IGZO based eDRAM: Bitcell and Array Optimization Enabling Denser Last Level Caches

A. Sharma[1], H. Oh[1], L. Verschueren[1], S. Subhechha[1], N. Rassoul[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 mail: arvind.sharma@imec.be, ORCID 0000-0003-1188-4924

***Abstract*—While SRAM density scaling saturates, the need for large on-chip memory capacities keeps increasing, calling for alternative memory options compatible with CMOS. In this paper we propose innovative, fully BEOL-compatible 2T0C and 3T0C IGZO-based memory bitcells featuring front- and back-gated write transistor contact configurations. The 3T0C architecture is explored for its scalability benefits in high-capacity arrays, while the use of back-gate contacts enhances storage node isolation through improved coupling shielding. We evaluate the power-performance-area trade-offs across multiple FinFET technology nodes and array sizes, demonstrating sub-2ns / sub-0.3ns access latency on 8 KiB N7 arrays at $27°$C / $85°$C, and ultra-compact $0.015\ \mu m^2$ bitcells—achieving metrics comparable to A10 SRAM, hence achieving denser alternatives.**

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

I. Introduction

As SRAM area scaling staggers on advanced nodes, its becomes increasingly slower and costly [1]. Therefore, there is growing interest in embedded memory solutions that offer high capacity at lower cost (Fig. 1). BEOL-compatible IGZO memories are promising, as they decouple memory fabrication from advanced logic, enabling dense and cost-effective integration. These benefits align well with the emerging "bonding era," where modular compute-memory integration is key. While emerging alternatives like STT-MRAM [2] offer high density, its write latency and energy still fall short of SRAM-class performance. In contrast, IGZO-based dynamic memories provide faster access while preserving density.

This paper presents a roadmap for IGZO-based eDRAM using in-house fabricated devices, focusing on design tradeoffs across FinFET nodes. We propose novel 2T0C and 3T0C IGZO bitcell architectures with both front- and back-gated (FG/BG) write transistor contact configurations. These innovations enhance performance by improving access control and shielding of the storage node. Notably, the 3T0C design while incurring a modest area overhead, decouples the read and write paths, enabling larger numbers of elements per bitline and thus high-capacity arrays. The proposed BG-process provides reduced area solutions, where a 3T0C-BG becomes denser than a traditional 2T0C-FG. Our optimized N7 implementations achieves sub-$0.3ns$ access latency in high performance $0.02\ \mu m^2$ cells, complemented by compact $0.015\ \mu m^2$ high density bitcells, demonstrating the feasibility of IGZO eDRAM as a scalable, energy-efficient and cheaper alternative to SRAM solutions.

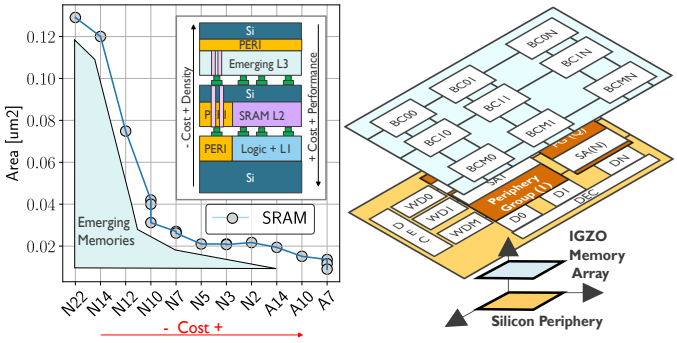

Fig. 1. Potential scenario for a high capacity, cost effective, hybrid cache hierarchy. Hybrid bonding construction with high-performance CFET-based L1 [1], dense Nanosheet L2, and cheaper high capacity emerging memory L3 in a cost-effective FinFET node. IGZO-based eDRAM memory array sitting above silicon periphery, minimizing area.

II. IGZO Device and Methodology

Our methodology (Fig. 2) begins by measuring in-house devices and calibrating compact models (CMs). For different nodes, BEOL layouts and 3D TCAD structures are generated [3] to extract parasitics [4]. Arrays are constructed, and bitcell architecture and array features are co-optimized to achieve maximum performance with minimal area after standard circuit analyses [5], with $V_{WWL} = 1.35/-1V$ and $V_{DD} = 1V$. In all experiments, the read delay is determined by the time from *WL* rise till the *RBL* develops a $130mV$ sensing margin (*SM*), and the retention time as the time till the storage node (*SN*) degrades by $100mV$.

The manufactured devices are designed for 2T0C/3T0C memory applications. These ultra-low leakage IGZO-TFTs [6] with raised source/drain architecture reach $I_{ON} > 30\mu A/\mu m$ at $V_{th} = 0.3V$ while maintaining an I_{OFF} as low as $4e{-}19A/\mu m$. By combining oxygen tunnel passivation with CAAC-IGZO raised contacts, the device preserves high doping near the contacts, enabling high-performance and low-power operation memory stacks. The device is first characterized, and then used for CM calibration across wide temperature ranges, enabling retention and performance temperature dependent analyses (Fig. 3). Gate leakage temperature dependence was carefully calibrated to accurately enable retention studies.

III. 2T0C and 3T0C Configurations and Tradeoffs

The present section discusses the proposed 2T0C/3T0C architectures, and introduces the tradeoffs on perfor-

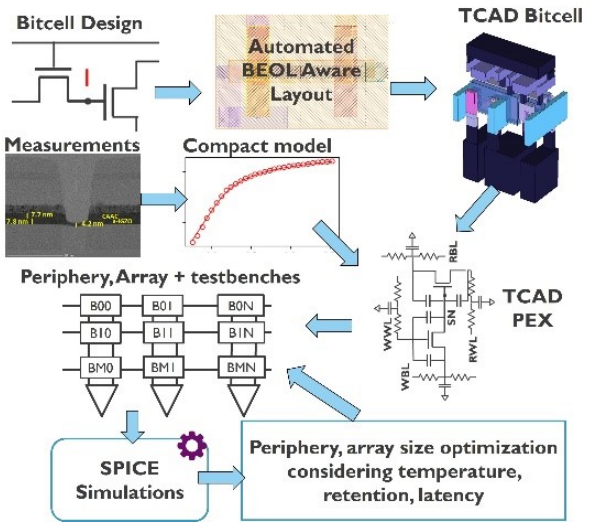

Fig. 2. Methodology: A parameterized layout generator is developed and used to generate bitcell layout based on design specifications. The layout is fed to Synopsys SDE to create 3D structure and Synopsys Raphael used for PEX generation. A BSIM-IMG CM is calibrated to measured data for length, width, and temperature variation. The PEX-enabled netlist and calibrated CM are used to simulate arrays in SPICE. Periphery and array size are optimized considering read access delay, retention time, and area. IGZO L_g remains unaltered as per the manufactured devices.

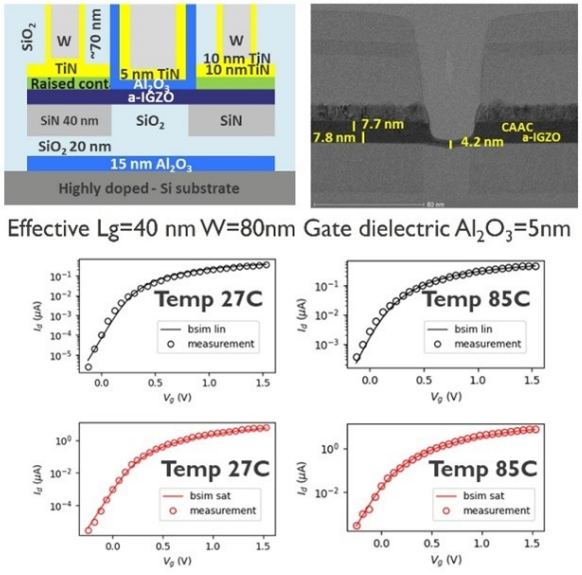

Fig. 3. (a) Schematic of IGZO-TFT used in this work and the corresponding TEM image (b). Calibration of the compact model to measured Id-Vg data for L_g ($40nm$ to $1\mu m$) [6], width, and temperature (27°C to 85°C). Gate leakage is also calibrated to the measurement data.

mance/density that are in depth explored in Section IV.

Fig. 4-a describes the eDRAM bitcell schematics. During write phase, the `Wtr` transistor charges the storage node at the gate of the storage transistor `Str`. In 2T0C, the absence of a dedicated read transistor (`Rtr`) compared to 3T0C reduces both area and bitcell parasitics, but large-array implementations suffer from increased read delay caused by IR drop along the *RWL*, and direct coupling between *RWL* and the *SN* [7],

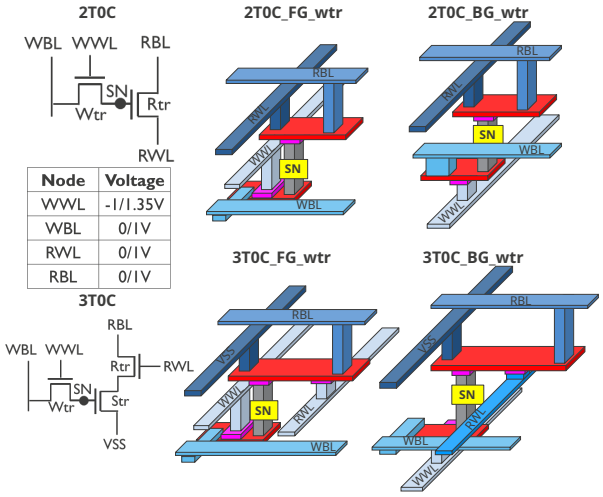

Fig. 4. Proposed 2T0C and 3T0C cells using front/back-gated `Wtr` transistor.

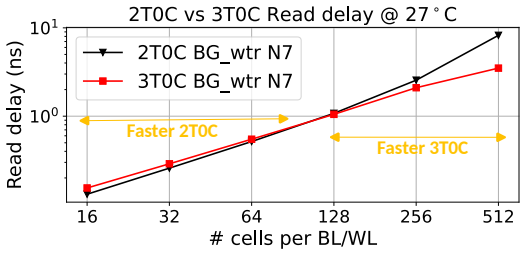

Fig. 5. Read access time v/s array capacity on N7 node. Shorter *BL* benefits 2T0C configurations due to reduced bitcell parasitics. 3T0C arrays outperform 2T0C on larger arrays.

thus limiting the performance at larger array sizes (Fig. 5). By contrast, 3T0C bitcells use a separate read path, improving read performance in larger arrays at the expense of bitcell area.

Fig. 4-b illustrates the proposed front- and back-gated (FG/BG) stacked-IGZO processes. As introduced before, BG-enabled bitcells reduce bitcell area up to 33% compared to FG-based alternatives. Table I compares the achieved cell areas across technology nodes, and describes the key process characteristics. As seen in Fig. 7 BG configurations in FinFET nodes provide better densities comparable to A10 SRAM bitcells. These results consider BEOL scaling effects, though transistor corresponds to non-scaled device, limiting area reduction to Y-dimension. For N7 and N40 nodes, $Lg = 40nm$ is used for `Rtr` and `Wtr`. Further, $W = 40nm$ for `Wtr` is established for both technology nodes, whereas for `Rtr`, W is $40nm$ and $80nm$ in the N7 and N40 nodes, respectively. Future gate scaling explorations for IGZO-based devices will be needed to fully leverage node scaling.

These bitcell scaling through nodes, architectures and processes has a significant impact in the related parasitics and therefore, in the final circuit performance. Furthermore, as described in Fig. 6-a, back gated write transistor-based cell has less *WWL-SN* coupling, which translates, as explored in Section IV, into higher performance. Moreover, BG enable the

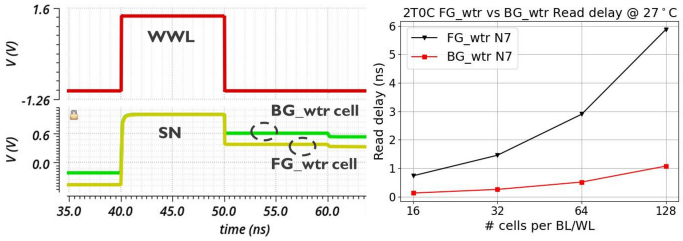

Fig. 6. (a) Impact of *WWL-SN* coupling on *SN* data in FG and BG 2T0C bitcells. (b) Comparison of 2T0C FG and BG bitcells read delays in N7 node.

TABLE I
PROCESS PITCHES AND BITCELL/SENSING AREA FIGURES.

Node	CPP (nm)	MPP (nm)	3T0C FG (μm^2)	3T0C BG (μm^2)	2T0C FG (μm^2)	2T0C BG (μm^2)	SA (μm^2)
N40	180	140	0.112	0.061	0.084	0.046	2.5
N12	78	64	0.037	0.022	0.028	0.016	0.5
N7	57	40	0.030	0.020	0.023	0.015	0.23

optimization of `Rtr` and `Wtr` processes separately.

IV. Bitcell and Array Explorations

This section explores the performance, energy and retention of memory arrays constructed employing 2T0C/3T0C bitcell configurations shown in Fig. 4, for N7 and N40 technology nodes. A differential sensing approach is implemented for readout operation, making use of a reference row located on the bank symmetric to the one being read. The reference row stores $0.7V_{DD}$ at *SN* node; this voltage level is based on exploration to minimize the worst read delay.

A. Sensitivity to Temperature, Bitcell & Process Architectures

To analyze tradeoffs between bitcell configurations, FG/BG processes, and array scaling at different temperatures, Fig. 9 examines read access time and energy efficiency. Both 2T0C and 3T0C configurations show a linear (exponential) relationship between delay and elements in the *WL* (*BL*), limiting ultra-fast (< 0.5 ns) readouts to *BL* sizes of $16/32$ elements. At larger capacities, 3T0C outperforms 2T0C due to its decoupled read path. While advanced nodes reduce RC delays through area scaling, higher resistances penalize larger configurations. The reduced area on BG configurations leads to reduced parasitics and improved performance. At $85°$C, where 3T0C benefits from higher current in IGZO devices due to negative V_{th} shift and greater number of tail states activation [6], while I_{ON} increases in 2T0C lead to greater *RWL* drops, causing slower, energy-intensive reads.

B. Retention

In the present experiments, cell retention within an array is defined as the time elapsed between a write operation and the point at which the SN degrades by $100mV$. Bitcell- and architecture-level experiments were conducted for both N7 and N40 technology nodes. Retention statistics for N7 exhibit minimal sensitivity to parasitic effects, consistent with

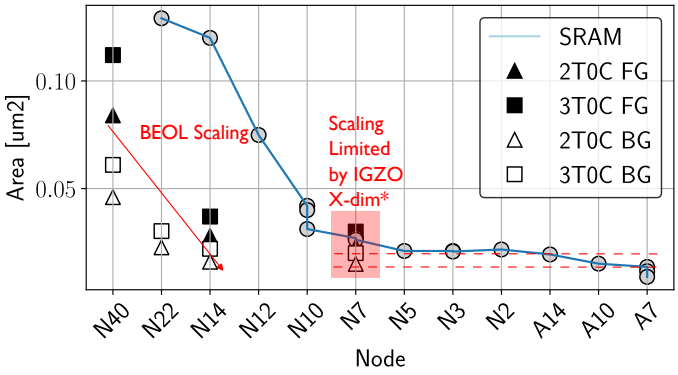

Fig. 7. Comparison of proposed 3T0C and 2T0C bitcell structures against SRAM across nodes. The bitcell scaling in this study is limited by the fixed L_g (*) of the manufactured IGZO devices. While L_g is expected to scale, our CM has been validated for the ranges of geometries discussed in Section III.

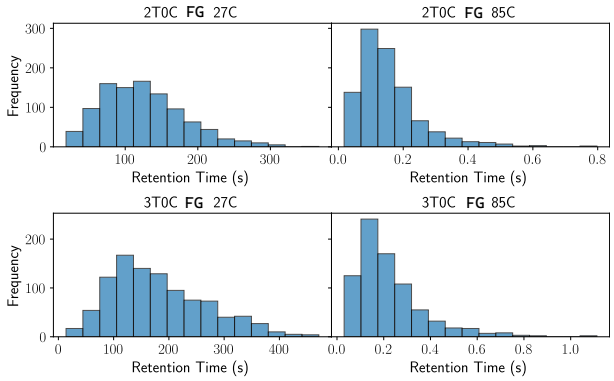

Fig. 8. 2T0C/3T0C FG retention trends as a function of the temperature. Our analysis indicates that further device engineering is required to reduce negative gate voltages to improve bitcell retention.

observations at other nodes. However, as shown in Fig. 8, elevated temperatures cause a marked degradation in retention, emphasizing the need for improved leakage control through advanced device engineering. Monte Carlo simulations reveal that the retention time distribution not only shifts and scales but also undergoes a change in shape, with values ranging from tens to hundreds of milliseconds at $85°$C.

C. 8 KiB Macros Exploration

8 KiB macros are enabled by using different arrays configurations, exploring the impact of the internal arrays configurations on the delivered performance. As seen in Fig. 10, despite having equal capacity, different architectures exhibit different performance characteristics, due to the following effects: ① 2T0C arrays with a large number of elements on *RWL/WWL* suffer from large IR-drop on *RWL*, reducing *Rtr*'s V_{GS} for the cells located far from *RWL* driver. Consequently, read delay increases. Performance degrades further at high temperatures due to increased IR-drop from higher leakage and I_{ON} in IGZO devices. This effect can be reduced by optimizing *RWL* resistance. ② During a read operation *RWL* toggles from low-to-high (high-to-low) for 3T0C (2T0C) arrays, causing *SN* level to increase (decrease) for 3T0C (2T0C) bitcells due to

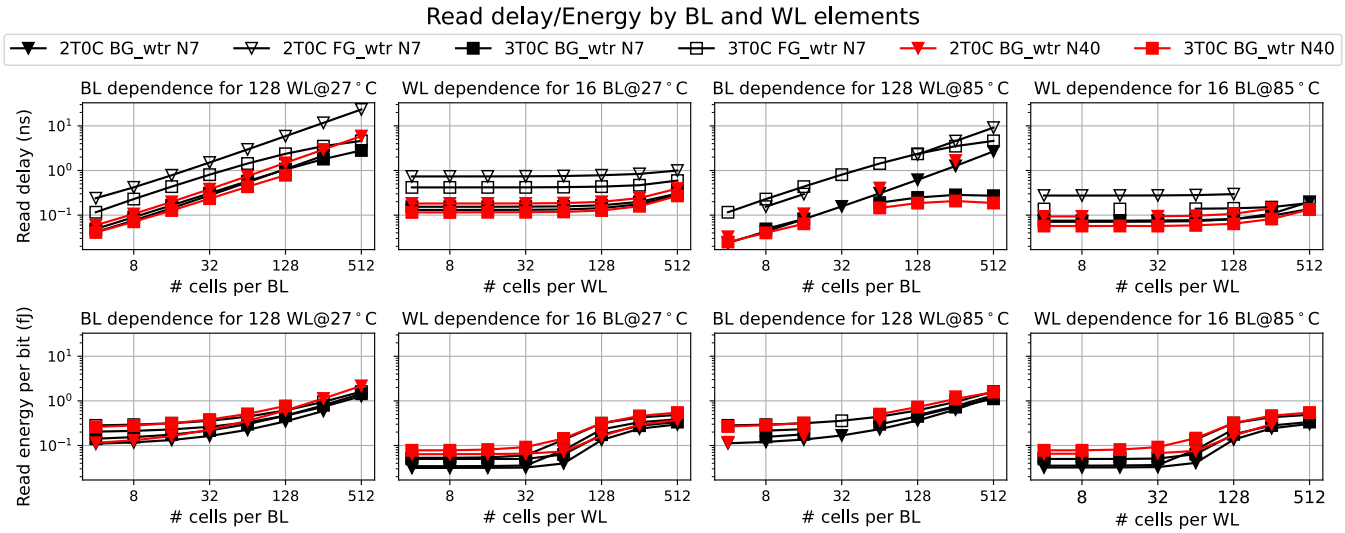

Fig. 9. Read delay and energy sensitivity to *WL* (fixed BL=16) and *BL* (fixed WL=128) elements for different bitcells and node, at both 27°C and 85°C, showing linear (BL) and exponential (WL) trends. 3T0C holds promise in terms of higher read speed due to decoupled read path and could be beneficial for larger macros.

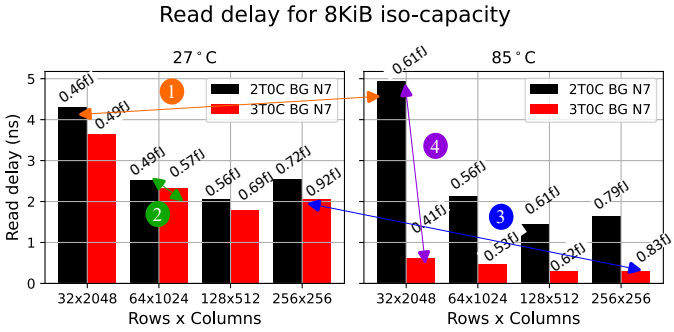

Fig. 10. Performance and energy comparison on different 8 KiB macros.

RWL-SN coupling. This improves the read performance of 3T0C arrays compared to 2T0C arrays. ③ In 3T0C cells two series transistors in read path limit its driving current; this is critical at advanced nodes where the read transistor's width is limited. However, at higher temperatures, the IGZO I_{ON} increases and significantly improves 3T0C arrays performance. ④ 3T0C bitcells, by using a dedicated read transistor, isolate/cut-off unselected cells on *RBL*, and becoming less sensitive to leakage of unselected cells. Consequently, at higher temperatures 3T0C memory arrays performance is significantly better compared to 2T0C counterparts.

Energy wise, 2T0C offer reduced consumption compared to 3T0C approaches as per area dependent reduced parasitic capacitances. Similarly, larger parasitic capacitances on *BL* penalize array configurations with larger number of rows.

V. Conclusions and Discussion

This work presents a comprehensive roadmap (Fig. 7) for IGZO-based BEOL-compatible eDRAM as a scalable, high-density memory solution. Flexible options for denser, ultra-fast and energy efficient SRAM replacements are achieved through bitcell, device, and array optimization. The roadmap highlights the density of BG-processes for both 2T0C small capacity high-speed arrays, and for 3T0C larger capacities, thus tailoring to diverse application requirements, reaching 0.015 μm^2 and 0.02 μm^2 cells for 2T0C and 3T0C configurations. Combined with CMOS BEOL scaling, N7 enabled 3T0C delivers sub-2ns/sub-0.3ns read delays on 8 KiB arrays at 27°C and 85°C respectively. However, key tradeoffs emerge, including the need for IGZO gate length scaling to fully leverage node scaling. Similarly, retention demands further device engineering for high-temperature operations. At the μArchitecture level, non-square arrays can be optimized for performance and energy efficient solutions. Overcoming these challenges is essential for advancing IGZO eDRAM as a next-generation memory solution in the bonding era.

Acknowledgment

This work has been enabled in part by the NanoIC pilot line (nanoic-project.eu), jointly funded by the Chips Joint Undertaking, through the European Union's Digital Europe (101183266) and Horizon Europe programs (101183277).

The authors would like to thank D. Abdi, C. Apte, K. English, R. Schreiber, and J. Wuu for their helpful discussions.

References

[1] H. Kükner *et al.*, "Double-Row CFET: Design Technology Co-Optimization for Area Efficient A7 Technology Node," in *IEEE IEDM*, 2024.
[2] F. Garcia-Redondo *et al.*, "STT-MRAM Stochastic and Defects-aware DTCO for Last Level Cache at Advanced Process Nodes," in *IEEE ESSDERC*, 2023.
[3] Sentaurus Structure Editor – Synopsys Inc., Mountain View, CA, USA.
[4] Raphael FX - Synopsys Inc., Mountain View, CA, USA.
[5] Cadence® spectre® circuit simulator/virtuoso® layout suite.
[6] S. Subhechha *et al.*, "Ultra-low Leakage IGZO-TFTs with Raised Source/Drain for Vt $>$ 0 V and Ion $>$ 30 µA/µm," in *IEEE VLSI*, 2022.
[7] L. Zheng, Z. Wang, Z. Lin and M. Si, "The Impact of Parasitic Capacitance on the Memory Characteristics of 2T0C DRAM and New Writing Strategy," in *IEEE EDL*, vol. 44, no. 8, pp. 1284-1287, 2023.