Energy-Efficient, High-Density, and High-Performance IGZO eDRAM in Advanced Technology for AI Memory

*A. Sharma[1], *C. Apte[3], D. Matsubayashi[1], L. M. Giacomini Rocha[1], H. Oh[1], K. English[3], R. Schreiber[4], J. Wuu[5], S. Subhechha[1], H. Watson[2], L. Verschueren[1], M. Garcia Bardon[1], A. Belmonte[1], G. S. Kar[1], G. Hellings[1], D. Biswas[1], F. Garcia-Redondo[2]
[1]*IMEC*, Leuven, Belgium, [2]*IMEC*, Cambridge, UK, [3]*AMD*, Boxborough, MA, USA,
[4]*AMD*, Austin, TX, USA, [5]*AMD*, Fort Collins, CO, USA
[*]Authors contributed equally to the work. mail: arvind.sharma@imec.be, ORCID 0000-0003-1188-4924

***Abstract*—We present energy-efficient, high-density, high-performance memory macros based on 2T0C and 3T0C IGZO eDRAM bitcells, implemented using an in-house 7 nm node (N7). IGZO TFTs are optimized for ultra-low leakage, achieving ~ 430 s / ~ 4 s retention at $27°$C / $85°$C. Bitcell architectures are designed to minimize WL-to-storage node coupling, and full-swing and differential sensing schemes are employed to optimize energy-performance-area tradeoffs. The macros integrate dense BEOL arrays with FEOL periphery, achieving A10 SRAM-comparable density. 32 KiB configurations optimized for high-performance (energy-efficiency) achieve 1.27 (4.17) ns access time at $85°$C and 49 (40) Mb/mm² density.**

***Index Terms*—IGZO, eDRAM, 2T0C, 3T0C, DTCO**

I. Introduction

SRAM has reliably supported on-chip memory for decades, but the explosive growth of AI workloads has exposed fundamental limits in capacity, bandwidth, energy scalability, and cost. SRAM capacity has increased by $100\times$ in three years, while SRAM bitcell area has only shrunk incrementally [1] (Fig. 1). Emerging memories are therefore gaining traction, offering different trade-offs in density, power, and retention. For instance, STT-MRAM improves density but suffers from high energy/latency [3]. SOT-MRAM, while faster, incurs into additional area penalties [4]. BEOL-compatible oxide-semiconductor (OS) transistors have been demonstrated [5], [6], yet integrating them in 1T1C cells still demands extreme aspect-ratio capacitors. In contrast, capacitor-less OS-based eDRAM combines high density, ultra-low leakage, BEOL integration, and energy-efficient operation without capacitor complexity [7]–[9].

This work presents a comprehensive optimization framework from device to macro-level memory deployment. We first present the optimized IGZO thin-film transistors (TFTs), engineered for ultra-low leakage at elevated temperatures. These devices are used to implement 2T0C and 3T0C bitcells, achieving retention times of $430(4)$ s at $27(85)°$C. We pursue column construction to demonstrate function and project to macro PPA with N7 FEOL periphery. Differential and a hierarchical array full-swing sensing schemes are employed to optimize energy-efficiency and density, and performance profiles, with access times of 4.17 ns and 1.27 ns, respectively, and SRAM-comparable density to A10 and N2 nodes.

Work enabled by NanoIC pilot line (nanoic-project.eu), jointly funded by the Chips Joint Undertaking, through the EU's Digital Europe (101183266) and Horizon Europe programs (101183277).

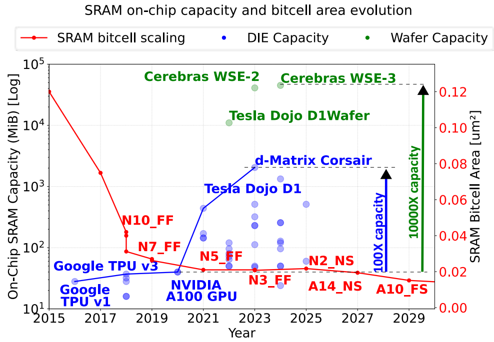
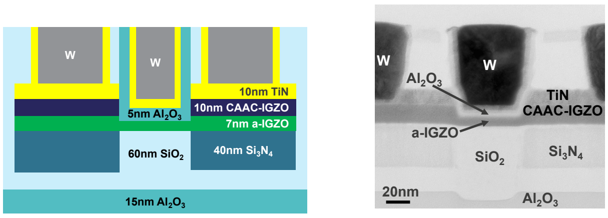


Fig. 1: On-chip memory demand is driven by AI accelerators, with $100\times$ / $10000\times$ die / wafer demand increase in just 4 years, while SRAM scaling becomes costlier [1], [2].

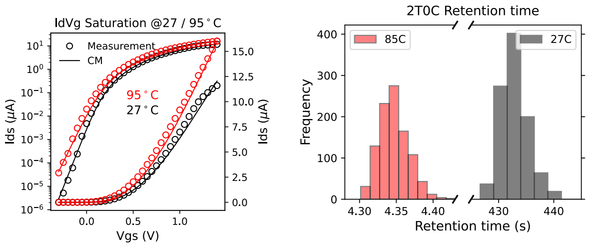


(a) (b)

Fig. 2: Gate-last IGZO-TFT (a) schematic and (b) cross-sectional TEM. The $L_g = 40\,$nm, $W = 210\,$nm fabricated devices were optimized for retention at $95°$C [10].

II. IGZO Devices and Bitcells

IGZO-TFTs optimized for retention with atomic layer etching as the active patterning technique were



(a) CM calibration (b) Retention time

Fig. 3: (a) CM calibration to fabricated devices ($Lg = 40nm$, $W = 210nm$) at 27°C and 95°C (b) Retention at 27°C/85°C for 2T0C.

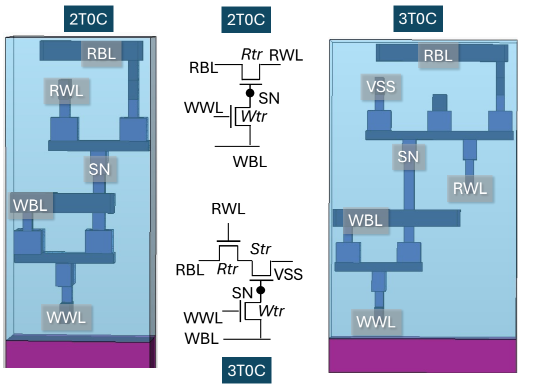


Fig. 4: 2T0C/3T0C schematic and corresponding 3D view [12].

in-house manufactured [10]. The device schematic and TEM image are shown in Fig. 2. Process details for the gate-last architecture, including the oxygen tunnel and CAAC-IGZO raised contacts, are followed from [11]. Post-processing O_2 ambient anneal was applied to optimize threshold voltage (V_{th}), and on/off-currents (Ion, Ioff). Devices with a gate length of 40 nm were characterized at 27°C and 95°C. As described in Fig. 3(a), a BSIM-based compact model (CM), used for the experiments described in Section III, was then calibrated to the median I–V curves.

A. Bitcell Architectures

Fig. 4 presents the stacked implementation of 2T0C and 3T0C bit cells using IGZO devices. Compared to 3T0C, the 2T0C cell benefits from reduced area and parasitics due to the absence of a dedicated read transistor. However, large-array's performance is limited by read delays induced by disturb currents on RBL and direct coupling between storage node (SN) and RWL [7]. In contrast, 3T0C employs an isolated read path, enhancing read latency in large arrays. Table I summarizes BEOL layers, routing CDs/pitches, and operating voltages, following in-house N7 BEOL rules and IGZO CPP/active-pitch constraints. Given an IGZO $V_{th} \sim 0.3$ V, the WWL ON voltage is set to 1.35 V. Based on IGZO devices and N7 BEOL stacks, 3D cell layouts were constructed [12] and parasitic components extracted using Synopsys Raphael FX [13]. The resulting bitcell areas are 0.015 µm² for 2T0C and 0.020 µm² for 3T0C.

TABLE I: Layer Pitches and Bitcell Node Voltages

Layer	CD (nm)	Pitch (nm)	Pitch 3T0C (nm)
Gate	40	120	120
WWL	25	180	240
WBL	25	85	85
RWL	25	180	240
RBL	25	85	85

Wite Node	Voltage	Read Node	Voltage
WWL	−1 / 1.35 V	RWL	0 / 1 V
WBL	0 / 1 V	RBL	0 / 1 V

B. Data Retention

We define retention as the time till 100 mV drop (increase) at storage node for data '1' ('0'). Both cells exhibit retention exceeding 400 s at 27°C. Due to the higher temperature sensitivity of IGZO, increased I_{off} at elevated temperature degrades retention relative to Si devices; however, optimized low-I_{off} devices achieve a mean retention of $\sim$ 4 s at 85°C, representing a $\sim 100\times$ improvement over prior work [7]. Fig. 3(b) shows the retention distributions from 1000 Monte Carlo samples for the 2T0C cell at 27°C and 85°C, with similar behavior observed for 3T0C. In [10], 200 s retention at 95°C is reported using $V_{\text{WWL}} = -1.4$ V and $W = 210$ nm, whereas this work employs $W = 40$ nm and a FEOL-compatible $V_{\text{WWL}} = -1$ V.

III. Optimized Subarrays and Macros Design

Fig. 5(a) shows the architecture of the proposed 2T0C/3T0C IGZO-based eDRAM bit array and its periphery. IGZO devices are monolithically integrated in the BEOL stack, while periphery and control logic are implemented in N7 CMOS, to integration efficiency. Arrays interface with periphery blocks at the boundaries, with WWL, WBL, RWL, RBL, and VSS (for 3T0C) routed through four metal layers (Fig. 4).

Both differential and full-swing sensing schemes have been implemented for comparative studies. Fig. 5(b) and 5(c) describe the energy efficient, differential sensing circuitry, and the hierarchical, full-swing sensing circuitry, respectively. In differential mode, a reference level is generated via WBL underdrive to balance '0'/'1' delays, with each array including a periodically refreshed reference row. A 130 mV differential voltage ensures robust reads. However, differential sensing is limited by low Ion and variability in IGZO read ports, making sub-ns read access challenging.

To address this, a full-swing single-ended sensing scheme with hierarchical bitlines is employed to reduce effective bitline capacitance and parasitics. This reduction is critical for IGZO's limited drive strength and enables sub-ns access even with single-ended sensing, at the cost of higher energy and reduced density.

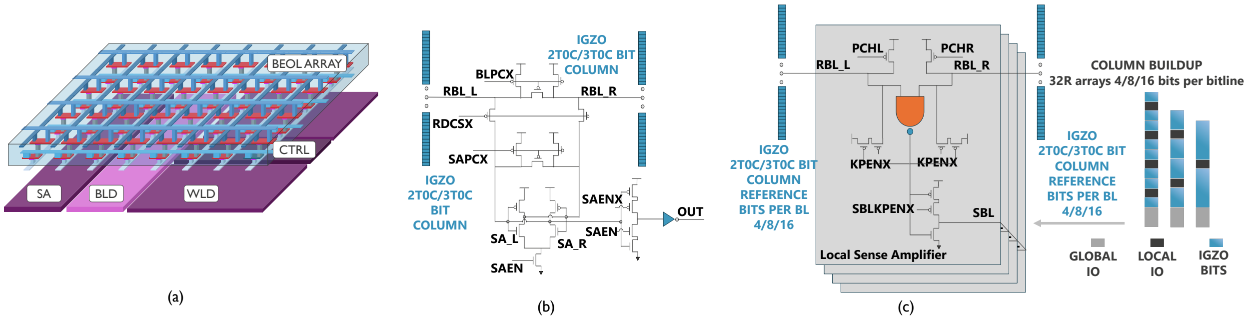


Fig. 5: (a) Organization of the BEOL array memory and the periphery underneath, including control, write drivers, bitline drivers and sense amplifiers. (b) Energy efficient differential and (c) Hierarchical full-swing approach for high-performance at the cost of extra periphery (multiplexing, sensing, etc.). The hierarchical multiplexing/sensing logic is arranged underneath the memory array and deeper multiplexing schemes may incur area penalties.

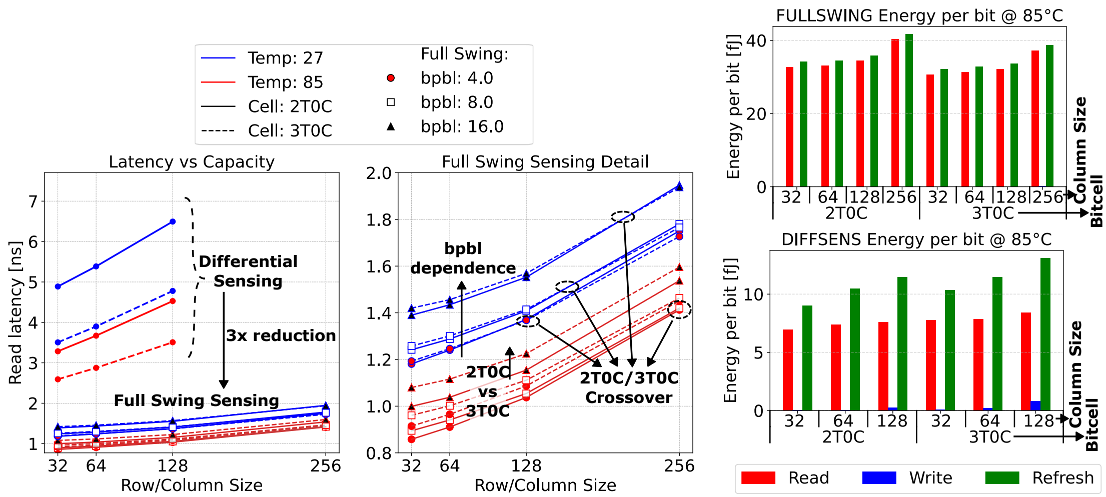


Fig. 6: (a) Differential vs full swing sensing latency analysis for different temperatures, bitcell configurations and number of elements in columns (column size). (b) Energy per bit across experiments. Read operation dominates energy consumption. Full-swing requires $3.5\times$ higher energy/op compared to differential sensing.

Local sense amplifiers are shared across a reduced bits-per-bitline (*bpbl*), evaluating data locally before driving short SBL segments and forwarding to the muxed I/O. Although additional periphery is required, BEOL array implementation allows the circuitry to be placed beneath the array, minimizing area overhead (Fig. 5).

A. Array Performance and Energy Analysis

Fig. 6 compares read latency for both schemes across temperature and subarray size. Circuit evaluations were simulated using Cadence Spectre [14], where sensitivity trends are appreciated. In differential mode, 3T0C offers 50% faster reads than 2T0C due to negligible disturb current on RBL from unselected cells. While latency improves with temperature, it degrades with larger subarrays sizes. Full-swing sensing present $2/3\times$ performance improvements over standard differential sensing. Here, smaller bpbl achieve better delays at the expense of higher areas. Regarding 2T0C vs 3T0C performance, in full-swing sensing, this difference is less noticeable, with 2T0C being slightly faster than 3T0C on smaller arrays, with a 3T0C crossover at larger capacities. This fast scheme maintains $<$ 1.5 ns access even for 2 KiB banks, supporting high-throughput.

Regarding energy consumption, Fig. 6 shows that energy per bit is dominated by read and pre-charge phases. Full-swing sensing consumes more energy ($\sim 3.5\times$) than differential due to full swing on BLs. Even though temperature has minimal effect, energy scales with subarray size.

B. Macro Power, Performance and Area Metrics

Column-level measurements are extrapolated to estimate macro-level metrics. Array efficiency—defined as the ratio of bitcell area to total macro area—is $0.84\times$ for high-density (HD) macros [15] and $0.75\times$ for high-performance (HP) macros, providing a basis for macro density estimation. With advanced CMOS periphery, macro read/write latency is dominated by nanosecond-scale column latency. Subarrays are optimized for energy efficiency (differential sensing) or high performance (full-swing single-ended sensing). A 32 KiB macro is stitched via a two-layer H-tree from 16 128×128 subarrays. Table II compares the array

TABLE II: Comparison of different L3 memory candidates based on 256x256 *row* x *column* size. Further, a comparison of 32 KiB macro implemented with IGZO 2T0C/3T0C and SRAM

	[1]	[3]	[8]	[9]	This work
Technology	A10 SRAM	N7 STT-MRAM	[1]IWO 2T0C	N7 IWO 2T0C	N7 IGZO 2T0C
Bitcell area (μm^2)	0.015	0.0158	0.032	0.021	0.0153
Retention time (s)	NA	70@85°C	1@85°C	0.32@27°C	4.3@85°C
Read time (ns)	0.1-0.2	1.5-2.5	> 3	< 1	1.5-2.5
Write time (ns)	< 0.3	5-10	> 3	< 0.5	< 0.5
Bitcell	FEOL	BEOL+FEOL	BEOL	BEOL	BEOL

[1]Considering single eDRAM layer

Macro Params	Energy-efficient	High-performance	High-density
Bitcell	N7 3T0C	N7 2T0C	A10 SRAM [1]
Sensing scheme	Differential	Full swing	Differential
Subarray size (bits)	128×128	128×128	256×256
#Subarrays	16	16	4
Read delay (ns) at 27°C / 85°C	5.7 / 4.17	1.68 / 1.27	0.26 / 0.25
Write delay (ns) at 27°C / 85°C	0.73 / 0.57	0.50 / 0.40	0.28 / 0.29
Read energy (fJ/bit) at 27°C / 85°C	15.11 / 15.63	39.77 / 37.97	8.10 / 8.64
Write energy (fJ/bit) at 27°C / 85°C	7.96 / 8.01	6.41 / 6.42	29.34 / 31.58
Refresh energy (fJ/bit) at 27°C / 85°C	26.79 / 27.56	47.41 / 45.65	NA
Leakage power (μW) at 27°C / 85°C	5.2 / 39	8.84 / 62.4	158 / 538
Retention (s) at 27°C / 85°C	434 / 4.3	434 / 4.3	NA
Bitcell area (μm^2)	0.020	0.0153	0.0151 [1]
Macro area efficiency	0.84	0.75	0.84 [15]
Density (Mb/mm^2)	40	49	55.6

specifications of the proposed work against literature, and summarizes energy-efficient (EE) and HP macro specifications; refresh energy includes read, write, and precharge.

For differential sensing approach based IGZO and A10 SRAM macros area efficiency of 0.84 is used [15]. High-performance 2T0C macro uses 128×128 subarrays with 8 bpbl, and requires larger periphery than differential approach–Fig. 5(c). With a macro area efficiency of 0.75–where peripherals can only be partially placed under the BEOL arrays, 49 Mb/mm2 density is achieved. Density can be further improved by optimizing under-array periphery placement, achieving densities closer to A10 SRAM. The proposed HP and EE macros match A10 and N2 SRAM density [15], respectively, while using a five generations older technology. This projects to $\sim 2.5\times$ improvement over N7 SRAM efficiency applying a 20% energy reduction per node [15].

IV. Conclusions

This work introduces an IGZO-based eDRAM scaling path for AI memory beyond SRAM limits. Ultra-low-leakage IGZO TFTs enable 2T0C/3T0C bitcells with 430 s/4 s retention at $27°$C/$85°$C. Differential and hierarchical full-swing sensing optimize energy and speed. Monolithic BEOL integration with N7 FEOL periphery achieves $40/49$ Mb/mm^2 density and $4.17/1.27$ ns access for energy-efficient/high-performance macros, matching A10/N2 SRAM density [1], [15].

V. Acknowledgment

The authors would like to acknowledge Dawit Abdi and Moritz Brunion for their help.

References

[1] D. Abdi *et al.*, "SRAM Scaling Opportunities Below 0.01 um² Using Double-Row CFET Architecture with Wordline-Folded Bitcell Design for Performance Optimization," in *IEEE Symp. on VLSI Tech. and Circ.,* pp. 1–3, 2025.

[2] A. Sharma *et al.*, "Analysis and Mitigation for Row Hammer in IGZO based eDRAM at Advanced FinFET Node," in *IEEE International Memory Workshop (IMW)*, 2026.

[3] D. Narducci *et al.*, "High Density, High Speed STT-MRAM N7 Macros: Material and DTCO Exploration," in *IEEE Symp. on VLSI Tech. and Circ.*, pp. 1–3, 2025.

[4] Y. Xiang *et al.*, "SOT-MRAM Bitcell Scaling With BEOL Read Selectors: A DTCO Study," in *IEEE TED,* pp. 1–7, 2025.

[5] K. H. Chiang *et al.*, "Integration of 0.75V VDD Oxide-Semiconductor 1T1C Memory with Advanced Logic for an Ultra-Low-Power Low-Latency Cache Solution," in *IEEE Symp. on VLSI Tech. and Circ.*, pp. 1–3, 2025.

[6] M. Kim *et al.*, "High Performance and Reliable 4F2 IGZO Vertical Channel Transistor (VCT) with Extremely Low Contact Resistance and 10 Year BTI lifetime for Sub-10nm DRAM," in *IEEE Symp. on VLSI Tech. and Circ.,* pp. 1–3, 2025.

[7] A. Sharma *et al.*, "IGZO based eDRAM: Bitcell and Array Optimization Enabling Denser Last Level Caches," in *IEEE ESSERC,* pp. 1–4, 2025.

[8] H. Ye *et al.*, "Double-gate W-Doped Amorphous Indium Oxide Transistors for Monolithic 3D Capacitorless Gain Cell eDRAM," in *IEEE IEDM*, pp. 28.3.1–28.3.4, 2020.

[9] F. Waqar *et al.*, "Optimization and Benchmarking of Monolithically Stackable Gain Cell Memory for Last-Level Cache," in *IEEE Trans. on Comp.*, vol. 75, no. 3, pp. 760–775, 2026.

[10] K. Izukashi *et al.*, "Pathways for Retention Boost in Atomic Layer Etched IGZO-based Capacitorless DRAM," in *IEEE Electron Device Letters,* 2025.

[11] S. Subhechha *et al.*, "Ultra-Low Leakage IGZO-TFTs with Raised Source/Drain for Vt $>$ 0 V and Ion $>$ 30 µA/µm," in *IEEE Symp. on VLSI Tech. and Circ.,* pp. 292–293, 2022.

[12] Sentaurus Structure Editor – Synopsys Inc., CA, USA.

[13] Raphael FX - Synopsys Inc., Mountain View, CA, USA.

[14] Cadence® spectre® circuit simulator/virtuoso® layout suite.

[15] T. -Y. J. Chang *et al.*, "A 38.1 Mb/mm 2 SRAM in a 2nm-CMOS-Nanosheet Technology for High-Density and Energy-Efficient Compute," in *IEEE ISSCC,* pp. 492–494, 2025.