gain-cell (NVGC) memory and InOx mixed-signal circuits, we further develop an OS-memory/OS-logic/OS-memory "sandwich 3-D" architecture for scalable BEOL-integrated AI acceleration. Unlike conventional CuA 3-D memory architectures, where peripheral and logic circuits remain confined to a single silicon CMOS tier, the proposed "sandwich 3-D" scheme vertically distributes memory arrays, mixed-signal peripherals, and lightweight logic across multiple OS BEOL tiers. Such an architecture is uniquely enabled by OS technologies, which simultaneously support both memory devices (OS-FeFETs/OS-eDRAM) and logic transistors (InOx OSFETs) within BEOL-compatible fabrication processes. This organization fundamentally differs from prior 3-D memory stacking approaches, as it makes each vertically integrated memory tier computation-capable in an efficient manner, and can be applied to enhance compute density and vertical scalability for other 3-D CIM designs. The proposed architecture alleviates CuA routing congestion and intermediate data movement bottlenecks while improving compute scalability, memory utilization, and intertier communication efficiency for LLM accelerators. With our experimentally calibrated OS standard cell library and OS process design kit (PDK), this work provides the first complete mixed-signal physical design implementation of a 3-D OS-on-CMOS LLM accelerator, considering practical implications of high-performance nMOS-only OSFETs and a relaxed OS node on silicon.

II. NVGC MEMORY

Leveraging the low-leakage characteristics of wide bandgap OSs, the OSFETs are well-suited for 2T gain-cell (GC) eDRAMs with separate access and read transistors [17], [18]. Compared with conventional silicon-based GC memories, the OS devices provide improved retention characteristics while maintaining BEOL compatibility for monolithic 3-D integration. Our 2T-NVGC retains this compact topology but replaces both transistors with OS-channel ferroelectric FETs (OS-FeFETs), enabling native reconfiguration between dynamic (charge-based) and NV modes for efficient LLM hardware mapping and context-switching workloads. Compared with a 1T-1FeFET cell [19], the 2T-NVGC offers significant circuit-level advantages when addressing the device-level challenges in programming the access-FeFET in the 2-FeFET GC topology, while also improving flexibility for mixed FP/integer (INT) CIM operations.

To leverage 2T-NVGC for FP-CIM, as shown in Fig. 1, exponent bits are stored in the read-FeFETs, enabled by reliable multibit storage via read-verify programming, while mantissa bits are stored bitwise in the access FeFETs, optimized for single-bit one-shot programming. This structure realizes exponent-weighted mantissa FP-CIM, where exponents encoded in the read-FeFET states are directly reflected



- SiO_2/Si substrates
- Al_2O_3 adhesion layer deposition
- Gate Metallization 40 nm Ni or 35 nm W
- Dielectric deposition ALD 6 nm HfO_2 or 10 nm HZO
- Channel semiconductor deposition ALD InO_x or IGZO In:Ga:Zn = 3:1:1
- 400 °C N_2 annealing (Ferroelectric)
- Source/Drain Metallization 40 nm Ni
- Channel dry etch isolation
- O_2 190 to 300 °C annealing

FIGURE 2. Device schematic, image, and fabrication flow of ALD InOx-based FETs and FeFETs.

in the read current, eliminating the need for mantissa prealignment and padding overhead in the state-of-the-art CIM circuits. Since exponent differences beyond 8 can be truncated with negligible accuracy loss, a 3-bit read-FeFET with eight states is sufficient to encode most FP formats used in practical LLM acceleration. The NVGC devices and OS-logic transistors are fabricated and characterized to support subsequent 3-D accelerator circuit design.

The device structure, optical image, and fabrication flow of the ALD InOx-based logic and memory devices (In2O3-FETs and IGZO-FeFETs) are illustrated in Fig. 2. The gate electrode consists of either a 40-nm Ni or a 35-nm W layer. The gate dielectric is formed by plasma-enhanced ALD (PE-ALD) of 6-nm HfO_2 or 10-nm $Hf_{0.5}Zr_{0.5}O_2$ (HZO), deposited at 200 °C. The semiconducting IGZO channel is deposited by thermal ALD at 225 °C, followed by the formation of 40-nm Ni source/drain contacts. For OS-FeFETs, a 400 °C annealing step in N_2 is applied after film deposition to enable the ferroelectricity. Channel isolation is achieved through dry etching, and postmetal annealing is conducted at temperatures ranging from 190 °C to 300 °C.

Fig. 3(a) presents the polarization–voltage characteristics of the ferroelectric–semiconductor–metal capacitor, indicating a high-quality ferroelectric gate-stack. Fig. 3(b) shows the standard transfer characteristics of an ALD IGZO-FeFET, demonstrating distinct memory states and a large memory window. The inset highlights the absence of hysteresis under ±1-V bias, confirming a nondestructive read operation. Fig. 3(c) illustrates the multilevel programming capability of the NVGC read-FeFET using programming voltages from 1.5 to 3 V, yielding eight clearly distinguishable states. Fig. 3(d) presents the retention characteristics of the GC mode of NVGC, with the access-FeFET in both low-Vth (LVT) and high-Vth (HVT) states, exhibiting stable charge retention without degradation up to 100 s. Logic circuits based on ALD InOx can be constructed using enhancement-mode (E-mode) and depletion-mode (D-mode) FETs. Fig. 3(e)–(g) shows the transfer and output characteristics of devices with a 5-μm channel length. The E-mode and D-mode FETs exhibit high mobilities of 76.9 and 67.7 $cm^2/V{\cdot}s$, respectively. Fig. 3(h) shows the successful single-bit programming of the NVGC access-FeFET without leakage or disturbance to the read-FeFET or adjacent cells. The inset illustrates array-level programmability using a modified V/2 scheme adapted from the 1T-1FeFET design [19], enabling reliable reverse programming from LVT to HVT.

III. 3-D LLM ACCELERATOR CIRCUIT DESIGN

Utilizing the proposed NVGC memory on a heterogeneous CMOS+OS platform in a monolithic 3-D integration scheme, we perform complete analog, digital, and mixed-signal 3D-integrated circuit (IC) implementation with full place-and-route (PnR) across the CMOS and interwoven OS tiers. Unlike many prior BEOL-based accelerator demonstrations that mainly focus on device characterization or isolated circuit blocks, this work emphasizes practical physical implementation considerations, including tier partitioning, mixed-voltage integration, routing congestion, intertier communication, and realistic area overheads. Such physical-design-level considerations are particularly important for evaluating the actual scalability and efficiency of BEOL-augmented 3-D AI accelerators.

First, the overall processing engine (PE) architecture is illustrated in Fig. 4(a). Each PE spans multiple vertically integrated tiers, including the NVGC memory tier, the OS-logic tier, and the Si CMOS compute tier. To balance performance, power, and routing complexity, only essential low-power peripheral circuits, such as row decoders, self-refresh controllers, and local sensing interfaces, are implemented on the OS-logic tier. This partitioning minimizes long-distance routing between the OS and CMOS domains while reducing the overall power overhead associated with peripheral operations. The analog front-end circuits including ADCs and DACs for near-memory processing are shared across multiple bit-lines through analog multiplexing structures to further improve area and energy efficiency. The main digital compute engines, scheduling logic, and control units remain on the CMOS tier to leverage the superior device performance and routing capability of advanced CMOS technology. Since the OS logic operates at a higher voltage domain, the dedicated level shifters are incorporated to interface the low-voltage CMOS logic with the 3.3-V OS circuits.

Next, we illustrate the operation modes and dataflows of the NVGC PE in Fig. 4(b)–(e). The proposed architecture supports four dynamically reconfigurable operating modes, including static FP/INT CIM using NV-FeFET storage, dynamic FP/INT CIM using GC-eDRAM operation, and an eDRAM buffer mode for storing intermediate

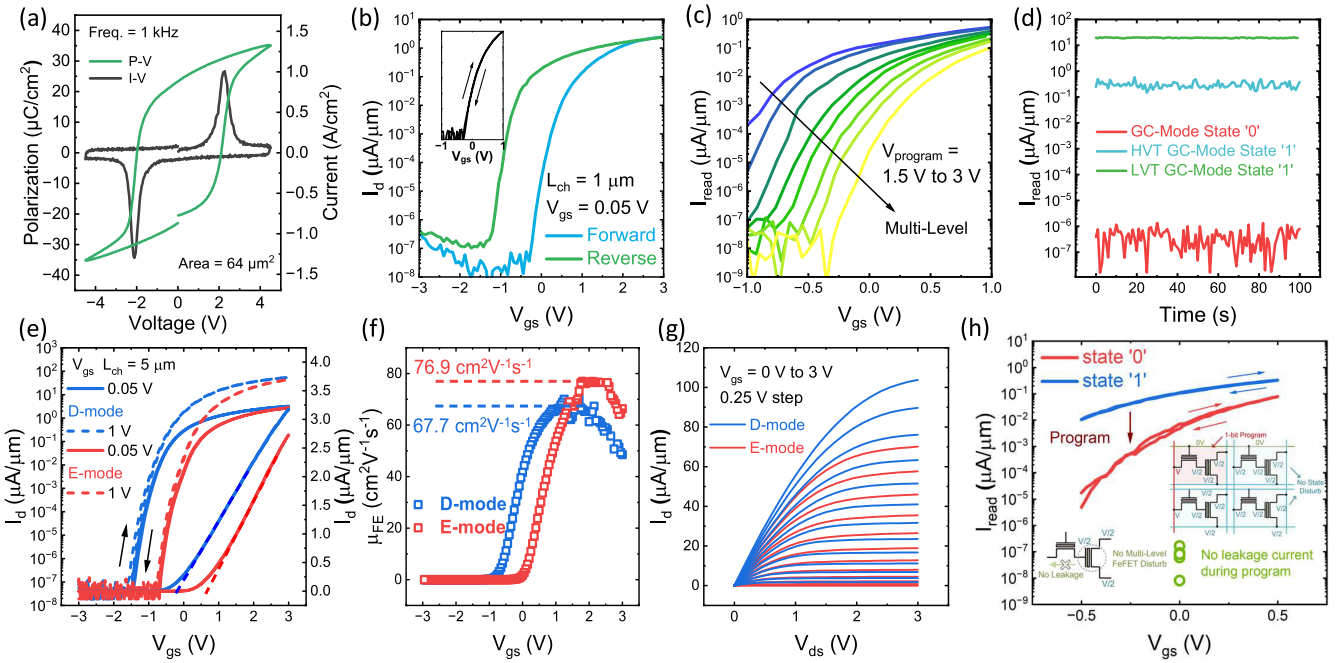


FIGURE 3. **(a) Polarization versus voltage characterization of ALD IGZO Fe-Cap. (b) IGZO FeFET transfer curves. Inset: no hysteresis below 1-V Vgs. (c) Multilevel characterization of OS-FeFET with eight distinct states. (d) GC DRAM mode charge-based retention characterization. (e) E-mode and D-mode InOx FETs transfer characteristics. (f) E-mode and D-mode InOx FETs field-effect mobility characteristics. (g) E-mode and D-mode InOx FETs output characteristics. (h) Programming single-level access FeFET without leakage or disturb. The measurement results are used for circuit simulations.**

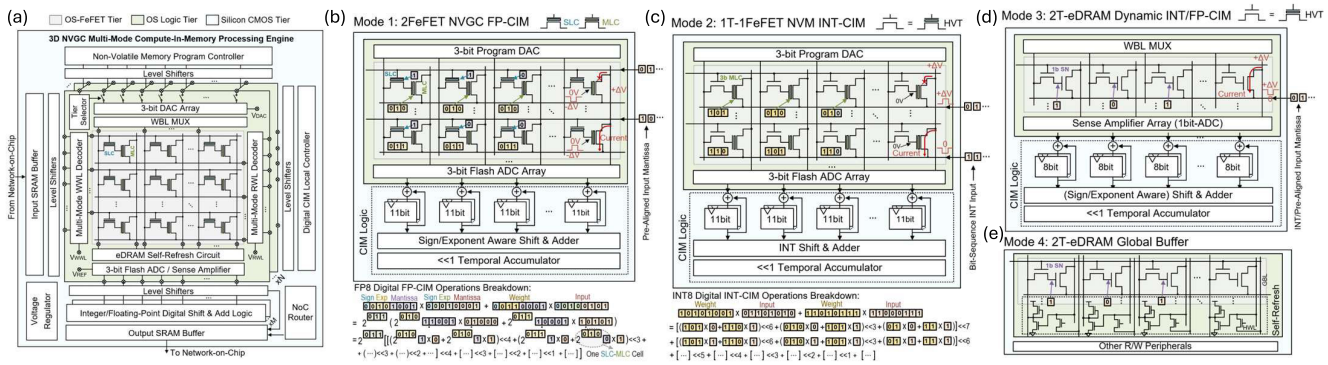


FIGURE 4. **(a) Floorplan and circuit design of 3-D NVGC CIM PE across OS-FeFET, OS-logic, and Si CMOS tiers. Operation modes of the NVGC PE (b) Mode 1, NV-FP mode: 2T-NVGC configured as a fused single-level multilevel 2-FeFET cell for FP CIM. (c) Mode 2, NV-INT mode: 2T-NVGC configured as 1T-1FeFET cell for INT CIM. (d) Mode 3, GC mode: 2T-NVGC configured as 2T-GC eDRAM cell for INT/FP CIM. (e) Mode 4, buffer mode: self-refreshing eDRAM global buffer.**

activations and temporary data during inference and training. This reconfigurable operation allows the same physical memory fabric to efficiently support workloads with different temporal characteristics and precision requirements. Static NV-mode operation is particularly suitable for storing long-term model weights and reusable parameters, while dynamic GC-mode operation enables efficient handling of activations, attention matrices, and temporary training states. Such flexibility is especially important for transformer-based workloads, where the ratio between static and dynamic data can vary significantly across models and sequence lengths.

Specifically, as shown in Fig. 4(b) capturing the "NV-FP mode," the 2T-NVGC is configured as a fused single-level/multilevel (SLC/MLC) 2-FeFET cell to support FP CIM. In this mode, the multilevel read-FeFET states encode exponent information while access-FeFETs store mantissa bits, enabling exponent-aware analog current accumulation without explicit mantissa prealignment. The accumulated outputs are digitized by the shared flash ADC array and processed by sign/exponent-aware shift-and-add logic together with temporal accumulation units. Fig. 4(c) shows the "NV-INT mode," where the same 2T-NVGC structure is

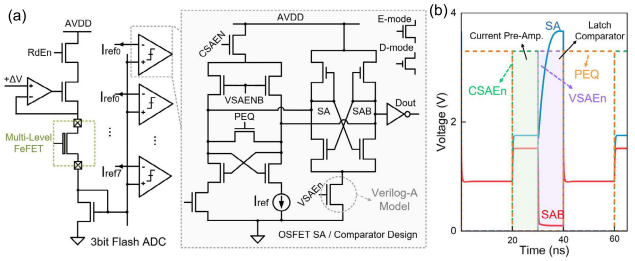


FIGURE 5. Analog circuit design with InOx transistors. (a) NMOS-only 3-bit flash ADC incorporating E-mode and D-mode OSFETs, with eight comparators that also function as sense amplifiers. (b) Simulated waveform of a two-stage current sense amplifier using a custom Verilog-A model with measured characteristics of fabricated OSFETs.

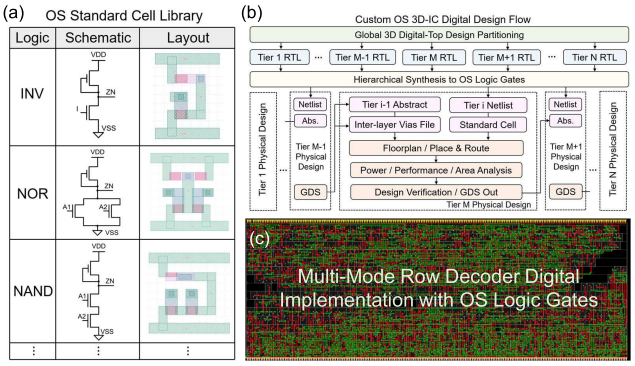


FIGURE 6. Digital circuit design with custom OS PDK. (a) Standard cell library based on InOx transistors. (b) Customized 3D-IC digital design flow. (c) Example result: multimode row decoder with OS gates.

reconfigured into a 1T-1FeFET style INT-CIM array for efficient INT inference with simplified bit-serial accumulation. Fig. 4(d) presents the "GC mode" for dynamic INT/FP-CIM operations, where the 2T-NVGC functions as a charge-based eDRAM compute fabric capable of handling dynamic activations, attention matrices, and temporary training states in the charge domain efficiently. In this mode, the flash ADC array is reconfigured into single-bit sense-amplifier arrays together with exponent-aware digital accumulation logic to support both INT and FP operations. Since dynamic weights and activations stored in the GC mode are frequently updated and rewritten, conventional weight prealignment introduces relatively small overhead compared with stationary NV-mode weights, making GC mode more suitable for flexible dynamic computation. Finally, Fig. 4(e) illustrates the buffer mode, where the 2T structure operates as a self-refreshing eDRAM global buffer with lightweight refresh circuitry for intermediate data storage and local communication across PE stages.

At the architectural level, these four modes allow the PE to dynamically adapt the same physical memory array to distinct computational roles, including long-term weight storage, high-utilization FP-CIM, conventional INT-CIM, dynamic activation processing, and temporary buffering. Such multimodal operation substantially improves memory utilization and efficiencies compared with fixed-function CIM architectures, while reducing both the demand for large on-chip SRAM buffers and the volume of intertier data movement.

To validate circuit functionality across multiple abstraction levels, all PE operations are verified at register-transfer-level (RTL) through mixed-signal simulations using experimentally calibrated behavioral models of the NVGC arrays. These simulations capture both digital control behavior and analog memory characteristics, including multilevel FeFET programming states, GC retention behavior, and analog sensing operations. Fig. 5(a) presents the analog implementation of a 3-bit flash ADC constructed using the E-mode and D-mode OSFETs. The comparator employs a two-stage architecture consisting of a current-mode preamplifier followed by a latch-based sense amplifier to improve sensing speed and robustness under OS device variations. The measured device I–V characteristics are used to extract Verilog-A models for accurate circuit-level simulation. The simulated waveform shown in Fig. 5(b) demonstrates high-speed rail-to-rail sensing with minimal offset and stable operation across multiple sensing cycles.

Fig. 6(a) further presents the custom OS standard cell library developed for digital synthesis and PnR on the OS-logic tier. Since conventional CMOS libraries are not directly compatible with nMOS-only OS logic, the library that we developed is specifically optimized for ALD InOx devices and incorporates experimentally calibrated transistor-level characteristics. The cell performance, power, and area (PPA) metrics are extracted through detailed transistor-level simulation and integrated into the digital design flow. Fig. 6(b) and (c) illustrates the overall 3D-IC design methodology and a representative multimode row decoder synthesized using the proposed OS-logic library, demonstrating the feasibility of large-scale digital implementation on the OS tier.

Leveraging the "sandwich 3-D" stacking scheme for the 3-D NVGC accelerator [Fig. 7(a)], our OS-on-CMOS physical implementation consists of three tightly integrated functional tiers that collaboratively distribute memory, mixed-signal processing, and control functionality across the vertical dimension. Unlike conventional CuA-based 3-D memory architectures, where vertically stacked layers mainly increase storage capacity while computation and peripherals remain confined to the silicon tier, the sandwich-3-D architecture allows each OS tier to participate directly in local data processing and orchestration. Such an organization is enabled by the unique capability of OS technologies to simultaneously support BEOL-compatible logic and memory devices within the same integration platform.

The top FeFET memory tier contains NVGC memory banks connected to the OS-logic tier through dense

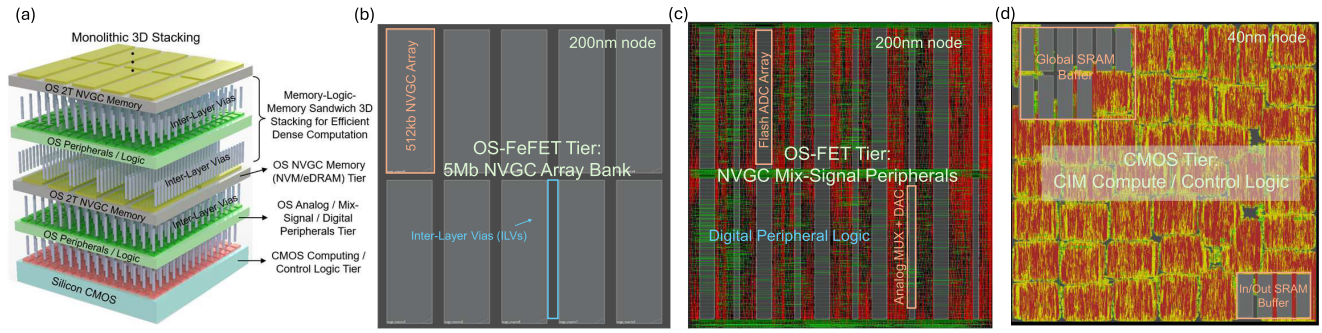


FIGURE 7. (a) Illustration of 3-D NVGC accelerator featuring OS-memory/OS-logic/OS-memory sandwich tiers on silicon CMOS. (b) OS-FeFET tier: 5-Mb NVGC array bank, composed of 512-kb subarrays, connected to the OS-logic tier through interlayer vias. (c) OS-FET tier: analog components (such as DACs and flash ADCs) with digital peripherals and essential functional logic. (d) CMOS tier: CIM auxiliary and control logic for upper tiers, system scheduler, global buffer, and input/output SRAM buffer.

interlayer vias [Fig. 7(b)], enabling high-bandwidth local communication between storage and mixed-signal processing circuits. The intermediate OS-logic tier integrates mixed-signal peripheral circuits, including row decoders, local controllers, sensing interfaces, ADCs/DACs, and lightweight digital logic [Fig. 7(c)], thereby offloading substantial local dataflow operations from the bottom CMOS tier. The bottom CMOS tier implements the main compute engines, scheduling logic, network-on-chip (NoC), and global control circuitry [Fig. 7(d)], leveraging silicon CMOS for large-scale digital orchestration.

This vertical partitioning transforms the stacked memory hierarchy from passive storage layers into computation-capable tiers with localized processing capability. As a result, the sandwich-3-D architecture simultaneously improves compute density, routing scalability, and memory utilization efficiency while reducing long-distance data movement and CuA congestion bottlenecks. Such findings suggest a scalable direction for future BEOL-integrated AI systems, where memory, mixed-signal peripherals, and lightweight logic are co-distributed across vertically stacked tiers rather than centralized in a single silicon layer.

The OS tiers are implemented using a relaxed 200-nm technology node while the CMOS tier utilizes a 40-nm foundry process, enabling realistic evaluation of lab-to-fab scaling implications for future BEOL-integrated AI accelerators. The vertical connectivity across tiers is not limited by this node mismatch, since the ILV count is set by the wiring pitch of the 200-nm OS tiers rather than by the 40-nm CMOS pitch. With 200 nm via diameter and 400 nm minimum pitch, a 512-kb subarray requires only about 1.85k ILVs between the memory and logic tiers, corresponding to a density more than three orders of magnitude below demonstrated monolithic-3-D capabilities, and each ILV contributes approximately 8 Ω and 0.3 fF, adding less than 0.5% to the bit-line load. Including landing pads and keep-out regions, ILVs occupy less than 0.5% of the memory tier and about 1%–2% of the logic tier, while on the CMOS side, they terminate on top-metal pads without consuming device-layer area. The complete design is implemented at the PE level, with inter-PE communication handled through an NoC infrastructure. For modeling simplicity, intra-PE communication across vertically integrated OS tiers is abstracted as NoC-level communication in the system-level evaluation framework. This abstraction allows scalable evaluation of large accelerator configurations while preserving the key effects of intertier communication overhead and vertical integration constraints.

IV. SYSTEM-LEVEL FEATURES AND INSIGHTS

To faithfully benchmark LLM inference and continual learning workloads at the system level, we develop a multitier 3-D modeling and design space exploration (DSE) framework [Fig. 8(a)], adapted from our recent work [23]. Unlike simplified analytical estimations that only consider isolated compute kernels, our framework captures interactions across multiple abstraction levels, including LLM model characteristics, accelerator architecture, circuit implementation, and device technology parameters. The DSE tool supports operation partitioning across different memory and compute resources, models PE-level kernel mapping and dataflow scheduling with inter/intratier communication overheads, and projects end-to-end power, performance, and area (PPA) metrics under realistic workload conditions. This framework further enables comprehensive system-technology co-optimization (STCO) analysis for the proposed 3-D LLM accelerator exploiting the reconfigurable NVGC memory and the "sandwich 3-D" architecture.

First, the reconfigurable NVGC memory enables on-demand optimization of energy and energy-delay-product (EDP) across LLM workloads with varying sequence lengths [Fig. 8(b)]. Unlike prior designs with fixed memory compositions, the proposed architecture dynamically adjusts the ratio between NV and GC operating modes according to workload behavior, thereby eliminating unnecessary FeFET rewrite overheads and OSFET refresh overheads. Since GC-mode operation is subcoercive [Fig. 3(b) inset], reconfiguration endurance is consumed

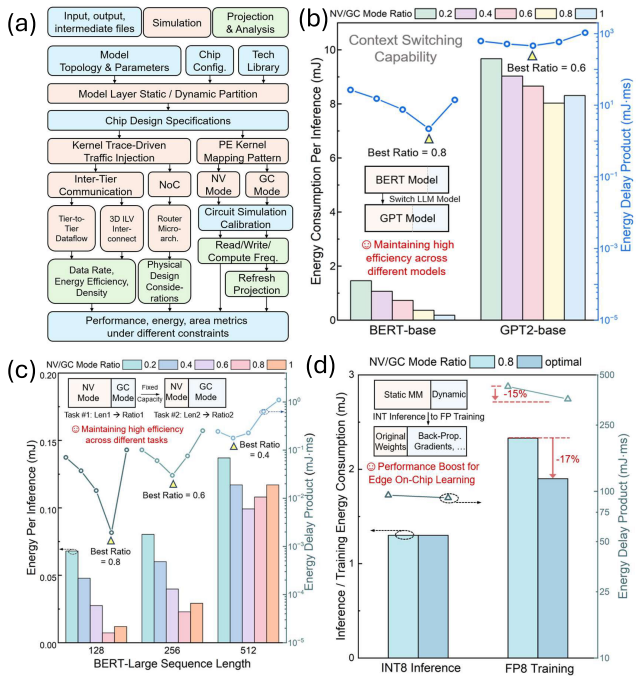


FIGURE 8. **(a) Full-stack system DSE and benchmarking framework. (b) BERT evaluation on 3-D NVGC with adaptive NV/GC mode partition ratios, revealing task-specific EDP-optimal configurations. (c) Efficient on-demand context switching across BERT/GPT-2 models via adapting NV/GC ratios, guided by system inference energy and EDP. (d) Reconfigurable NVGC PE reduces FP8 training energy by 17% for BERT, enabled by INT/FP adaptation with minimal overhead.**

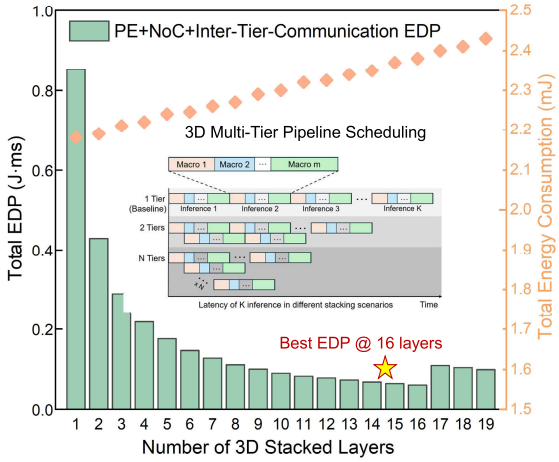


FIGURE 9. **3-D stacking design exploration considering CMOS resources, NoC, and intertier communication with cross-tier pipelining, showing best performance with 16 layers.**

only by infrequent NV reprogramming at context-switching boundaries, well within demonstrated OS-FeFET endurance. Longer sequence lengths tend to favor lower NV/GC ratios because they introduce a larger proportion of dynamic matrix operations associated with attention computation and activation updates. In contrast, the shorter sequence lengths benefit from higher NV ratios due to increased reuse of static model parameters. This workload-adaptive reconfiguration capability enables the accelerator to maintain near-optimal efficiency across a broad range of LLM operating conditions.

Second, the accelerator supports rapid context and task switching through dynamic reconfiguration of the NV/GC memory composition. As shown in Fig. 8(c), the optimal NV/GC ratio for minimizing energy consumption and EDP shifts significantly when the workload changes from BERT to GPT-2, reflecting fundamental differences in model architecture, dataflow characteristics, and temporal memory behavior. Such adaptability is particularly important for emerging edge and continual-learning applications, where heterogeneous AI tasks may execute sequentially on the same hardware platform. The proposed NVGC architecture therefore provides a flexible hardware substrate capable of efficiently supporting diverse transformer-based models without requiring dedicated accelerator redesign.

Third, the proposed system simultaneously supports energy-efficient INT-based inference and FP-based training workloads through the inherent flexibility of the NVGC memory architecture. As illustrated in Fig. 8(d), FP training workloads require a larger proportion of GC-mode memories to accommodate dynamic weight updates during backpropagation and gradient accumulation, while inference-oriented workloads benefit from larger NV-mode allocation for static weight storage. By enabling seamless transitions between these operating modes, the accelerator efficiently supports both inference and on-device training within a unified hardware framework, which is increasingly important for future adaptive and personalized AI systems. Finally, we implement pipelined inference with multitier scheduling on the proposed 3-D architecture while evaluating vertical scalability from an STCO perspective. The analysis considers tradeoffs among compute density, communication overhead, tier utilization, and CMOS area constraints across the OS-FeFET memory tier, OS-logic tier, and CMOS compute tier. As shown in Fig. 9, a pareto-optimal operating point achieving the best balance between EDP and compute density is obtained at approximately 16 vertically stacked layers. Beyond this point, additional layers introduce diminishing returns due to increasing communication overhead and limited CMOS resource scalability.

Thermal integrity of the 16-tier stack is evaluated with a conservative 1-D thermal-resistance network at peak compute utilization (Fig. 10). The dominant compute power is dissipated on the bottom CMOS tier next to the heat sink, while each OS tier carries only tens of mW/mm^2 of memory-access and peripheral power. Heat from upper tiers must traverse the ILD stack below, so the temperature rise accumulates toward the top tier; even so, the worst case top-tier rise is only 7.6 K above the CMOS tier, 3.4 K with ILVs serving as thermal vias, and below 15.2 K under a 2× OS-power overestimate. Such a rise degrades neither

TABLE 1. Comparison With Prior 2-D/3-D AI Hardware Designs Using OS-eDRAM or NVM. This Work Achieves Ultrahigh Compute Density and Supports LLM Context/Task Switching via Inherent Reconfigurability, Enabling Both Inference and Training Across INT and FP Formats.

Reference	IEDM'23 [17]	ISSCC'23 [20]	IEDM'24 [18]	ISSCC'24 [21]	VLSI'24 [22]	This Work
Cell Type	2T0C	3T1C	2T0C	1T1R	STT-MRAM	**2T NVGC**
Multi-Bits / Memory	1b / OS-eDRAM	1b / OS-eDRAM	1b / OS-eDRAM	1b / RRAM	1b / MRAM	**3b / OS-FeFET**
Memory Retention	Volatile	Volatile	Volatile	Non-Volatile	Non-Volatile	**Non-Volatile / Dynamic Reconfig.**
Data Format	INT	INT	INT	INT	INT/FP	**INT/FP**
ML Workload	CNN	Endpoint AI	Transformer	Robotics	CNN	**LLM**
Inference/Training	Inference	Inference	Inference	Inference	Inference	**Inference/Training**
3D Memory Bank	Yes	Yes	Yes	No	No	**Yes**
BEOL Peripherals	No	Yes	No	No	No	**Yes**
3D-Layer Logic	No	No	No	No	No	**Yes**
System-Technology Co-Optiomized	No	No	No	No	No	**Yes**
Compute Density (TOPS/mm^2) [1–3]	50	NR [4]	0.007	0.013	0.018	**102.4**
Energy Efficiency (TOPS/W) [1–3]	256	4.44	0.099	1.14 [5]	63.9	**32.9**

[1] 1 MAC = 2 OP [2] 1.1V, 200MHz working condition [3] Normalized to 8-bit operations. [4] Not reported. [5] w/ SRAM CIM.

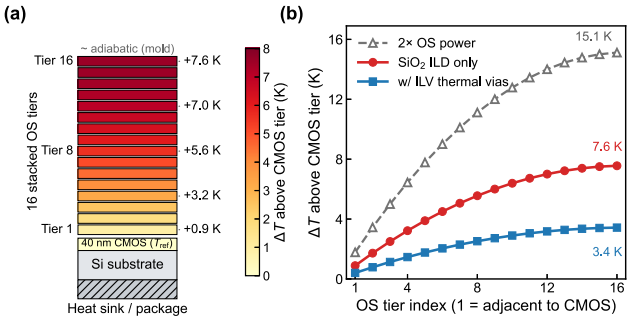


FIGURE 10. Thermal-resistance-network analysis of the 16-tier stack at peak utilization. (a) Steady-state temperature profile. (b) Per-tier rise above the CMOS tier: at most 7.6 K (SiO_2; ILD), 3.4 K with ILV thermal vias, and less than 15.2 K under a 2× OS-power sensitivity check.

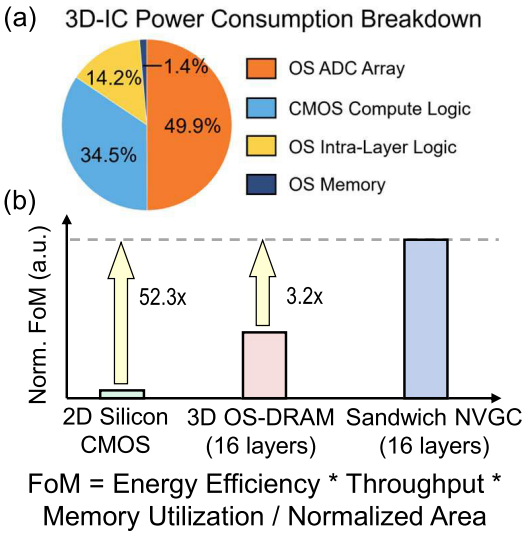


FIGURE 11. (a) 3D-IC cross-tier power breakdown. (b) FoM comparison among 2-D silicon CMOS, 3-D OS-DRAM, and this work.

the thermally activated OSFET mobility nor the multilevel retention margin of the read-FeFETs.

With the figure-of-merit (FoM) defined in Fig. 11, the proposed 3-D NVGC accelerator achieves a 52.3× improvement in holistic system efficiency over planar CMOS-based counterparts and a 3.2× improvement over 3-D DRAM-based solutions for LLM hardware acceleration. Table 1 further summarizes the comparisons with the state-of-the-art emerging memory-based AI accelerators. Compared with prior works primarily targeting fixed-function INT inference with either volatile or NV memories, the proposed 3-D NVGC architecture uniquely combines dynamic/NV reconfigurability, native FP/INT CIM capability, BEOL-integrated peripherals and logic, and full-stack STCO within a monolithic 3-D implementation. These features enable efficient support for LLM inference, training, and context-switching workloads while simultaneously achieving ultrahigh compute density and practical mixed-signal physical design scalability for future 3-D AI accelerators.

V. CONCLUSION

We presented a device-circuit-system co-designed 3-D LLM accelerator based on OS-logic and NVGC memories. The key contributions of this work are threefold: 1) a new reconfigurable 2T-NVGC memory with native FP compute capability, enabling seamless transitions between dynamic

and NV operating modes; 2) a new "sandwich 3-D" BEOL stacking scheme for scalable and high-density 3-D CIM architectures; and 3) the first full-PnR monolithic 3-D LLM accelerator that captures the practical lab-to-fab implications of OS-logic and memory technologies. Through device characterizations for experimental calibration, mixed-signal circuit implementation, and full-stack system-level evaluation, the co-designed solution efficiently supports both INT and FP LLM workloads across inference and training. Together, this work establishes a scalable, LLM-native technology platform that can co-evolve and scale with future 3-D-ICs.

REFERENCES

[1] T.-H. Wen et al., "A 22nm 16Mb floating-point ReRAM compute-in-memory macro with 31.2TFLOPS/W for AI edge devices," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2024, pp. 580–582.

[2] L. Zheng et al., "CENTAUR: A 38.5-TFLOPS/W 600MHz floating-point digital compute-in-memory engine with 40nm fusion RRAM-eDRAM macros featuring 3D-MAC operation," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2025, pp. 139–141.

[3] L. Zheng et al., "PROTEUS: A 40 nm programmable general-purpose digital compute-in-memory accelerator with eNVM and hierarchical ISA for versatile edge AI," *IEEE J. Solid-State Circuits*, vol. 61, no. 9, pp. 5125–5141, Sep. 2026.

[4] L. Zheng, Z. Wang, S. Du, M. Chen, A. M. Bavani, and H. Li, "Heterogeneous compute-in-memory fabrics for efficient, scalable edge inference and learning," in *Proc. 27th Int. Symp. Quality Electron. Design (ISQED)*, Apr. 2026, pp. 1–6.

[5] C. Niu et al., "Monolithic 3D integration of atomic-layer-deposited oxide semiconductors on 200-mm silicon wafers," *Nature Nanotechnol.*, Sep. 2026. [Online]. Available: https://www.nature.com/articles/s41565-026-02276-0#citeashttps://www.nature.com/articles/s41565-026-02276-0#citeas

[6] L. Zheng and H. Li, "CMOS+X technologies for neuro-vector-symbolic computing," in *Proc. Device Res. Conf. (DRC)*, Jun. 2024, pp. 1–2.

[7] J.-G. Lee et al., "Analog computation in ultra-high density 3D FeNAND for TB-level hyperscale AI models," in *IEDM Tech. Dig.*, Dec. 2024, pp. 1–4.

[8] K. S. Choi et al., "A three dimensional DRAM (3D DRAM) technology for the next decades," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Jun. 2024, pp. 1–2.

[9] M. Si, Z. Lin, Z. Chen, X. Sun, H. Wang, and P. D. Ye, "Scaled indium oxide transistors fabricated using atomic layer deposition," *Nature Electron.*, vol. 5, no. 3, pp. 164–170, Feb. 2022.

[10] C. Niu et al., "First demonstration of BEOL wafer-scale all-ALD channel CFETs using IGZO and te for monolithic 3D integration," in *IEDM Tech. Dig.*, Dec. 2024, pp. 1–4.

[11] Z. Lin, M. Si, and P. D. Ye, "Ultra-fast operation of BEOL-compatible atomic-layer-deposited In_2O_3 Fe-FETs: Achieving memory performance enhancement with memory window of 2.5 V and high endurance > 109 cycles without V_T drift penalty," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 1–2.

[12] Z. Lin et al., "High-peformance BEOL-compatible atomic-layer-deposited In_2O_3 Fe-FETs enabled by channel length scaling down to 7 nm: Achieving performance enhancement with large memory window of 2.2 V, long retention > 10 years and high endurance > 108 cycles," in *IEDM Tech. Dig.*, Dec. 2021, pp. 17.4.1–17.4.4.

[13] S. Datta et al., "Amorphous oxide semiconductors for monolithic 3D integrated circuits," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2024, pp. 1–2.

[14] A. Belmonte and G. S. Kar, "Disrupting the DRAM roadmap with capacitor-less IGZO-DRAM technology," *Nature Rev. Electr. Eng.*, vol. 2, no. 4, pp. 220–221, Mar. 2025.

[15] H. Mulaosmanovic et al., "Ferroelectric field-effect transistors based on HfO_2: A review," *Nanotechnol.*, vol. 32, no. 50, 2021, Art. no. 502002.

[16] Z. Huo, W. Cheng, and S. Yang, "Unleash scaling potential of 3D NAND with innovative Xtacking® architecture," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 254–255.

[17] W. Tang et al., "30 Mb/mm^2/layer 3D eDRAM computing-in-memory with embedded BEOL peripherals and local layer-wise calibration based on first-demonstrated vertically-stacked CAA-IGZO 4F2 2T0C cell," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4.

[18] S. Liu et al., "Edge continual training and inference with RRAM-gain cell memory integrated on Si CMOS," in *IEDM Tech. Dig.*, Dec. 2024, pp. 1–4.

[19] S. G. Kirtania et al., "Amorphous indium oxide channel FeFETs with write voltage of 0.9 V and endurance > 10^{12} for refresh-free 1T-1FeFET embedded memory," in *IEDM Tech. Dig.*, Dec. 2024, pp. 1–4.

[20] 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.

[21] S. D. Spetalnick et al., "30.1 A 40nm VLIW edge accelerator with 5MB of 0.256pJ/b RRAM and a localization solver for bristle robot surveillance," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2024, pp. 482–484.

[22] D.-Q. You et al., "A 22nm nonvolatile AI-edge processor with 21.4TFLOPS/W using 47.25Mb lossless-compressed-computing STT-MRAM near-memory-compute macro," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2024, pp. 1–2.

[23] S. Du et al., "3D-CIMlet: A chiplet co-design framework for heterogeneous in-memory acceleration of edge LLM inference and continual learning," in *Proc. 62nd ACM/IEEE Design Autom. Conf. (DAC)*, Jun. 2025, pp. 1–7.