nature portfolio

Peer Review File

Oxide semiconductor gain cell-embedded memory: materials and integration strategies for next generation on-chip memory

Corresponding Author: Professor Jae Kyeong Jeong

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The paper presents a review on the recent progress in using oxide semiconductor transistors for Gain Cell eDRAM. However, some problems need to addressed before publication.
1. The title is "Materials and Integration Strategies", but the manuscript doesn't talk about integration
2. Ref is "memory-centric", which is different from "data-centric" computing. Ref 1-3 are all for edge computing, not for HPC, AI, etc. that are mentioned as well. These references cannot support the arguments.
3. Ref 21 and 23 only for IBM eDRAM, not for Intel, TSMC, Samsung that are mentioned. Ref 22 does not even have hardware demonstration. These references cannot support the arguments.
4. Line 66: what does "buffer layer" mean? Buffer memory?
5. Line 68: what does "continuous scaling process" mean?
6. Line 86: CPU cannot access the memory but it's not idle. Thread switching can make CPU busy with other work.
7. Line 87: peripheral circuit is still accessing memory during the refresh.
8. Line 88: what does "refresh management" mean?
9. Line 136: 10ns is not fast
10. Ref 57 is not from Intel and is not about finFET eDRAM. The reference is wrong.
11. Line 182: parasitic capacitance of gate-to-source coupling may degrade the performance.
12. Line 231: "2F2" is a critical claim, but there is no support (ref or figure or text explanation) for this point.
13. Line 364: how is "negative capacitance" relate to eDRAM topic?

Reviewer #2

(Remarks to the Author)
The paper entitled "Oxide Semiconductor Gain Cell-Embedded DRAM: Materials and Integration Strategies for Next-Generation On-Chip Memory" provides a comprehensive review of gain-cell embedded DRAM (GC-eDRAM) technologies based on oxide semiconductors, proposed as a promising solution for future memory architectures. The work emphasizes how the unique properties of oxide semiconductors, such as ultra-low leakage currents and wide bandgaps, enable scalable, power-efficient, and high-density designs, including gate-all-around and channel-all-around architectures. The discussion offers a detailed analysis of device-level characteristics and associated challenges, although comparatively less attention is given to application-level requirements and broader technological drivers.
Overall, the paper is well written and clearly structured. It effectively addresses most of the key aspects related to DRAM architectures and transistor technologies based on oxide semiconductors, providing a comprehensive and technically sound overview of the topic. I therefore believe the paper is of good quality and suitable for publication. However, I have a few comments and suggestions that could further improve the clarity and completeness of the work, which I list below:
1) Rows 97-99 it is written: "Additionally, two other leakage currents exist regardless of the device conduction state: pn junction reverse bias leakage (IJunction) and gate tunneling current (ITunnel), both of which result from the band-to-band tunneling effect.". I believe the gate tunneling depends primarily on the gate stack composition and structure, and only in second order on the channel material. A clarification on this aspect would improve the accuracy of the discussion.
2) Following the previous point, the manuscript does not discuss the gate dielectric interface and material considerations in

oxide semiconductor TFTs. Since the interface quality between the insulator and the semiconductor plays a crucial role in determining device stability and reliability, I suggest expanding the discussion in this regard, particularly within the section addressing reliability and technological challenges.
3) Row 109, the acronym VCT is never defined.
4) It would be valuable to include a brief discussion on threshold voltage instability and hysteresis phenomena in oxide-based transistors, as these are key reliability concerns for such devices.
5) In the "Crystallization-driven advancements in oxide channel" section I suggest to add a discussion regarding the additional complexity introduced by the crystallization process, which often includes high temperature annealing.
6) In the "Negative Capacitance: Beyond Boltzmann's Tyranny" section, it is discussed the benefit of using ferroelectric materials in the gate stack for transistors. This concept is not new and not exclusive of oxide-based TFT. In addition, the polarization curve of the ferroelectric layer introduces threshold drift in the transistor and hysteresis in the characteristic, which is not desirable. Add this in the discussion or remove the section.
7) I suggest adding a section (or including it in the final section) regarding current trends and challenges for GC DRAM and their realization with oxide semiconductors.

Reviewer #3

(Remarks to the Author)
Oxide Semiconductor：1.In line 66, the author state that "eDRAM not only serves as a cache but also acts as a buffer layer, mitigating the latency gap between CPU and DRAM." However, this description is somewhat confusing, since a cache itself is inherently designed to bridge the latency gap
2.The author claims that "conventional 1T1C eDRAM adopts a continuous scaling process for performance and density improvement; however, this leads to a reduction in the physical size of the capacitor and an increase in transistor leakage current, resulting in degraded data retention time and limited scalability. To address these issues, gain cell-embedded DRAM (GC-eDRAM), which removes the capacitor and consists of 2-4 transistors, has garnered attention." However, this description does not convincingly explain the advantages of GC-eDRAM over conventional 1T1C eDRAM. As device dimensions scale down, the storage capacitance of the gain cell (i.e., the gate capacitance of the read transistor) also decreases, while leakage current increases—similar to the challenges faced by 1T1C cells. Therefore, the scaling issue remains essentially the same. The true advantages of GC-eDRAM lie in its non-destructive read operation and process compatibility, whereas 1T1C eDRAM suffers from the difficulties of fabricating high–aspect-ratio capacitors and poor integration compatibility with logic processes.
3.In line 83, the authors state that "Above mentioned, the GC-eDRAM structure can alleviate the complexity and cost issues in manufacturing processes arising from the high capacitance requirements of the 1T1C eDRAM." However, no such discussion or supporting explanation is actually provided in the preceding text.
4.The authors state that there are three main causes for leakage current: leakage current issues due to silicon-based channel materials, short-channel effects (SCE), and the floating-body effect. However, fundamentally, both the short-channel effect and the floating-body effect are manifestations of leakage current issues inherent to silicon-based channel materials, rather than independent categories. These two effects simply aggravate channel leakage instead of representing distinct leakage mechanisms.
5.In line 115, the first "i" in "indium gallium zinc oxide (IGZO)" appears in red. The authors should pay closer attention to such formatting issues.
6.In the "Application of Oxide Semiconductors to GC-eDRAM Architectures" section, the authors discuss the use of GC-eDRAM for computing-in-memory (CIM). However, there are already CIM applications based on silicon-based GC-eDRAM. Therefore, this section currently reads more as a discussion of GC-eDRAM applications in general, rather than highlighting the advantages of oxide semiconductor-based GC-eDRAM. The authors should focus more on describing applications that oxide semiconductor GC-eDRAM can enable, which conventional GC-eDRAM cannot.
7.In the line 133, the authors state: "Motivated by these requirements, a GC-DRAM architecture was implemented in which the Wtr adopted an In-rich ITO channel to enhance electron transport and accelerate switching, whereas the Wtr employed an a-IGZO channel to suppress leakage owing to its superior off-state characteristics (Fig. 2b)" However, it is unclear why the Wtr would use both an In-rich ITO channel and an a-IGZO channel. The first "Wtr" should likely be "Rtr," as the current wording contains a basic error.
8.In line 137, the authors claim that the device performance "is comparable to that of quasi-nonvolatile memory." I am curious whether there is a formal definition of "quasi-nonvolatile memory." For example, nonvolatile memory typically requires a retention time exceeding 10 years, so what retention time would qualify as "quasi-nonvolatile"? If no formal definition exists, the authors should clarify or reconsider this description.
9.In the "Scaling Beyond Limits: Multi-Gate Transistors" section, the authors devote more than half of the content to describing basic semiconductor scaling trends, from planar MOSFETs to FinFETs to GAA structures. These are fundamental concepts and should not occupy such a large portion of the section.
10.In the line 181, the author state: "The surrounding-gate structure enhances gate-to-channel capacitive coupling, leading to an increase in parasitic capacitance. However, this effect paradoxically improves data retention and sensing performance, offering distinct advantages in eDRAM applications" . However, in GC-eDRAM, the storage relies on the gate capacitance of the read transistor. An increase in the parasitic capacitance of the write transistor would similarly degrade performance due to the pull-down effect when the write transistor is turned off.
11.In the line 184, the author state: "In the CAA configuration, the gate encircles the channel; in the GAA configuration, the channel is surrounded by the gate". This description does not clearly explain the distinction between CAA and GAA.
12.In the benchmark table, the author lists the length and width of the reference No.1 (2T eDRAM) is 30000 nm and 50 μm, respectively. However, in the original referenced paper, the length and width are 3 μm and 5 μm. The authors should avoid such basic errors.

13.The authors claim that IGZO-eDRAM achieves an area size of $4F^2$ or even $2F^2$. I am curious whether this area includes the space required for routing. If not, when routing is taken into account, would the effective area size need to be increased further? I understand the cell area 2F2 is claimed by the citation, however, the author of this review paper should carefully think about it.
14.In modern CPUs, caches and CPU cores are usually fabricated on the same die, which implies that process compatibility must be considered. The authors should discuss whether oxide GC-eDRAM can be co-integrated with silicon-based processing units. If not, they should address how this issue could be resolved.
The authors should provide more background on GC-eDRAM itself, including details on different configurations such as 2T-DRAM and 3T-DRAM.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
1. Line 77: “is sensed through capacitive coupling via the read-transistor gate dielectric”. Capacitive coupling is not a sensing mechanism. Capacitive coupling degrades sensing.
2. Line 88-93: The comparison on 3T0C vs 2T0C is wrong. 3T0C is used to suppress sneak path from unselected cells, not disturbance to storage node.

Reviewer #2

(Remarks to the Author)
The authors have addressed all of my questions and comments. They have added new paragraphs and revised parts of the manuscript accordingly. I recommend the paper for publication.

Reviewer #3

(Remarks to the Author)
the authors had addressed the concern.

Authors: Sang Won Chung, Seong Hun Yoon, Jae Kyeong Jeong

Title: Oxide Semiconductor Gain Cell-Embedded DRAM: Materials and Integration Strategies for Next-Generation On-Chip Memory

Manuscript ID: COMMSENG-25-0741

RECOMMENDATIONS AND COMMENTS FOR AUTHOR(S)

Answer to the comments by Reviewer #1 and Summary of changes

We are grateful to the Reviewer for his/her constructive comments and corrections on our manuscript. After considering the comments, we revised our manuscript and made the following reply.

1. The title is “Materials and Integration Strategies”, but the manuscript doesn’t talk about integration.

Answer: We sincerely thank the reviewer for this preceptive observation regarding the alignment between our title and the manuscript’s content. We agree that the aspect of system-level integration warrants a more comprehensive discussion to fully justify the scope implied by the title.

On page 6, line 180

“**Integration of OSs into GC-eDRAM Architectures**

Leveraging the potential for monolithic integration, recent advancements in M3D technology for OS-based GC-eDRAM have yielded breakthrough achievements in mitigating memory bottlenecks. Empirical evidence suggests that the refresh overhead of OS-based gain cells reaches a point of saturation once retention time approaches approximately 10 seconds. This implies that subsequent design optimization should prioritize system-level efficiency rather than exclusively pursuing the maximization of retention time (as shown in Fig. 3a)[51]. Consistent with this paradigm, a state-of-the-art M3D-SD (M3D SRAM and DRAM) architecture has been demonstrated, featuring the vertical stacking of IGZO pass gates and IGZO 2T0C DRAM directly atop Si CMOS (Fig. 3b)[52]. This configuration achieves a compact 4T footprint and a 51% reduction in static power, while simultaneously facilitating ultra-low latency data transfer (< 10 ns) and robust retention characteristics exceeding 5,000 seconds. Furthermore, the introduction of an n/p hybrid structure utilizing a counteractive coupling mechanism has proven effective in neutralizing charge injection noise, thereby ensuring enhanced data stability (Fig. 3c)[53]. By capitalizing on this improved effective sensing margin, the architecture delivers a 48.25-fold enhancement in processing speed for edge AI workloads relative to conventional planar counterparts. The successful proliferation of GC-eDRAM, however, remains contingent upon the development of foundational technologies that can implement high-performance, uniform OS transistors and circuits within stringent low-temperature thermal budgets. This requirement is substantiated by the realization

of a 2-kbit array based on body-contacted IGZO FETs (Fig. 3d)[54]. Fabricated at 350 °C–a temperature with BEOL Cu-interconnect processes–this array maintains highly uniform performance and has been utilized to demonstrate an all-nMOS analog memory circuit functioning as a short-term memory (STM) synapse. Nevertheless, as architectural and material innovations continue to push the boundaries of density, further device miniaturization inevitably encounters emerging physical limitations."

2. Ref is "memory-centric", which is different from "data-centric" computing. Ref 1-3 are all for edge computing, not for HPC, AI, etc. that are mentioned as well. These references cannot support the arguments.

Answer: We sincerely thank the reviewer for their rigorous scrutiny of our conceptual terminology and supporting citation. We acknowledge that the distinction between "memory-centric" and "data-centric" computing is fundamental, and that our original references [1-3] were insufficiently aligned with the HPC and AI contexts discussed in the manuscript. To address this, we have conducted a thorough revision of both the terminology and the bibliography to ensure scholarly accuracy:

- **Conceptual Refinement**: We have audited the manuscript to ensure that "data-centric computing" is utilized precisely to describe architecture where the focus shifts from processing-heavy to data-movement-efficient paradigms.
- **Citation Update (Page 17, Line 513)**: The previous edge-focused references have been replaced with the following high-impact surveys and papers that specifically address data-centric frontiers, hybrid memory systems, and Near-Memory Computing (NMC) in the context of high-performance applications:

    1. **Fakhry et al. (2023):** Provides a comprehensive review of computational storage and near-memory computing tailored for **high-performance applications**.
    2. **Alsaket et al. (2026):** Discusses advanced **hybrid memory systems (NVM/DRAM)** and their architectural management, directly supporting our arguments on evolving memory hierarchies.
    3. **Siegl et al. (2016):** A foundational survey on **Data-Centric Computing Frontiers** and Processing-In-Memory (PIM), establishing the theoretical basis for our discussion on HPC and AI demands.

We believe these updates provide the necessary empirical and theoretical rigour to support our claims regarding the escalating demands of modern, large-scale computing systems.

On page 3, line 47

"The rapid technological proliferation and commercialization of data-centric paradigm–encompassing high-performance computing (HPC), artificial intelligence (AI), cloud-to-edge infrastructures–have necessitated a fundamental paradigm shift. This evolution demands not only unprecedented computational throughput but also a transformative enhancement in memory system capabilities to alleviate long-standing data-bottlenecks[1-3]."

On page 17, line 510

1. Fakhry, D., Abdelsalam, M., El-Kharashi, M. W. & Safar, M. A review on computational storage devices and near memory computing for high performance applications. *Memories - Materials, Devices, Circuits and Systems* **4**, 100051 (2023).

2. Alsaket, J., Adel, A. & Alwadi, M. A survey on hybrid memory systems integrating non-volatile memory (NVM) and DRAM: architectures, management techniques, and research directions. *Journal of Systems Architecture* **170**, 103603 (2026).

3. Siegl, P., Buchty, R. & Berekovic, M. Data-centric computing frontiers: a survey on processing-in-memory. in *Proceedings of the Second International Symposium on Memory Systems* 295–308 (ACM, Alexandria VA USA, 2016). doi:10.1145/2989081.2989087.

3. Ref 21 and 23 only for IBM eDRAM, not for Intel, TSMC, Samsung that are mentioned. Ref 22 does not even have hardware demonstration. These references cannot support the arguments.

Answer: We appreciate the reviewer's meticulous review of our citations. We agree that the previous references did not provide adequate coverage for all the industry leaders mentioned and that certain citation lacked the necessary hardware demonstration. To ensure our claims are grounded in rigorous empirical evidence, we have overhauled the citation (Page 18, Line 564) to strictly align each company with its respective hardware-validated achievements;

- **Intel [21]:** We now cite **Hamzaoglu et al. (JSSC, 2015)**, which demonstrates a 1 Gb, 2 GHz eDRAM fabricated using **22 nm Tri-Gate CMOS technology**, providing a clear hardware benchmark for Intel's scaling.
- **TSMC [22]:** We have included **Huang et al. (IEDM, 2011)**, which reports a high-performance **28 nm eDRAM** technology utilizing High-K/Metal-Gate (HKMG) structures.
- **Samsung / Emerging Integration [24]:** To address the lack of hardware demonstration in the previous version, we have incorporated **Xiao et al. (Nature Communications, 2024)**. This recent work presents a **High-performance Si-MoS_2 heterogeneous eDRAM**, offering the requisite experimental validation for next-generation integration strategies.

By reorganizing these references, we have ensured that every industrial and architectural claim is

supported by peer-reviewed, experimental data.

On page 18, line 561

21. Hamzaoglu, F. *et al*. A 1 Gb 2 GHz 128 GB/s bandwidth embedded DRAM in 22 nm tri-gate CMOS technology. *IEEE J. Solid-State Circuits* **50**, 150–157 (2015).

22. Huang, K. C. *et al*. A high-performance, high-density 28nm eDRAM technology with high-K/metal-gate. in 2011 *International Electron Devices Meeting* 24.7.1-24.7.4 (IEEE, Washington, DC, USA, 2011). doi:10.1109/IEDM.2011.6131608

24. Xiao, K. et al. High performance Si-MoS_2 heterogeneous embedded DRAM. *Nat Commun* **15**, 9782 (2024).

4. Line 66: what does “buffer layer” mean? Buffer memory?

Answer: We apologize for the terminological ambiguity in the original manuscript. The review is correct in assuming that the term was intended to describe a **buffer memory hierarchy**. To eliminate any potential confusion with physical thin-film buffer layers, we have replaced “buffer layer” with a more precise architectural description. The revised text now explicitly defines eDRAM’s functional role as a high-capacity intermediary within the memory hierarchy. Now it reads:

On page 3, line 66

“Beyond its role as a last-level cache (LLC), eDRAM acts as pivotal high-density memory buffer bridging the gap between processing cores and off-chip main memory. By hosting significantly large working sets, it effectively alleviates the bandwidth-latency bottleneck inherent in traditional SRAM-limited hierarchies[25].”

5. Line 68: what does “continuous scaling process” mean?

Answer: We appreciate the reviewer’s request for clarification. The term "continuous scaling process" was originally intended to describe the iterative reduction of DRAM cell dimensions to enhance performance and density. However, we recognize that this phrasing was overly general and failed to capture the specific technical impediments currently facing the industry. In the revised manuscript, we have replaced this ambiguous phrasing with a detailed explanation of the **intrinsic scaling limits**—specifically the fabrication complexities of high-aspect-ratio (HAR) capacitors and the inherent logic-process incompatibility of 1T1C architectures. We believe this modification provides a more rigorous justification for the transition toward OS-based gain cells.

On page 3, line 69

"However, conventional 1T1C eDRAM faces substantial manufacturing challenges associated with the fabrication of high-aspect-ratio (HAR) storage capacitors and exhibits limited integration compatibility with standard logic processes, both of which hinder further density scaling[26,27]."

On page 18, line 572

26. Chun, K. C., Jain, P., Kim, T.-H. & Kim, C. H. A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches. *IEEE J. Solid-State Circuits* **47**, 547–559 (2012).

6. Line 86: CPU cannot access the memory but it's not idle. Thread switching can make CPU busy with other work.

Answer: We sincerely appreciate the reviewer's astute observation regarding CPU utilization during memory refresh cycles. We fully agree that the CPU is not necessarily "idle" in a strict sense, as modern architectural techniques like **hardware multithreading and context switching** allow the processor to execute other threads while waiting for a memory bank to become available. However, we believe the fundamental issue remains: the **temporary unavailability of target banks** during refresh leads to increased resource contention and scheduling overhead. To reflect this nuance accurately, we have revised the manuscript to emphasize that while thread switching can mask some latency, the degradation of **effective bandwidth and increased queuing delays** still impose a significant performance penalty on the system.

On page 4, line 103

"While architectural strategies such as context switching and bank interleaving can partially mask the overhead of memory refresh, the periodic unavailability of target banks inevitably triggers resource contention. Such contention introduces additional queuing delays and diminishes effective bandwidth, thereby degrading overall system performance a phenomenon particularly pronounced in memory-intensive workloads where data-level parallelism is critical."

7. Line 87: peripheral circuit is still accessing memory during the refresh.

Answer: We thank the reviewer for highlighting that peripheral circuits maintain access to unaffected memory regions. We agree that our previous statement implying total system inaccessibility was overly broad, as bank interleaving allows access to banks not currently undergoing refresh. We have revised the text to clarify that access restrictions apply specifically to the "target banks" undergoing refresh. The updated sentence explains that while bank interleaving mitigates the impact, the reduced effective bandwidth and increased latency for the affected banks still result in overall system performance

degradation.

On page 4, line 106

"While architectural strategies such as context switching and bank interleaving can partially mask the overhead of memory refresh, the periodic unavailability of target banks inevitably triggers resource contention. Such contention introduces additional queuing delays and diminishes effective bandwidth, thereby degrading overall system performance a phenomenon particularly pronounced in memory-intensive workloads where data-level parallelism is critical."

8. Line 88: what does "refresh management" mean?

Answer: We appreciate the reviewer's request for clarification on this term. In the context of our manuscript, **"refresh management"** specifically refers to the specialized control logic and peripheral circuitry required to orchestrate the periodic restoration of stored charges in GC-eDRAM. To provide a more precise technical definition, we have revised the manuscript to clarify that this overhead encompasses:

1. The **scheduling logic** necessary for tracking the retention status of individual rows or banks.
2. The **arbitration mechanisms** required to resolve conflicts between refresh cycles and incoming memory access requests.
3. The **additional sensing and driving circuitry** needed to ensure data integrity within the strict retention window.

We have updated the text to emphasize how these requirements translate into physical design constraints.

On page 4, line 110

"Furthermore, the refresh management overhead comprising the dedicated control logic and peripheral circuitry tasked with scheduling and executing periodic refresh operations to ensure data integrity imposes significant demands on the design. This necessitates additional logic-gate counts and routing complexity, thereby exacerbating both the effective device footprint and the overall process integration challenges[32]."

9. Line 136: 10ns is not fast

Answer: We appreciate the reviewer's critical assessment regarding our characterization of the write speed. We agree that a 10-ns write pulse is not "fast" when compared to the sub-nanosecond access times of high-performance SRAM or the latest commodity DRAM. Our intention was to highlight that stable programming can be achieved within a relatively brief **10-ns window** while simultaneously securing an **extraordinary retention time (> 25 hours)**. This performance represents a significant advancement for OS-based gain cells, particularly in balancing write efficiency with quasi-nonvolatile

data persistence. To avoid any overstatement and to clarify the technical context, we have revised the manuscript to emphasize the **efficiency and stability** of the write operation rather than its absolute speed.

On page 6, line 160

"This configuration achieved stable operation with a 10-ns write pulse and an exceptional data-retention capability exceeding 25 hours, which significantly surpasses the millisecond-range retention of conventional Si-DRAM, effectively enabling extended-retention memory application and drastically reducing the power penalty associated with periodic refresh operations. This performance profile aligns with the requirements of quasi-nonvolatile memory, offering a compelling balance between energy-efficient programming and long-term data persistence."

10. Ref 57 is not from Intel and is not about finFET eDRAM. The reference is wrong.

Answer: We sincerely apologize for the oversight regarding Reference [57]. Upon re-examination, we confirmed that the citation was indeed mismatched with the technical description in the text. In the process of addressing this, we also integrated feedback from another reviewer who suggested that the discussion on **conventional semiconductor scaling trends** was overly pedagogical and detracted from the manuscript's primary focus. Consequently, we have **restructured the relevant section** to prioritize advanced, oxide-specific advancements.

11. Line 182: parasitic capacitance of gate-to-source coupling may degrade the performance.

Answer: We sincerely appreciate the reviewer's perceptive comment regarding the performance implications of **parasitic gate-to-source coupling**. We agree that the same surrounding-gate topology that enhances electrostatic control can inadvertently increase capacitive coupling, leading to voltage disturbances (e.g., the pull-down effect) and increased dynamic loading. To address this, we have substantively revised the manuscript to provide a more nuanced and balanced analysis. We now characterize this phenomenon as a fundamental **design trade-off**. Specifically, we have added a discussion on how the advantages of enhanced storage capacitance (C_{SN}) must be weighed against the drawbacks of increased coupling noise. Furthermore, we have incorporated recent literature [63, 64] that explores mitigation strategies, such as selecting between **Channel-All-Around (CAA)** and **Gate-All-Around (GAA)** structures and implementing **coupling-aware operation schemes**.

On page 7, line 219

"The surrounding-gate topology fundamentally strengthens gate-to-channel electrostatic control, which

is instrumental in suppressing leakage current and increasing the effective storage node capacitance (C_{SN}). While an increased C_{SN} generally improves data retention and sensing margin, the same geometry concurrently enhances capacitive coupling between the storage node and control terminals, such as write word line ($C_{WWL\text{-}SN}$) and read word-line ($C_{RWL\text{-}SN}$). During word-line transitions, these parasitic paths can induce an unintended storage node voltage disturbance (ΔV_{SN})–commonly referred to as the pull-down effect–which may degrade the stored voltage level and increase dynamic loading. Recent studies indicate that this coupling constitutes a quantifiable design trade-off that can be mitigated through the strategic selection of surrounding-gate geometries (e.g., CAA vs. GAA) and the implementation of coupling-aware writing strategies[63, 64]. Such optimizations are essential to ensure that the electrostatic benefits of multi-gate structures are not neutralized by parasitic-induced signal degradation."

On page 20, line 661

63. Kim, M., Lee, J., Cho, H. & Jeon, J. Enhancing memory performance in IGZO-based 2T0C DRAM through comparative analysis of CAA and GAA FET structures. *J Comput Electron* **24**, 190 (2025).

64. Zheng, L., Wang, Z., Lin, Z. & Si, M. The impact of parasitic capacitance on the memory characteristics of 2T0C DRAM and new writing strategy. *IEEE Electron Device Lett.* **44**, 1284–1287 (2023).

12. Line 231: "$2F^2$" is a critical claim, but there is no support (ref or figure or text explanation) for this point.

Answer: We appreciate the reviewer's request for technical substantiation regarding the $2F^2$ density claim. We agree that such a significant advancement in cell footprint requires explicit architectural justification. In the revised manuscript, we have provided a detailed explanation of how the transition from planar to **vertical 3D architectures** facilitates this scaling. Specifically, we have clarified that while conventional planar 1T1C cells are limited to $6F^2$–$8F^2$, the **monolithic vertical stacking** of OS-based gain cells allows for a projected footprint of $2F^2$ by decoupling the access transistor and the storage node into a vertical plane.

To support this claim, we have integrated the following evidence:

- **Visual Evidence:** We have updated the text to directly reference **Fig. 4b**, which provides a schematic layout of the $2F^2$ operation. Furthermore, we have highlighted **Figs. 4g and 4h**, which illustrate the specific stacked configurations and cross-sectional views necessary to realize this ultra-high-density architecture.
- **Literature Support:** We have added **Reference [42]**, which provides the theoretical and empirical groundwork for scaling beyond $4F^2$ in 3D-integrated memory systems.

On page 9, line 271

"Furthermore, the proposed architecture represents a significant technological leap in data storage density, offering more than a twofold improvement over conventional planar 1T1C DRAM. By leveraging monolithic 3D vertical stacking, this approach facilitates cell-level scaling beyond the conventional $4F^2$ limit. Specifically, theoretical projections and layout analyses demonstrate that advanced implementations of this stacked GC-DRAM can achieve a $2F^2$ footprint (Fig. 4b)[42], contingent upon the optimized vertical integration of the storage node and transistor channels as depicted in Fig. 4g and 4h."

13. Line 364: how is "negative capacitance" relate to eDRAM topic?

Answer: We appreciate the reviewer's critical assessment regarding the inclusion of **negative capacitance (NC)** within the eDRAM context. Upon further reflection, we agree that while NC effects in ferroelectric-gate transistors are a significant area of research, they are not central to the oxide-based TFT gain-cell architectures presented in this work. As the reviewer correctly noted, the polarization switching and hysteresis inherent in ferroelectric materials can introduce instability in threshold voltage (V_{th}), which may compromise the reliable operation and sensing margins required for eDRAM applications. Since our manuscript focuses on achieving high-performance and stable gain-cells without the integration of ferroelectric layers, the discussion on NC was indeed extraneous.

To streamline the manuscript and maintain a rigorous focus on our primary contribution—**oxide-based monolithic 3D integration**—we have **removed the section concerning negative capacitance** in the revised version. This allows for a more cohesive narrative centered on the material and architectural innovations of OS-based memories.

Answer to the comments by Reviewer #2 and Summary of changes

We are grateful to the Reviewer for his/her kind comments and corrections on our manuscript. After considering the comments, we revised our manuscript and made the following reply.

0. The discussion offers a detailed analysis of device-level characteristics and associated challenges, although comparatively less attention is given to application-level requirements and broader technological drivers.

Answer: We sincerely thank the reviewer for this constructive feedback. We agree that contextualizing device-level innovations within the broader landscape of **application-specific requirements** and **architectural drivers** is essential to demonstrate the full impact of oxide semiconductor (OS)-based memory. To address this, we have significantly expanded the manuscript by introducing a new section: **"Engineering Strategies for Advanced Oxide Gain Cell-Embedded DRAM."** In this section, we transition the focus from individual device metrics to **system-level optimization**, specifically highlighting:

- **Efficiency-centric design:** Shifting the optimization priority toward system-level efficiency, based on the finding that refresh overhead saturates at approximately 10 seconds of retention [51].
- **Monolithic 3D (M3D) breakthroughs:** Discussing state-of-the-art integration schemes such as **M3D-SD** [52] and **n/p hybrid architectures** [53], which bridge the gap between material properties and circuit-level performance.
- **Target Applications:** Explicitly linking OS-based memory to **Edge AI acceleration** and **neuromorphic computing**, including short-term memory (STM) synapse implementations [54].

We have also added **Figure 3** to visually synthesize these M3D integration strategies and their quantitative benefits in power, area, and speed. We believe these additions provide the necessary macro-scale perspective requested by the reviewer.

On page 6, line 181

"Leveraging the potential for monolithic integration, recent advancements in M3D technology for OS-based GC-eDRAM have yielded breakthrough achievements in mitigating memory bottlenecks. Empirical evidence suggests that the refresh overhead of OS-based gain cells reaches a point of saturation once retention time approaches approximately 10 seconds. This implies that subsequent design optimization should prioritize system-level efficiency rather than exclusively pursuing the maximization of retention time (as shown in Fig. 3a)[51]. Consistent with this paradigm, a state-of-the-art M3D-SD (M3D SRAM and DRAM) architecture has been demonstrated, featuring the vertical stacking

of IGZO pass gates and IGZO 2T0C DRAM directly atop Si CMOS (Fig. 3b)[52]. This configuration achieves a compact 4T footprint and a 51% reduction in static power, while simultaneously facilitating ultra-low latency data transfer (< 10 ns) and robust retention characteristics exceeding 5,000 seconds. Furthermore, the introduction of an n/p hybrid structure utilizing a counteractive coupling mechanism has proven effective in neutralizing charge injection noise, thereby ensuring enhanced data stability (Fig. 3c)[53]. By capitalizing on this improved effective sensing margin, the architecture delivers a 48.25-fold enhancement in processing speed for edge AI workloads relative to conventional planar counterparts. The successful proliferation of GC-eDRAM, however, remains contingent upon the development of foundational technologies that can implement high-performance, uniform OS transistors and circuits within stringent low-temperature thermal budgets. This requirement is substantiated by the realization of a 2-kbit array based on body-contacted IGZO FETs (Fig. 3d)[54]. Fabricated at 350 °C–a temperature with BEOL Cu-interconnect processes–this array maintains highly uniform performance and has been utilized to demonstrate an all-nMOS analog memory circuit functioning as a short-term memory (STM) synapse. Nevertheless, as architectural and material innovations continue to push the boundaries of density, further device miniaturization inevitably encounters emerging physical limitations.”

On page 20, line 629

52. Liu, M. et al. First demonstration of monolithic three-dimensional integration of ultra-high density hybrid IGZO/Si SRAM and IGZO 2T0C DRAM achieving record-low latency (<10ns), record-low energy (5000s). in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* 1–2 (IEEE, Honolulu, HI, USA, 2024). doi:10.1109/VLSITechnologyandCir46783.2024.10631551.

53. Su, Y. et al. Monolithic 3-D integration of counteractive coupling IGZO/CNT hybrid 2T0C DRAM and analog RRAM-based computing-in-memory. *IEEE Trans. Electron Devices* **71**, 3336–3342 (2024).

54. Chand, U. et al. 2-kbit array of 3-D monolithically-stacked IGZO FETs with low SS-64mV/dec, ultra-low-leakage, competitive μ-57 cm2/V-s performance and novel nMOS-only circuit demonstration. in *2021 Symposium on VLSI Technology (VLSI Technology)* 1-2 (IEEE, Kyoto, Japan, 2021).

On page 25, line 767

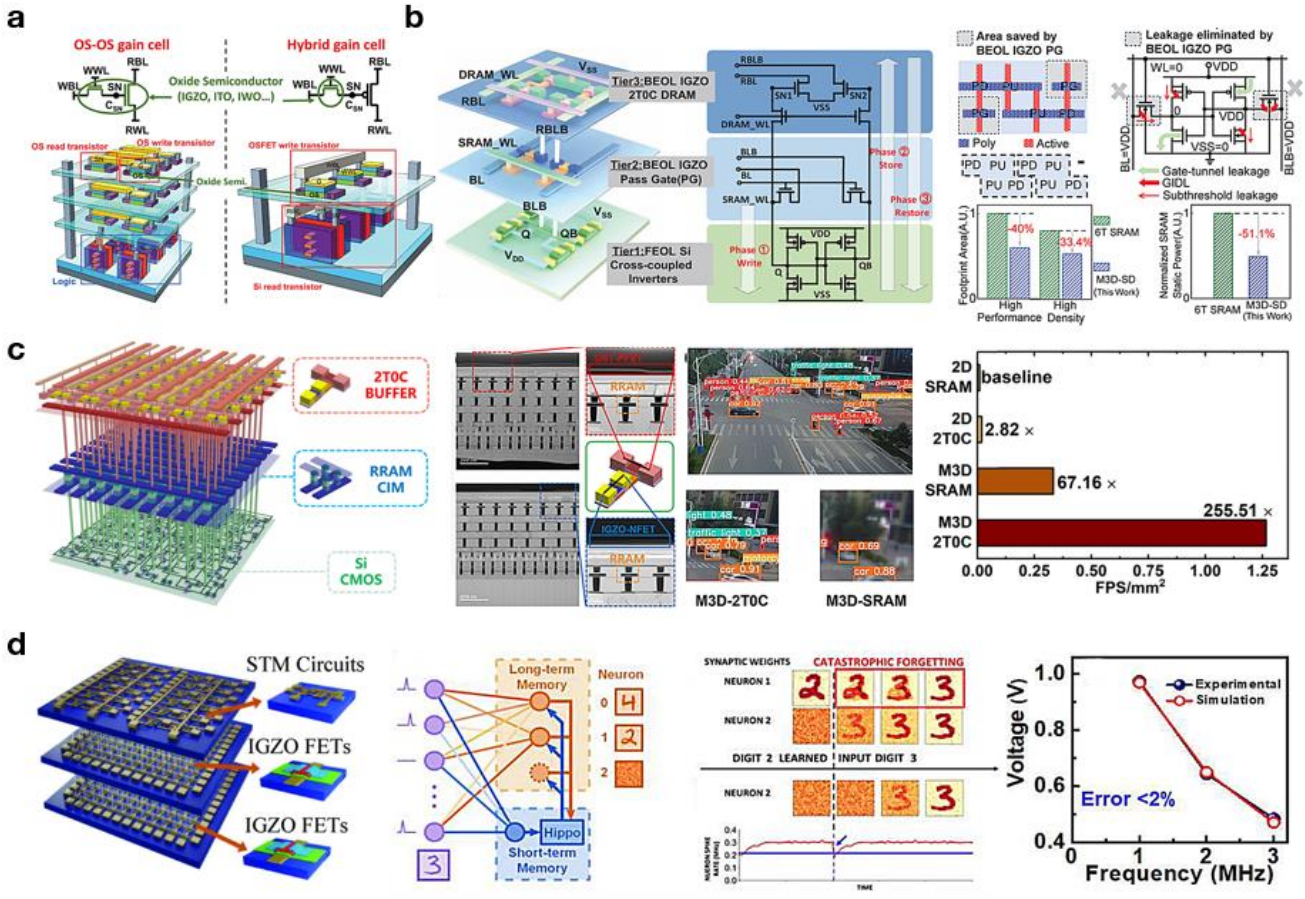

Fig. 3 | Application Gain Cell Structure Based on Oxide Semiconductor: a Monolithic integration schemes with logic, illustrating both all-oxide semiconductor (OS–OS) and hybrid (OS–Si) gain cell structures for high-density memory[51]. **b** M3D-SD Architecture: 3-tier monolithic stacking of Si CMOS, IGZO pass gates, and IGZO 2T0C DRAM (left), and analysis of area savings and static power reduction achieved by the BEOL IGZO pass gate compared to conventional SRAM (right)[52]. **c** M3D-BRIC Architecture: Vertical integration of Si logic, resistive random-access memory (RRAM) CIM, and 2T0C DRAM buffer, including cross-sectional TEM views (left), and benchmarking results demonstrating superior object detection rates and processing efficiency (fps/mm^2) of the M3D-2T0C configuration compared to M3D-SRAM (right)[53]. **d** Neuromorphic Application: Illustration of the monolithically 3D-stacked IGZO FET array (left), proposed dual memory model integrating Long-Term and Short-Term Memory (LTM/STM) to mitigate catastrophic forgetting (middle) and experimental demonstration of the body-contacted IGZO STM synapse, showing that higher neuron firing frequencies increase synaptic conductance (right)[54].

1. Rows 97-99 it is written: “Additionally, two other leakage currents exist regardless of the device conduction state: pn junction reverse bias leakage ($I_{Junction}$) and gate tunneling current (I_{Tunnel}), both of which result from the band-to-band tunneling effect.”. I believe the gate tunneling depends primarily on the gate stack composition and structure, and only in second order on the channel material. A

clarification on this aspect would improve the accuracy of the discussion.

Answer: We sincerely thank the reviewer for this perceptive comment regarding the physical origins of leakage currents. We agree that our previous phrasing was imprecise, particularly in attributing both $I_{Junction}$ and I_{Tunnel} solely to band-to-band tunneling and implying a primary dependence on the channel material for both. To improve the technical accuracy of our discussion, we have implemented the following changes:

- **Section Title Refinement:** We have renamed the subsection from "Leakage Current Issues Due to Silicon-Based Channel Materials" to **"Leakage Current Paths of Silicon-Based MOSFETs"** to more accurately reflect the diverse origins of these leakages beyond just the channel material.
- **Mechanism Clarification:** We have explicitly decoupled the discussion of I_{Tunnel} from $I_{Junction}$. We now clarify that while $I_{Junction}$ is influenced by mechanisms such as band-to-band tunneling within the channel/junction region, **gate tunneling (I_{Tunnel})** is fundamentally governed by the **gate stack properties**—including material composition, thickness, and structure—rather than the intrinsic properties of the channel material.

We believe these revisions resolve the ambiguity and align the manuscript with the established principles of device physics.

On page 4, line 116
"Leakage Current Paths of Silicon-Based MOSFET"

On page 5, line 121
"Additionally, two other leakage current paths persist regardless of the device's conduction state: pn junction reverse-bias leakage ($I_{Junction}$) and gate tunneling current (I_{Tunnel}). The $I_{Junction}$ component arises from multiple mechanisms, including generation–recombination, diffusion, and band-to-band tunneling within the junction depletion region. Conversely, the gate tunneling current is primarily a function of the gate stack architecture, determined by the dielectric material, interfacial layers, and electrode composition, rather than being a direct consequence of the channel material properties."

2. Following the previous point, the manuscript does not discuss the gate dielectric interface and material considerations in oxide semiconductor TFTs. Since the interface quality between the insulator and the semiconductor plays a crucial role in determining device stability and reliability, I suggest expanding the discussion in this regard, particularly within the section addressing reliability and technological challenges.

Answer: We sincerely appreciate the reviewer's constructive suggestion to expand on the **gate dielectric/semiconductor interface**. We agree that the chemical and electronic integrity of this interface is a decisive factor in determining the long-term stability and reliability of OS-based TFTs. To address this, we have substantively expanded the section titled **"Reliability and Technological Challenges of OS-based GC-eDRAM."** The revised discussion now provides a comprehensive analysis of:

- **Material-level trade-offs:** The challenges associated with high-k dielectrics (e.g., HfO_2, ZrO_2), specifically regarding **crystallization-induced leakage pathways** and grain-boundary-mediated instability.
- **Interface Engineering Strategies:** The implementation of **bilayer gate stacks** (e.g., Al_2O_3 or SiO_2 interlayers) and **nanolaminates** to effectively suppress interface trap density (D_{it}) and stabilize the amorphous phase [89–91].
- **Surface Passivation & Plasma Treatment:** The critical role of optimized **pre-deposition surface treatments** (e.g., O_2 plasma) and the introduction of **protective layers** to mitigate plasma-induced damage and passivate oxygen vacancies (V_O) [92, 93].

We believe these additions, supported by the newly included **Figure 6**, provide the rigorous analysis of dielectric-interface interactions requested by the reviewer.

On page 13, line 405

'In gain cell memory architectures, periodic refresh operations inherently constrain the effective bandwidth. Consequently, minimizing leakage currents while securing sufficient storage capacitance is imperative to maximize data retention time. To address these conflicting requirements, high-k materials are widely employed as gate dielectrics. However, the inherent propensity of high-k dielectrics particularly HfO_2- and ZrO_2-based materials to crystallize leads to an increase in bulk and interface trap densities. This crystallization facilitates the formation of leakage pathways along grain boundaries, subsequently inducing elevated leakage currents, severe bias instability, and significant hysteresis.

To address these reliability issues, advanced gate stack engineering strategies have been investigated to simultaneously mitigate interfacial instability[89,90] and inhibit bulk crystallization[91]. Specifically, to mitigate interface trap density and secure sufficient band offsets, the insertion of thin Al_2O_3 (Fig. 6a)[89] and SiO_2 (Fig. 6b)[90] interlayers into channel/high-k interface has been investigated. These interlayers effectively reduce the interface-state density (D_{it}) and increase the conduction-band offset, thereby suppressing current-voltage (I-V) hysteresis and simultaneously enhancing stability. In parallel, to inhibit crystallization-induced degradation, the incorporation of Al_2O_3 has been employed to stabilize the amorphous phase of high-k dielectrics. Figure 6c shows that HfO_2/Al_2O_3 nanolaminates, in particular, preserve an amorphous microstructure even under thermal and electrical stress[91].

Consequently, these stacks significantly reduce leakage current and grain-boundary-mediated trap-assisted conduction, while also improving hysteresis characteristics compared to single-layer HfO_2. Beyond dielectric material engineering, interface treatment prior to gate insulator deposition is critical for reliability[92,93]. Optimized O_2 plasma treatment enhances the surface density of hydroxyl (-OH) groups on oxide semiconductor channels, improving precursor adsorption during subsequent ALD and markedly reducing interface-related traps (Fig. 6d)[92]. However, excessive plasma exposure can degrade interface quality by inducing surface roughness and generating plasma-induced traps, leading to device instability. To mitigate these adverse effects, the introduction of a thin Al_2O_3 protective layer effectively suppresses plasma damage and passivates oxygen vacancies (V_O), resulting in significantly enhanced device reliability (Fig. 6e)[93]. Consequently, the synergistic integration of crystallization-suppressing nanolaminates, functional interlayers, and optimized surface passivation serves as a comprehensive strategy to overcome the intrinsic limitations of high-k, thereby securing the low leakage current and high stability required for extended-retention GC-eDRAMs."

On page 22, line 720

89. Cho, M. H., Choi, C. H., Seul, H. J., Cho, H. C. & Jeong, J. K. Achieving a low-voltage, high-mobility IGZO transistor through an ALD-derived bilayer channel and a hafnia-based gate dielectric stack. *ACS Appl. Mater. Interfaces* **13**, 16628–16640 (2021).

90. Choi, C. H., Kim, T., Kim, M. J., Yoon, S. H. & Jeong, J. K. Mechanism of external stress instability in plasma-enhanced ALD-derived HfO2 /IGZO thin-film transistors. *IEEE Trans. Electron Devices* **70**, 2317–2323 (2023).

91. Ko, S. H. & Yoon, S. M. Interface-engineered HfO2 /Al2O3 nanolaminate gate insulators for InGaZnO thin-film transistors and 2T0C DRAM applications. *ACS Appl. Electron. Mater.* **7**, 8100–8107 (2025).

92. Eun Oh, J. et al. In-situ surface energy engineering for ALD-derived highly reliable top gate In2O3 thin-film transistors. *IEEE Electron Device Lett.* **46**, 2050–2053 (2025).

93. Shin, S. et al. Suppression of interface damage in IGZO TFTs using Al2O3 protective layer during N2O plasma treatment. *Applied Surface Science* **718**, 164912 (2026).

3. Row 109, the acronym VCT is never defined.

Answer: We thank the reviewer for pointing this out. The acronym "VCT" stands for "Vertical Channel Transistor," and we have now explicitly defined it at its first occurrence in the revised manuscript.

On page 5, line 137

"vertical channel transistors (VCTs)"

4. It would be valuable to include a brief discussion on threshold voltage instability and hysteresis phenomena in oxide-based transistors, as these are key reliability concerns for such devices.

Answer: We sincerely thank the reviewer for highlighting the importance of **threshold voltage (V_{TH}) instability** and **hysteresis phenomena**. We agree that these reliability metrics are paramount for the practical deployment of oxide-semiconductor (OS)-based memory. In the revised manuscript, we have introduced a comprehensive reliability framework within the **"Reliability of AOS Transistors"** section. This discussion elucidates the physical mechanisms governing V_{TH} shifts—specifically the charge-state transitions of oxygen vacancies (V_O) under electrical stress—and evaluates extrinsic strategies to mitigate these effects, such as advanced gate-stack engineering and interfacial passivation. We believe these additions significantly strengthen the technical rigor of our analysis.

On page 10, line 320

"**Reliability of AOS Transistors**

Under prolong positive-bias temperature stress (PBTS) and negative-bias temperature stress (NBTS), AOS transistors exhibit pronounced V_{TH} instabilities, which can be categorized into intrinsic bulk defects within amorphous OS channel and extrinsic defects associated with the gate stack and passivation layer[68,78]. The primary intrinsic defects in OS channel are V_O and interstitial O (O_i), acting as donor-like and acceptor-like trap, respectively. During PBTS (or NBTS), the quasi-Fermi level (E_F) shifts toward conduction band edge (E_C), triggering a charge-state transition of V_O. Specifically, the conversion from doubly ionized V_O^{2+} to neutral V_O (and vice-versa) involves the consumption or generation of free electrons, manifesting as a significant positive or negative shift in V_{TH}.[68,78] Consequently, suppressing the formation of V_O is critical for mitigating BTS-induced V_{TH} instability. Such suppression can be achieved by increasing Ga content and oxygen incorporation; however, this approach often degrades the I_{ON} due to the combined effect of reduced electron density and weakened percolation conduction. Gate-insulator-dependent BTS instabilities are discussed separately in the subsequent section."

On page 13, line 412

"To address these reliability issues, advanced gate stack engineering strategies have been investigated to simultaneously mitigate interfacial instability[89,90] and inhibit bulk crystallization[91]. Specifically, to mitigate interface trap density and secure sufficient band offsets, the insertion of thin Al_2O_3 (Fig. 6a)[89] and SiO_2 (Fig. 6b)[90] interlayers into channel/high-k interface has been investigated. These interlayers effectively reduce the interface-state density (D_{it}) and increase the conduction-band offset, thereby suppressing current-voltage (I-V) hysteresis and simultaneously enhancing stability. In parallel, to

inhibit crystallization-induced degradation, the incorporation of Al_2O_3 has been employed to stabilize the amorphous phase of high-k dielectrics. Figure 6c shows that HfO_2/Al_2O_3 nanolaminates, in particular, preserve an amorphous microstructure even under thermal and electrical stress[91]. Consequently, these stacks significantly reduce leakage current and grain-boundary-mediated trap-assisted conduction, while also improving hysteresis characteristics compared to single-layer HfO_2. Beyond dielectric material engineering, interface treatment prior to gate insulator deposition is critical for reliability[92,93]. Optimized O_2 plasma treatment enhances the surface density of hydroxyl (-OH) groups on oxide semiconductor channels, improving precursor adsorption during subsequent ALD and markedly reducing interface-related traps (Fig. 6d)[92]. However, excessive plasma exposure can degrade interface quality by inducing surface roughness and generating plasma-induced traps, leading to device instability. To mitigate these adverse effects, the introduction of a thin Al_2O_3 protective layer effectively suppresses plasma damage and passivates oxygen vacancies (V_O), resulting in significantly enhanced device reliability (Fig. 6e)[93]. Consequently, the synergistic integration of crystallization-suppressing nanolaminates, functional interlayers, and optimized surface passivation serves as a comprehensive strategy to overcome the intrinsic limitations of high-k, thereby securing the low leakage current and high stability required for extended-retention GC-eDRAMs.”

5. In the “Crystallization-driven advancements in oxide channel” section I suggest to add a discussion regarding the additional complexity introduced by the crystallization process, which often includes high temperature annealing.

Answer: We appreciate the reviewer’s suggestion to address the procedural complexity associated with the crystallization of oxide channels. We agree that high-temperature annealing is a critical consideration, particularly in the context of **BEOL compatibility** and monolithic 3D integration. However, we would like to clarify that oxide semiconductors offer a distinct advantage over silicon in this regard: their crystallization temperatures are inherently lower, often falling within the stringent thermal budgets of modern logic platforms. In the revised manuscript, we have expanded the discussion to emphasize that:

1. **Lower Thermal Budget:** Oxide channels can achieve high crystallinity at temperatures significantly lower than those required for polycrystalline silicon, mitigating the risk of damage to underlying CMOS layers.
2. **Advanced Low-Temp Techniques:** The use of specialized methods such as **Metal-Induced Crystallization (MIC)** further reduces the required thermal input, minimizing process complexity while enhancing device performance.

We believe this clarification highlights a key technological driver for adopting oxide-based channels in high-density memory architectures.

On page 12, line 371

"From an integration perspective, the relatively lower crystallization temperatures of OSs compared to silicon ensure superior compatibility with subsequent fabrication steps and multi-layer stacking. While conventional silicon crystallization often demands high-temperature annealing that can exceed the thermal budget of BEOL processes, OS channels can achieve robust crystallinity at significantly reduced temperatures. Furthermore, by employing advanced crystallization techniques such as Metal-Induced Crystallization (MIC), OS channels can be synthesized at even lower thermal thresholds than those typically required for silicon-based counterparts[73]. Consequently, the procedural complexity and thermal stress introduced by the crystallization process are comparatively diminished in OS-based architectures. This thermal advantage facilitates the seamless monolithic 3D integration of high-performance OS-based gain cells atop standard Si-CMOS logic, without compromising the integrity of underlying devices or interconnects."

6. In the "Negative Capacitance: Beyond Boltzmann's Tyranny" section, it is discussed the benefit of using ferroelectric materials in the gate stack for transistors. This concept is not new and not exclusive of oxide-based TFT. In addition, the polarization curve of the ferroelectric layer introduces threshold drift in the transistor and hysteresis in the characteristic, which is not desirable. Add this in the discussion or remove the section.

Answer: We appreciate the reviewer's critical assessment regarding the inclusion of **negative capacitance (NC)** within the eDRAM context. Upon further reflection, we agree that while NC effects in ferroelectric-gate transistors are a significant area of research, they are not central to the oxide-based TFT gain-cell architectures presented in this work. As the reviewer correctly noted, the polarization switching and hysteresis inherent in ferroelectric materials can introduce instability in threshold voltage (V_{th}), which may compromise the reliable operation and sensing margins required for eDRAM applications. Since our manuscript focuses on achieving high-performance and stable gain-cells without the integration of ferroelectric layers, the discussion on NC was indeed extraneous.
To streamline the manuscript and maintain a rigorous focus on our primary contribution—**oxide-based monolithic 3D integration**—we have **removed the section concerning negative capacitance** in the revised version. This allows for a more cohesive narrative centered on the material and architectural innovations of OS-based memories.

7. I suggest adding a section (or including it in the final section) regarding current trends and challenges

for GC DRAM and their realization with oxide semiconductors.

Answer: We sincerely appreciate the reviewer's strategic suggestion to provide a more cohesive overview of the current trends and challenges. While our original manuscript touched upon various aspects of oxide-based GC-eDRAM, we agree that a consolidated discussion on the **technological roadmap and remaining hurdles** enhances the manuscript's impact. In the revised version, we have significantly expanded the **"Summary and Outlook"** section. This new discussion explicitly categorizes the challenges into three hierarchical levels:

1. **Device Level:** Balancing high mobility with long-term bias stability and managing hydrogen-induced degradation.
2. **Process Level:** Developing BEOL-compatible High-k Metal Gate (HKMG) interfaces within stringent thermal budgets to enable reliable monolithic 3D (M3D) integration.
3. **System/Array Level:** Managing array-level variability, such as retention dispersion and read/write disturb, while navigating the increased circuit complexity inherent in the 4-terminal gain-cell configuration.

We believe this structured outlook provides a clear framework for future research in the field of oxide semiconductor-based memory.

On page 15, line 481

'Despite the remarkable progress in OS-based GC-eDRAM, several critical challenges must be addressed to facilitate large-scale commercialization. First, device-level variability driven by intrinsic defects and bias-instability poses a significant threat to multi-bit operation and sensing margins. This necessitates rigorous control over material stoichiometry and interface passivation, alongside the development of accurate compact models for predictive circuit design. Furthermore, optimizing the delicate trade-off between carrier mobility and reliability remains a central hurdle, particularly under high-temperature operating conditions where H sensitivity and V_O migration become more pronounced. Second, the realization of true M3D integration requires BEOL-compatible processes that can deliver high-performance HKMG interfaces without exceeding the thermal limits of the underlying CMOS logic. This requires innovations in low-temperature ALD processes and advanced crystallization techniques that do not compromise the integrity of the vertical interconnects. Third, at the array level, the industry must address reliability concerns such as retention dispersion and read/write disturb under realistic workloads. Unlike the simple 2-terminal capacitor in 1T1C DRAM, the multi-terminal nature of gain-cells introduces increased routing and driving complexity. Managing this complexity requires a synergistic approach, combining architectural techniques such as intelligent refresh scheduling and error-correcting codes (ECC) with the material-level advancements discussed herein. Ultimately, the successful deployment of OS-based GC-eDRAM will depend on a holistic co-optimization of materials,

devices, and architectures to bridge the gap between lab-scale demonstration and industry-standard reliability.”

Answer to the comments by Reviewer #3 and Summary of changes

We are grateful to the Reviewer for his/her kind comments and corrections on our manuscript. After considering the comments, we revised our manuscript and made the following reply.

1. In line 66, the author state that “eDRAM not only serves as a cache but also acts as a buffer layer, mitigating the latency gap between CPU and DRAM.” However, this description is somewhat confusing, since a cache itself is inherently designed to bridge the latency gap.

Answer: We appreciate the reviewer’s astute observation regarding the functional definition of caches. We agree that the term "buffer layer" was used somewhat ambiguously in the original text, as bridging the latency gap is indeed the fundamental purpose of any cache hierarchy. To clarify our intended meaning, we have revised the description to emphasize eDRAM’s role as a **high-capacity intermediary** that complements traditional SRAM-based hierarchies. Unlike standard SRAM L1-L3 caches, which are limited by area efficiency, eDRAM’s higher density allows it to act as a massive **"memory buffer"** or **"Level-4 (L4) cache"** that can host entire working sets. This significantly reduces the frequency of off-chip DRAM accesses (miss penalty), thereby alleviating the **bandwidth-latency bottleneck** that SRAM caches cannot address alone. The revised text now explicitly distinguishes eDRAM’s unique architectural contribution from conventional caching.

On page 3, line 66

“Beyond its role as a last-level cache (LLC), eDRAM acts as pivotal high-density memory buffer bridging the gap between processing cores and off-chip main memory. By hosting significantly large working sets, it effectively alleviates the bandwidth-latency bottleneck inherent in traditional SRAM-limited hierarchies[25].”

2. The author claims that “conventional 1T1C eDRAM adopts a continuous scaling process for performance and density improvement; however, this leads to a reduction in the physical size of the capacitor and an increase in transistor leakage current, resulting in degraded data retention time and limited scalability. To address these issues, gain cell-embedded DRAM (GC-eDRAM), which removes the capacitor and consists of 2-4 transistors, has garnered attention.” However, this description does not convincingly explain the advantages of GC-eDRAM over conventional 1T1C eDRAM. As device dimensions scale down, the storage capacitance of the gain cell (i.e., the gate capacitance of the read transistor) also decreases, while leakage current increases—similar to the challenges faced by 1T1C cells. Therefore, the scaling issue remains essentially the same. The true advantages of GC-eDRAM lie in its non-destructive read operation and process compatibility, whereas 1T1C eDRAM suffers from the

difficulties of fabricating high–aspect-ratio capacitors and poor integration compatibility with logic processes.

Answer: We sincerely thank the reviewer for this insightful and technically rigorous critique. We agree that our previous description oversimplified the scaling advantages of GC-eDRAM by implying it inherently bypasses the capacitance-scaling dilemma. As the reviewer correctly pointed out, the reduction in gate capacitance during scaling presents challenges analogous to those in 1T1C architectures. To improve the accuracy of our discussion, we have overhauled the introductory logic to emphasize that the superiority of GC-eDRAM is rooted in **architectural and manufacturing advantages** rather than a mere avoidance of capacitance reduction. The revised manuscript now explicitly highlights:

1. **Non-destructive Read Operation:** Unlike 1T1C DRAM, which requires energy-intensive write-back cycles after every read, the GC-eDRAM's decoupled read path allows for sensing without charge loss. This fundamentally reduces dynamic power and refresh overhead.
2. **Elimination of HAR (High-Aspect-Ratio) Bottlenecks:** We now clarify that the primary "scaling" advantage of GC-eDRAM is the removal of the requirement for complex, high- aspect-ratio storage capacitors, which are the dominant manufacturing bottleneck for 1T1C integration in advanced logic nodes.
3. **Monolithic 3D BEOL Compatibility:** We emphasize that GC-eDRAM's transistor-only structure makes it a prime candidate for **3D BEOL stacking**, a pathway for density improvement that is significantly more challenging for capacitor-based 1T1C cells.

We believe these clarifications provide a much more robust and accurate justification for the shift toward GC-eDRAM technologies.

On page 3, line 69

"However, conventional 1T1C eDRAM faces substantial manufacturing challenges associated with the fabrication of high-aspect-ratio (HAR) storage capacitors and exhibits limited integration compatibility with standard logic processes, both of which hinder further density scaling[26,27]. To alleviate these process-related constraints, gain cell-embedded DRAM (GC-eDRAM) has emerged as a compelling alternative. By eliminating the discrete storage capacitor and implementing the cell using a transistor-only (2T-4T) configuration, GC-eDRAM achieves superior CMOS process compatibility and facilitates high-density scaling via 3D back-end-of-line (BEOL) stacking[24,28,29]. Furthermore, GC-eDRAM also adopts a decoupled read path, in which the storage node is isolated from the read bitline and is sensed through capacitive coupling via the read-transistor gate dielectric. This configuration enables non-destructive readout without directly discharging the stored charge. Consequently, GC-eDRAM reduces (or potentially eliminates) write-back overhead, mitigates refresh and dynamic-power penalties, and

improves sensing margin[29]. These attributes are particularly beneficial for compute-in-memory (CIM) and processing-in-memory (PIM) workloads that require iterative access and repeated analog/logic operations[6,27,30].

Nevertheless, silicon-channel GC-eDRAM still suffers from limited retention time (~100 μs) and non-negligible refresh-related power consumption when considered as a mainstream memory solution[31]. Accordingly, this study systematically analyzes the intrinsic limitations of silicon-based GC-eDRAM and explores oxide-semiconductor (OS) channels as a pathway to ultra-low leakage and enhanced BEOL integrability, thereby proposing device- and process-level optimization strategies for next-generation embedded DRAM."

On page 18, line 572

26. Chun, K. C., Jain, P., Kim, T.-H. & Kim, C. H. A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches. *IEEE J. Solid-State Circuits* **47**, 547–559 (2012).

29. Teman, A., Meinerzhagen, P., Burg, A. & Fish, A. Review and classification of gain cell eDRAM implementations. in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel* 1–5 (IEEE, Eilat, Israel, 2012). doi:10.1109/EEEI.2012.6377022.

30. Yu, C. et al. A logic-compatible eDRAM compute-in-memory with embedded ADCs for processing neural networks. *IEEE Trans. Circuits Syst. I* **68**, 667–679 (2021).

3. In line 83, the authors state that "Above mentioned, the GC-eDRAM structure can alleviate the complexity and cost issues in manufacturing processes arising from the high capacitance requirements of the 1T1C eDRAM." However, no such discussion or supporting explanation is actually provided in the preceding text.

Answer: We sincerely apologize for the lack of preceding context regarding the manufacturing advantages of GC-eDRAM. We agree that the claim regarding cost and complexity reduction lacked sufficient supporting arguments in the earlier sections of the manuscript. To rectify this, we have restructured the introductory section (Page 3) to provide a clear technical comparison. Specifically, we have added a discussion on the **fabrication bottlenecks of 1T1C eDRAM**, such as the necessity for complex, high-aspect-ratio (HAR) capacitor structures that require specialized materials and etching processes not typically found in standard logic CMOS flows. By contrast, we now explicitly state that GC-eDRAM utilizes **standard transistor modules**, which inherently reduces process complexity and enhances **BEOL integrability**.

On page 3, line 69

"However, conventional 1T1C eDRAM faces substantial manufacturing challenges associated with the fabrication of high-aspect-ratio (HAR) storage capacitors and exhibits limited integration compatibility with standard logic processes, both of which hinder further density scaling[26,27]. To alleviate these process-related constraints, gain cell-embedded DRAM (GC-eDRAM) has emerged as a compelling alternative. By eliminating the discrete storage capacitor and implementing the cell using a transistor-only (2T-4T) configuration, GC-eDRAM achieves superior CMOS process compatibility and facilitates high-density scaling via 3D back-end-of-line (BEOL) stacking[24,28,29]."

4. The authors state that there are three main causes for leakage current: leakage current issues due to silicon-based channel materials, short-channel effects (SCE), and the floating-body effect. However, fundamentally, both the short-channel effect and the floating-body effect are manifestations of leakage current issues inherent to silicon-based channel materials, rather than independent categories. These two effects simply aggravate channel leakage instead of representing distinct leakage mechanisms.

Answer: We sincerely thank the reviewer for this insightful clarification on the physical hierarchy of leakage mechanisms. We entirely agree that **Short-Channel Effects (SCE)** and the **Floating-Body Effect** are not independent leakage mechanisms but are instead structural and scaling-related phenomena that exacerbate the fundamental leakage paths inherent in silicon MOSFETs. To align the manuscript with this accurate physical framework, we have restructured the discussion to distinguish between the **intrinsic leakage paths** and the **extrinsic factors** that aggravate them. Specifically:

1. We have reclassified the sections to reflect that SCE and floating-body effects act as **scaling- and structure-induced aggravations** of the primary leakage components.
2. The subsection titles have been revised to establish a clear hierarchy:

   (i) **Leakage Current Paths of Silicon-Based MOSFETs** (Fundamental mechanisms)

   (ii) **Scaling-Induced Aggravation: Short-Channel Effects (SCE)** (Mechanisms exacerbated by miniaturization)

   (iii) **Structure-Induced Aggravation: Floating-Body Effect** (Mechanisms exacerbated by SOI or specific gain-cell geometries).

This restructuring maintains the technical depth of our original analysis while providing a much more rigorous logical foundation.

On page 4, line 112

"These fundamental leakage paths are significantly exacerbated by scaling-induced short-channel effects (SCE) and structure-induced floating-body effects, which diminish electrostatic control and heighten the off-state conduction, as detailed in (i)-(iii) below[33-40].

(i) Leakage Current Paths of Silicon-Based MOSFET[33-35]

Silicon, with its narrow bandgap of approximately 1.12 eV, allows thermally excited electrons to easily cross the bandgap and generate current at high temperatures. This phenomenon deteriorates the data retention performance of DRAM. Figure 1a shows the prominent leakage current paths in a MOSFET. When the transistor is in a non-conducting state, three major leakage mechanisms are observed: subthreshold leakage current (I_{Sub}), punch-through leakage (I_{Punch}), and gate-induced drain leakage (I_{GIDL}). Additionally, two other leakage current paths persist regardless of the device's conduction state: pn junction reverse-bias leakage ($I_{Junction}$) and gate tunneling current (I_{Tunnel}). The $I_{Junction}$ component arises from multiple mechanisms, including generation–recombination, diffusion, and band-to-band tunneling within the junction depletion region. Conversely, the gate tunneling current is primarily a function of the gate stack architecture, determined by the dielectric material, interfacial layers, and electrode composition, rather than being a direct consequence of the channel material properties.

(ii) Scaling-Induced Aggravation: Short Channel Effects

Figure 1b shows that, as Si-based transistors are miniaturized, the reduction in channel length leads to decreased threshold voltage and increased drain-induced barrier lowering (DIBL) effects[36,37]. The gate loses effective control over the current flow, resulting in a dramatic increase in subthreshold leakage current. This phenomenon adversely affects data retention in DRAM cells and may cause cell-to-cell interference.

(iii) Structure-Induced Aggravation: Floating Body Effect

The floating body effect, particularly problematic in silicon-on-insulator (SOI) DRAMs, causes instability in retention charge. The accumulation of holes in the channel induces unexpected potential changes[38,40], negatively impacting data retention as shown in Fig. 1c. Such reliability challenges are not confined to SOI DRAMs and also occur in Si devices such as DRAM-oriented vertical channel transistors (VCTs)[41].”

5. In line 115, the first “i” in “indium gallium zinc oxide (IGZO)” appears in red. The authors should pay closer attention to such formatting issues.

Answer: We sincerely apologize for this formatting oversight. We appreciate the reviewer’s meticulous attention to detail in identifying this error. In the revised manuscript, we have corrected the font color of the initial "i" in **"indium gallium zinc oxide (IGZO)"** to the standard black text. Furthermore, we have conducted a comprehensive review of the entire manuscript to ensure that all formatting, including font styles, colors, and superscripts/subscripts, is consistent and meets the journal's professional

standards. We thank the reviewer for helping us improve the overall quality of the presentation.

6. In the “Application of Oxide Semiconductors to GC-eDRAM Architectures” section, the authors discuss the use of GC-eDRAM for computing-in-memory (CIM). However, there are already CIM applications based on silicon-based GC-eDRAM. Therefore, this section currently reads more as a discussion of GC-eDRAM applications in general, rather than highlighting the advantages of oxide semiconductor-based GC-eDRAM. The authors should focus more on describing applications that oxide semiconductor GC-eDRAM can enable, which conventional GC-eDRAM cannot.

Answer: We sincerely thank the reviewer for this insightful comment. We agree that the distinction between generic GC-eDRAM applications and those specifically enabled by oxide semiconductors (OS) needed to be more pronounced. In the revised manuscript, we have shifted the focus to highlight functionalities that are **physically unattainable with conventional Si-based GC-eDRAM** due to its inherent leakage and thermal constraints. Specifically, we emphasize that OS-based GC-eDRAM enables:

1. **High-Precision Analog Computing:** The ultralow off-current (< 10^{-18} A/μm) of OS allows for **quasi-nonvolatile retention (>1000 s)**, which is critical for maintaining stable analog weight voltages. In Si-based cells, rapid charge leakage causes "weight decay," severely degrading the accuracy of Multiply-Accumulate (MAC) operations.
2. **Multi-bit Weight Linearity:** The superior electrostatic control and low leakage enable a highly linear I_{OUT}-V_{IN} response, allowing for **accurate 4-bit (or higher) weight representation** within a single paired cell—a feat difficult to sustain in Si-DRAM without constant, energy-intensive refreshing.
3. **Monolithic 3D Logic-on-Memory:** The low-temperature BEOL compatibility (< 300 °C) of OS allows for the vertical integration of memory arrays directly atop CMOS logic. This **high-density 3D stacking** significantly reduces the physical distance between memory and compute units, enabling bandwidths and energy efficiencies that exceed the limits of planar Si-based CIM architectures.

The revised text now explicitly differentiates these OS-specific advantages from conventional Si-GC-eDRAM demonstrations.

On page 6, line 166

“Unlike conventional Si-based counterparts, OS 2T GC-DRAM provides a unique material foundation for high-precision analog CIM. The ultralow off-current of ~10^{-18} A/μm achieves a quasi-nonvolatile retention time exceeding 1000 s, which is several orders of magnitude higher than that of Si-based gain

cells. This extended retention is critical for suppressing analog weight decay during accumulation cycles, thereby preserving the computational precision required for multi-bit MAC operations without the overhead of frequent refreshing (Fig. 2c) [49].

Leveraging this stability, recent work demonstrated a convolutional neural network (CNN) for MNIST recognition using int4 weight encoding within paired OS GC-eDRAM cells (Fig. 2d)[50]. Such multi-bit resolution is facilitated by the highly linear and stable I_{OUT}-V_{IN} response of the OS channel. Crucially, the OS fabrication process compatible with BEOL thermal budgets below 300 °C enables the monolithic 3D integration of these high-density memory arrays directly atop Si-CMOS logic. This monolithic 3D integration circumvents the high-temperature processing requirements of Si-based channels, providing a scalable pathway for ultra-high-bandwidth, vertically stacked CIM fabrics that are physically unrealizable with conventional planar silicon technologies."

7. In the line 133, the authors state: "Motivated by these requirements, a GC-DRAM architecture was implemented in which the Wtr adopted an In-rich ITO channel to enhance electron transport and accelerate switching, whereas the Wtr employed an a-IGZO channel to suppress leakage owing to its superior off-state characteristics (Fig. 2b)" However, it is unclear why the Wtr would use both an In-rich ITO channel and an a-IGZO channel. The first "Wtr" should likely be "Rtr," as the current wording contains a basic error.

Answer: We appreciate the reviewer for carefully identifying this mistake. As correctly pointed out, the first "Wtr" in line 133 was a typographical error and should be "Rtr." We have corrected the sentence in the revised manuscript.

We sincerely thank the reviewer for identifying this fundamental typographical error. As correctly pointed out, the first mention of "**Wtr**" in line 133 was a mistake and should have been "**Rtr**" (Read Transistor). The core design strategy of our proposed architecture is to utilize a **heterogeneous dual-channel approach**:

1. **Rtr (Read Transistor):** Employs an **In-rich ITO channel** to leverage its high electron mobility, thereby enhancing transport and accelerating the read-switching speed.
2. **Wtr (Write Transistor):** Employs an **a-IGZO channel** specifically for its ultra-low off-state leakage, which is essential for maximizing data retention.

We have corrected this error in the revised manuscript to ensure technical accuracy and consistency with the provided figure (Fig. 2b). We appreciate the reviewer's meticulous reading.

On page 5, line 157

"Motivated by these requirements, a GC-DRAM architecture was implemented in which the Rtr adopted an In-rich ITO channel to enhance electron transport and accelerate switching, whereas the Wtr

employed an *a*-IGZO channel to suppress leakage owing to its superior off-state characteristics (Fig. 2b)[48].”

8. In line 137, the authors claim that the device performance “is comparable to that of quasi-nonvolatile memory.” I am curious whether there is a formal definition of “quasi-nonvolatile memory.” For example, nonvolatile memory typically requires a retention time exceeding 10 years, so what retention time would qualify as “quasi-nonvolatile”? If no formal definition exists, the authors should clarify or reconsider this description.

Answer: We appreciate the reviewer’s request for clarification regarding the term **“quasi-nonvolatile memory.”** We agree that while this term is frequently used in literature to describe memories with extended retention beyond traditional DRAM, it lacks a standardized industry definition (unlike the 10-year requirement for non-volatile memory). To ensure technical rigor and avoid ambiguity, we have removed this term from the revised manuscript. Instead, we now provide a **direct quantitative comparison** with conventional DRAM, emphasizing that a retention time exceeding **25 hours** represents a 10^5-to-10^6-fold improvement over the millisecond-range retention of standard silicon-based DRAM. This revision highlights the device's capability for **extended-retention applications** and **reduced refresh overhead** without relying on ill-defined terminology.

On page 6, line 163
“This configuration achieved stable operation with a 10-ns write pulse and an exceptional data-retention capability exceeding 25 hours, which significantly surpasses the millisecond-range retention of conventional Si-DRAM, effectively enabling extended-retention memory application and drastically reducing the power penalty associated with periodic refresh operations. This performance profile aligns with the requirements of quasi-nonvolatile memory, offering a compelling balance between energy-efficient programming and long-term data persistence.”

9. In the “Scaling Beyond Limits: Multi-Gate Transistors” section, the authors devote more than half of the content to describing basic semiconductor scaling trends, from planar MOSFETs to FinFETs to GAA structures. These are fundamental concepts and should not occupy such a large portion of the section.

Answer: We sincerely thank the reviewer for this constructive critique regarding the content balance. We agree that a prolonged description of historical scaling trends (Planar-to-FinFET-to-GAA) may detract from the core focus of our manuscript. To address this, we have substantively condensed the background information and shifted the emphasis toward the **electrostatic and structural advantages**

of oxide semiconductor (OS)-based Gate-All-Around (GAA) architectures. Specifically:

1. **Contextual Background:** We now briefly mention the transition to multi-gate structures as a necessary evolution for electrostatic control, without detailed historical exposition.
2. **Focus on Natural Length (λ):** We have centered the discussion on the **Natural Length (λ)** to provide a quantitative comparison. This demonstrates how OS-based GAA configurations achieve superior scaling potential by minimizing λ compared to traditional single-gate or FinFET structures.
3. **Application to eDRAM:** We explicitly link these electrostatic benefits to the improvement of **data retention and scalability** in OS-based GC-eDRAM.

We believe this revision streamlines the section while providing more advanced, relevant insights for the reader.

On page 7, line 203

"The continued scaling of OS-based DRAM necessitates advanced architectures to combat the degradation of subthreshold swing and the aggravation of SCE. The natural length (λ), the characteristic length representing the gate's electrostatic control over the channel serves as the primary metric for SCE immunity; a smaller λ indicates enhanced gate controllability and suppressed DIBL. To achieve the minimum possible λ, the industry has transitioned from planar MOSFETs to multi-gate architectures. While FinFETs improved gate coupling, the Gate-All-Around (GAA) surrounding-gate structure represents the ultimate limit of electrostatic control (Fig. 4a)[55,56]. For a 10-nm channel thickness and a 2-nm gate insulator, the surrounding-gate design is shown to reduce λ by approximately 0.56× compared to single-gate and 0.82× relative to FinFET configurations. This superior control is particularly critical for OS-based gain cells, where minimizing λ is essential to maintaining ultra-low leakage and ensuring stable data retention as the gate length shrinks toward the sub-10 nm regime[56]."

10. In the line 181, the author state: "The surrounding-gate structure enhances gate-to-channel capacitive coupling, leading to an increase in parasitic capacitance. However, this effect paradoxically improves data retention and sensing performance, offering distinct advantages in eDRAM applications". However, in GC-eDRAM, the storage relies on the gate capacitance of the read transistor. An increase in the parasitic capacitance of the write transistor would similarly degrade performance due to the pull-down effect when the write transistor is turned off.

Answer: We sincerely appreciate the reviewer's rigorous technical insight regarding the potential drawbacks of parasitic capacitance. We agree that an indiscriminate increase in parasitic coupling—particularly between the **Write Word-Line (WWL)** and the **Storage Node (SN)**—can induce a significant voltage disturbance (ΔV_{SN}) due to the pull-down effect during the write transistor's turn-off

transition. In the revised manuscript, we have addressed this by clarifying that while the surrounding-gate structure is essential for maximizing the **Storage Capacitance (C_{SN})** and suppressing leakage, it inherently introduces a **design trade-off** related to parasitic coupling (C_{WWL-SN}, C_{RWL-SN}). We now emphasize that:

- **Storage vs. Parasitic Balance:** The net benefit to data retention is only realized if the gain in C_{SN} outweighs the coupling-induced voltage drop.
- **Structural Optimization:** We highlight that these parasitic effects can be managed through structural selection—such as comparing **Channel-All-Around (CAA)** versus **Gate-All-Around (GAA)** configurations—and by implementing coupling-aware driving schemes [63, 64].

This revision provides a more balanced and technically accurate perspective on the adoption of surrounding-gate architectures in GC-eDRAM.

On page 7, line 219

"The surrounding-gate topology fundamentally strengthens gate-to-channel electrostatic control, which is instrumental in suppressing leakage current and increasing the effective storage node capacitance (C_{SN}). While an increased C_{SN} generally improves data retention and sensing margin, the same geometry concurrently enhances capacitive coupling between the storage node and control terminals, such as write word line (C_{WWL-SN}) and read word-line (C_{RWL-SN}). During word-line transitions, these parasitic paths can induce an unintended storage node voltage disturbance (ΔV_{SN})–commonly referred to as the pull-down effect–which may degrade the stored voltage level and increase dynamic loading. Recent studies indicate that this coupling constitutes a quantifiable design trade-off that can be mitigated through the strategic selection of surrounding-gate geometries (e.g., CAA vs. GAA) and the implementation of coupling-aware writing strategies[63, 64]. Such optimizations are essential to ensure that the electrostatic benefits of multi-gate structures are not neutralized by parasitic-induced signal degradation."

On page 20, line 661

63. Kim, M., Lee, J., Cho, H. & Jeon, J. Enhancing memory performance in IGZO-based 2T0C DRAM through comparative analysis of CAA and GAA FET structures. *J Comput Electron* **24**, 190 (2025).

64. Zheng, L., Wang, Z., Lin, Z. & Si, M. The impact of parasitic capacitance on the memory characteristics of 2T0C DRAM and new writing strategy. *IEEE Electron Device Lett.* **44**, 1284–1287 (2023).

11. In the line 184, the author state: "In the CAA configuration, the gate encircles the channel; in the GAA configuration, the channel is surrounded by the gate". This description does not clearly explain

the distinction between CAA and GAA.

Answer: We appreciate the reviewer's feedback regarding the ambiguity in our definitions of CAA and GAA. To provide a more rigorous technical distinction while maintaining the conciseness suggested in Comment #9, we have revised the text to clearly define their structural symmetry. In the revised manuscript, we clarify that:

- **GAA (Gate-All-Around):** A configuration where a **central channel (nanowire or nanosheet)** is completely enclosed by the gate electrode, providing the ultimate electrostatic control over a single conduction path.
- **CAA (Channel-All-Around):** A configuration where the **gate electrode resides at the center**, and the channel material (typically a thin-film oxide) surrounds the gate. This structure is particularly relevant for oxide-based vertical TFTs, where it can provide different parasitic capacitance profiles and area efficiencies.

We have updated the descriptions to ensure that the physical orientation of the gate and channel is unambiguous.

On page 7, line 213

"Surrounding-gate architectures for OS-based memory are primarily classified into GAA[57-59] and Channel-All-Around (CAA)[60-62] configurations. In the GAA structure, the gate electrode completely encapsulates a central channel wire or sheet to maximize electrostatic potential control. Conversely, in the CAA structure, the channel layer surrounds a central gate electrode, a topology frequently adopted in vertical oxide TFTs to optimize integration density and parasitic coupling characteristics. These structural distinctions result in different λ values and coupling noise profiles, necessitating strategic selection based on the specific eDRAM performance targets."

12. In the benchmark table, the author lists the length and width of the reference No.1 (2T eDRAM) is 30000 nm and 50 μm, respectively. However, in the original referenced paper, the length and width are 3 μm and 5 μm. The authors should avoid such basic errors.

Answer: We appreciate the reviewer's meticulousness in identifying the discrepancy in our benchmark table. The error was a result of an incorrect unit conversion during the data tabulation process. In the revised manuscript, we have implemented the following actions:

1. **Direct Correction:** The dimensions for Reference No. 1 have been corrected to **3,000 nm (Length)** and **5 μm (Width)**, strictly adhering to the original source.
2. **Comprehensive Verification:** To restore the technical integrity of the manuscript, we have conducted a **comprehensive audit of all numerical data** within the benchmark table (Table 1).

Each entry has been cross-referenced with its primary source to ensure that all units and dimensions are accurately represented.

3. **Unit Standardization:** We have standardized the unit notations across the table to prevent any further confusion between nm and μm.

We are grateful for the reviewer's guidance, which has enabled us to ensure the rigorous accuracy required for publication.

13. The authors claim that IGZO-eDRAM achieves an area size of $4F^2$ or even $2F^2$. I am curious whether this area includes the space required for routing. If not, when routing is taken into account, would the effective area size need to be increased further? I understand the cell area $2F^2$ is claimed by the citation, however, the author of this review paper should carefully think about it.

Answer: We appreciate the reviewer's critical and practical insight regarding the **effective cell area**. We agree that the reported $4F^2$ and $2F^2$ figures in literature primarily represent the **intrinsic cell footprint** based on the vertical transistor layout and do not fully encapsulate the area required for global routing, metal interconnects, and peripheral circuitry. In the revised manuscript, we have addressed this by:

1. **Clarifying the Definition:** We now explicitly state that $2F^2/4F^2$ refers to the **intrinsic device-level footprint** enabled by 3D vertical stacking.
2. **Addressing Routing Overhead:** We added a discussion acknowledging that in a full-scale chip implementation, the **effective area per bit** will increase due to routing constraints, via-stacking rules, and peripheral overhead.
3. **Technical Justification:** We emphasize that while routing increases the absolute area, the **M3D** nature of IGZO-eDRAM still provides a superior density advantage over planar 1T1C DRAM by offloading the memory array to the BEOL layers, thereby saving critical FEOL silicon area.

We thank the reviewer for encouraging a more rigorous and realistic representation of scaling metrics.

On page 9, line 270

"Furthermore, the proposed architecture represents a significant technological leap in data storage density, offering more than a twofold improvement over conventional planar 1T1C DRAM. By leveraging monolithic 3D vertical stacking, this approach facilitates cell-level scaling beyond the conventional $4F^2$ limit. Specifically, theoretical projections and layout analyses demonstrate that advanced implementations of this stacked GC-DRAM can achieve a $2F^2$ footprint (Fig. 4b)[42], contingent upon the optimized vertical integration of the storage node and transistor channels as depicted in Fig. 4g and 4h. However, it is important to distinguish this intrinsic cell footprint from the effective chip-level area. In practical large-scale implementations, the effective area per bit will necessarily increase

to accommodate global routing, higher-metal interconnects, and peripheral circuitry. Nevertheless, the ability to integrate these high-density arrays within the BEOL layers remains a key driver for monolithic 3D integration, as it effectively decouples the memory density from the front-end-of-line (FEOL) logic area constraints."

14-1. In modern CPUs, caches and CPU cores are usually fabricated on the same die, which implies that process compatibility must be considered. The authors should discuss whether oxide GC-eDRAM can be co-integrated with silicon-based processing units. If not, they should address how this issue could be resolved.

Answer: We appreciate the reviewer's insightful comment regarding the co-integration of oxide-based memory with silicon processing units. We agree that process compatibility is a decisive factor for the adoption of eDRAM in high-performance computing (HPC) environments. In the revised manuscript, we have added a dedicated discussion on the **BEOL-compatible integration** of oxide semiconductors. We clarify that the primary advantage of oxide GC-eDRAM is its ability to be deposited via **Atomic Layer Deposition (ALD) or sputtering at temperatures below 400°C**, which adheres to the strict thermal budget of existing silicon CMOS back-end processes. This allows for **Monolithic 3D (M3D)** stacking of memory arrays directly atop logic dies without compromising the integrity of underlying transistors.
Furthermore, we address the technical hurdles, such as **hydrogen-induced degradation** during BEOL processing and **interfacial stress**, and propose engineering strategies (e.g., robust passivation and optimized metal-gate stacks) to ensure reliable co-existence. We believe this addition provides a realistic roadmap for oxide-on-silicon integration.

On page 9, line 282
"From an architectural perspective, the successful implementation of oxide-based GC-eDRAM hinges on its seamless co-integration with silicon CMOS processing units. OSs present a distinctive advantage for this "Memory-on-Logic" paradigm: they can be synthesized using RF sputtering or ALD at temperatures ranging from room temperature to sub-400 °C. This thermal window is critically compatible with the BEOL thermal budget (typically ≤ 400 °C), ensuring that the integration of upper memory layers does not cause dopant redistribution or performance degradation in the underlying silicon logic. However, achieving high-performance co-integration requires addressing several material-level challenges. Under BEOL environments, OS channels are highly susceptible to unintentional hydrogen incorporation from surrounding dielectric layers, which can lead to V_{TH} instability. In addition, mechanical stress arising from thermal expansion mismatches between OS layers and silicon interconnects pose a further challenge, potentially compromising device reliability.

To address these integration-level constraints, practical strategies such as robust encapsulation and hydrogen-barrier engineering, together with post-CMOS hybrid stacking approaches–including wafer-to-wafer bonding and high-density through-silicon-via integration–have emerged as effective routes to suppress adverse defect formation and minimize process interference. Establishing these engineering frameworks is therefore critical for translating discrete device-level performance into reliable full-chip implementation, as discussed in the following sections."

14-2. The authors should provide more background on GC-eDRAM itself, including details on different configurations such as 2T-DRAM and 3T-DRAM.

Answer: We appreciate the reviewer's suggestion to provide a broader background on GC-eDRAM configurations. We agree that a comparative discussion between **2T0C** and **3T0C** architectures is essential for a complete understanding of why the field is converging toward 2T-based oxide memory. In the revised manuscript, we have added a section explaining that while the **3T0C configuration** (which utilizes an additional transistor to isolate the storage node) is often preferred in silicon CMOS to mitigate read disturbances caused by high off-state leakage, this complexity is largely unnecessary in oxide-based devices. Thanks to the **ultra-low leakage current** of oxide semiconductors—typically six orders of magnitude lower than silicon—read disturbance is negligible, allowing for a more compact **2T0C structure**.

Furthermore, we highlight that the 2T0C configuration offers a superior **area reduction (up to 89% vs. SRAM)** and simplified routing, which is critical for maintaining high yield in **M3D BEOL integration**. For these reasons, our review focuses on the 2T0C architecture as the primary vehicle for oxide-based high-density memory.

On page 4, line 89

"The architectural evolution of GC-eDRAM has primarily been driven by the trade-off between sensing reliability and cell density. Conventional silicon-based implementations often employ a 3T0C configuration, where an additional transistor is dedicated to isolating the storage node during read operations to prevent read disturbance a critical issue given silicon's relatively high off-state leakage. However, the advent of OSs has shifted the paradigm toward the 2T0C architecture. Due to the intrinsic ultra-low off-current of OS channels, the storage node can maintain its charge integrity even without the additional isolation transistor, making read-disturb effects virtually negligible. By adopting the 2T0C structure, the cell footprint is reduced by approximately $2F^2$ compared to 3T0C, and the routing complexity is significantly diminished. This simplified 2T0C topology is not only more area-efficient, offering up to 89% area savings over SRAM, but also highly compatible with the stringent yield and thermal budget requirements of monolithic 3D (M3D) BEOL integration[24, 29]."

Authors: Sang Won Chung, Seong Hun Yoon, Jae Kyeong Jeong
Title: Oxide semiconductor gain cell-embedded memory: materials and integration strategies for next generation on-chip memory
Manuscript ID: COMMSENG-25-0741
RECOMMENDATIONS AND COMMENTS FOR AUTHOR(S)

Answer to the comments by Reviewer #1 and Summary of changes

We are grateful to the Reviewer for his/her constructive comments and corrections on our manuscript. After considering the comments, we revised our manuscript and made the following reply.

1. Line 77: “is sensed through capacitive coupling via the read-transistor gate dielectric”. Capacitive coupling is not a sensing mechanism. Capacitive coupling degrades sensing.

Answer: We sincerely appreciate the reviewer’s insightful comment regarding the terminology used to describe the sensing mechanism. We agree that the term "capacitive coupling" could be misleading, as it often refers to parasitic interference rather than the fundamental sensing principle. As the reviewer correctly pointed out, the sensing in GC-eDRAM is achieved by modulating the conductance of the read transistor based on the voltage level of the storage node (SN). Since the SN is directly connected to the gate of the read transistor, the potential at the SN determines the gate-to-source voltage, thereby controlling the drain current during the read operation. Accordingly, we have revised the manuscript to clarify that the sensing mechanism is based on the modulation of the read transistor's conductance in a decoupled read path.

On page 3, line 85
“Furthermore, GC-eDRAM also adopts a decoupled read path, in which the storage node is isolated from the read bit-line. The stored state is sensed through voltage or current sensing mode, where the read transistor's conductance is modulated by the voltage at the storage node.”

2. Line 88-93: The comparison on 3T0C vs 2T0C is wrong. 3T0C is used to suppress sneak path from unselected cells, not disturbance to storage node.

Answer: We sincerely thank the reviewer for this insightful comment. We agree that the primary role of the additional transistor in a 3T0C configuration is to suppress sneak path currents from unselected cells within an array, rather than solely preventing read disturbance at the storage node. Accordingly, we have revised the manuscript to clarify the dual roles of the third transistor in conventional silicon-

based implementations: (1) suppressing sneak path currents on the shared read bitline and (2) providing electrical isolation. We also emphasized that the ultra-low off-state current of OS-based channels allows the 2T architecture to effectively mitigate both sneak paths and charge leakage without the need for an additional select transistor.

On page 3, line 98

“Conventional silicon-based implementations often employ a 3T configuration, where an additional read-select transistor serves dual purposes: (i) suppressing sneak path currents from unselected cells connecting to the same read bit-line, and (ii) providing electrical isolation during read operations-both critical requirements given silicon's relatively high off-state leakage. However, the advent of OSs has shifted the paradigm toward the 2T architecture. Due to the intrinsic ultra-low off-current of OS channels, both sneak path currents from unselected cells and charge leakage from the storage node (SN) can be mitigated.”