Towards Scalable 3D Integration of 2T-nC FeRAM with Hundreds of Layer Stacking

Jiahui Duan[1*], Shan Deng[1*], Rudra Biswas[2*], Sadik Yasir Tauki[2], Sizhe Ma[1], Rajiv Joshi[3], Thomas Kampfe[4], Xiao Gong[5], Vijaykrishnan Narayanan[2], and Kai Ni[1]
[1]University of Notre Dame; [2]The Pennsylvania State University; [3]IBM Thomas J. Watson Research Center; [4]Fraunhofer IPMS; [5]National University of Singapore;
*Equal contribution; (email: jduan3@nd.edu)

***Abstract*— In this work, we study the limits of the number of capacitors and read history dependence in a 2T-nC ferroelectric random-access memory (FeRAM) cell, paving the way for its high-density integration toward hundreds of stacked layers. Through a comprehensive experimental and simulation study on the scaling behavior of the 2T-nC FeRAM architecture, we demonstrate: (i) successful fabrication of 2T-64C cells with robust memory operation and clearly distinguishable '0' and '1' states, even in 64-capacitor configurations; (ii) that the parasitic capacitance of the floating node originates predominantly from the linear component of the ferroelectric capacitor, and its impact on n-scaling—due to degraded sense margin—can be mitigated by floating unselected capacitors with enough TΩ isolation; (iii) that sharing write and read transistors among n capacitors introduces a read history dependence issue due to fluctuating floating node voltage (V_{FN}); and (iv) that a proposed FN discharge scheme can effectively eliminate read-sequence dependence, at the cost of reduced read endurance.**

I. Introduction

The growing demand for high-performance computing has intensified the memory wall—the bottleneck caused by the speed and energy gap between fast volatile memory (SRAM, DRAM) and slower nonvolatile storage (NAND). Bridging this gap requires a memory solution that offers speed, nonvolatility, scalability, and endurance. Ferroelectric capacitor-based memories, particularly those using doped HfO_2, show strong potential due to their fast switching, low-voltage operation, and inherent nonvolatility. As shown in Fig. 1(a), they can effectively span the performance-density gap between DRAM and NAND. To exploit ferroelectric behavior, several architectures have been proposed, including sequential 1T-1C, parallel 1T-1C, and 1T-nC cells [1]. Among them, the parallel 2T-nC structure stands out for offering the highest density, as it shares multiple ferroelectric capacitors across write and read transistors (T_W&T_R), while retaining independent control of write and read paths (Fig. 1(b)).

The 2T-nC memory cell operates by decoupling the write and read paths to for excellent write reliability and read performance [2]. A T_W helps program the ferroelectric capacitor while a separate T_R is responsible for charge amplification during readout. After each write, the internal node (i.e., FN) is reset to 0 V (V_{FN}=0) (Fig. 1(c)) to maintain excellent reliability. During read, the difference in ferroelectric switching behavior between logic states '0' and '1' leads to different amounts of polarization switching (ΔQ_0 and ΔQ_1), which is then translated into a distinguishable current by the read transistor. This charge-to-current amplification mechanism enhances signal margin and enables excellent scalability, making the 2T-nC cell a strong candidate for dense, low-power, and high-speed nonvolatile memory applications.

II. The Limit of n in the 2T-nC Cell

To increase the density of the 2T-nC FeRAM, one cell should include as much multiple Metal-Ferroelectric-Metal (MFM) capacitors as possible, i.e., highest *n*. (Fig.2(a)). However, it is unclear what is the limit of *n* and how it limits the sense margin and how it is related with the parasitics. TCAD models of 9-string 2T-*n*C arrays with *n* up to 256 are built to extract the parasitic capacitance on the FN, including the C_X, C_Y, the diagonal C_{XY}, and also that from linear dielectric of FE, C_{FE}, as shown in Fig.2(b). The parameters used for the vertical 2TnC FeRAM simulation is also shown in Fig.2(b). Decomposition of the total FN capacitance, C_{FN}, to different components, shown in Fig.2(c), indicate that the majority parasitic capacitances are from C_{FE}. In fact, which. total C_{FN} grows linearly with the number of MFM capacitors. Other components are negligible due to screening.

To evaluate the scalability of the 2T-nC FeRAM architecture, we analyze the sensing margin as a function of *n* under two biasing conditions for unselected cells: grounded and floating. As shown in Fig. 3(a)–(c), when unselected MFM capacitors are grounded, the sensing margin (ΔV_{FN} or ΔI_{RBL} between logic states '0' and '1, and sensed I_{RBL}) degrades noticeably with increasing *n*. This degradation is primarily attributed to charge sharing between the selected and unselected capacitors, which effectively reduces the voltage differential across the sense node. Such behavior imposes a scalability limit on array size unless mitigation techniques—such as hierarchical bitline segmentation or biasing schemes—are implemented. In contrast, Fig. 3(d)–(f) presents the ideal case where unselected capacitors are electrically floating. Here, the sense margin remains stable across all values of *n*, indicating that in the absence of charge leakage or parasitic coupling, the 2T-nC cell is inherently scalable as other caps do not participate in the charge sharing. This comparison highlights the critical role of cell isolation and array biasing strategy in achieving reliable sensing performance in large-scale 2T-nC FeRAM arrays.

In real circuit, isolation is never perfect, so we also evaluate the impact of the isolation resistance on the sense margin. Fig. 4 illustrates the impact of this non-ideal isolation: as the number of connected capacitors increases, a low isolation resistance leads to a noticeable reduction in the final V_{FN} (Fig. 4(a)) and the corresponding sensed I_{RBL} (Fig. 4(b)).

Consequently, the differential sensing margins—ΔV_{FN} or ΔI_{RBL} between logic states '0' and '1' diminish with n, as shown in Fig. 4(c) and (d). This degradation stems from increased leakage and charge redistribution through the resistive path, which blurs the distinction between stored logic states. To ensure robust sensing across large arrays, it is therefore critical to maintain a sufficiently high isolation resistance—on the order of teraohms ($T\Omega$) or greater.

III. Farbication of 2T-nC FeRAM

For experimental demonstration, we fabricated the proposed 2T-nC structure on a p-type silicon wafer (Fig. 5a)) [3]. After ion implantation and activation of the source and drain regions, the isolation oxide in the gate area was etched, and a 10 nm HfO_2 layer was deposited via atomic layer deposition (ALD) as the gate dielectric. Tungsten electrodes were then sputtered onto the source, drain, and gate regions after via formation using both dry and wet etching processes. Subsequently, a 10 nm $Hf_{0.5}Zr_{0.5}O_2$ layer was deposited by ALD, followed by sputtering of the top tungsten electrode and dry etching to open the contact vias. Finally, the device underwent rapid thermal annealing in a N_2 and H_2 ambient. A top-view SEM image of 2T-3C is shown in Fig.5(b). Fig.5(c) presents the I_{RBL}–V_{WPL} curve, demonstrating the correct operation of the T_W and T_R. In Fig.5(d), the DC I_D –V_G characteristics of the MFM capacitor combined with the T_R exhibit the expected hysteresis window. Fig.5(e) shows the PUND response of the MFM capacitor; both results validating its ferroelectric nature. Finally, Fig.5(f) displays the experimental waveforms for write and read operations in a 2T-3C cell, along with the corresponding I_{RBL} for data states '1' and '0', showing a clear and distinguishable margin.

To evaluate the scalability of the 2T-nC cell in terms of capacitor stacking, 2T-16C and 2T-64C structures were also fabricated. In Fig. 6(a), the pulse I–V characteristics of the 2T-16C cell with a capacitor and T_R combination exhibit uniform FeFET-like switching behavior, indicating consistent device operation. For comparison, Fig. 6(b) shows the corresponding measurement from a configuration with separated write and read paths. The '0' and '1' states I_{RBL} distribution of the 2T-16C cell, shown in Fig.6(c), further confirms clearly distinguishable states. Fig.6(d) and Fig.6(e) present top view SEM overview and zoom-in images of the fabricated 2T-64C cells, demonstrating the feasibility of high-density integration. The Q_{FE}-V_{WPL} characteristics of the 2T-64C, shown in Fig. 6(f), validate the expected ferroelectric switching behavior. The DC I_D-V_G curves from 64 MFM capacitors and 1 T_R combinations in Fig. 6(g) (displayed 32 devices) show consistent hysteresis windows, highlighting reliable memory operation across multiple cells. In Fig.6(h), clearly distinguishable '0' and '1' states are observed in all 64 devices, indicating excellent readout margins. Finally, Fig.6(i) shows the I_{RBL} distribution of the 2T-64C cell, further demonstrating the robustness and scalability of the proposed architecture.

IV. Dependence of Sense Margin on Read History

In multi-capacitor 2T-nC architectures, the sensing results of each ferroelectric capacitor can be influenced by the read history, primarily due to the shared floating node. As illustrated in Fig. 7(a), the voltage on V_{FN} after sensing one capacitor can impact the readout of subsequent capacitors. Simulation results in Fig. 7(b)–(c) show that for a two-capacitor cell storing '00' and '10', the sensed V_{FN} and I_{RBL} of the second capacitor (C2) after sensing C1 vary depending on what C1 stores. The C2 read current is raised if C1 stores '0'. This indicates a read-sequence dependency that can affect sensing accuracy. Experimental data in Fig.7(d) further confirms this behavior, while Fig.7(e) isolates the effect to be solely due to the read order, independent of swapping of MFM.

To eliminate the read-sequence dependence observed in 2T-nC cells, we propose an effective mitigation technique: actively discharging the shared V_{FN} node before each read operation. As shown in Fig.8(a), this approach resets the V_{FN} potential to a known state, preventing charge accumulation from prior reads. Fig. 8(b)–(c) compares the standard delay-based read scheme with the proposed discharge method. Fig.8(d)–(g) demonstrate that, while V_{FN} and the read current of C3 remain data-dependent under the delay-only approach, the discharge method fully removes such dependence. Measurements in Fig. 8(h)–(k) confirm that even after a 1 ms delay, C3 read current under the delay-only scheme still reflects C1&C2 data patterns, whereas the discharge scheme ensures correct and consistent sensing with as little as 1 μs discharge time. Similar results are observed in two-capacitor measurements (Fig. 8(l)–(o)), where C2 read current is strongly data-dependent with delay but remains unaffected when preceded by a V_{FN} discharge. Fig. 8(p)–(q) summarizes the C2 current across varying delay/discharge times, highlighting the robustness of the discharge technique in restoring sequence-independent readout. This method is both circuit-compatible and scalable, offering a practical solution to enhance read reliability in dense 2T-nC memory arrays.

While V_{FN} before each read effectively eliminates read-sequence dependence, it introduces a trade-off in terms of read endurance. As shown in Fig.9(a)–(b), the read current degrades more rapidly over cycles in the discharge scheme compared to the delay-only case. This behavior stems from the fact that discharging V_{FN} removes the ferroelectric switching induced charge on the FN, causing permanent loss of the charge, especially under repeated cycling. Consequently, although the discharge method enhances read accuracy and consistency, it reduces the number of reliable read cycles. This highlights a fundamental trade-off between sensing fidelity and device endurance, which must be balanced through circuit-level optimization or adaptive read schemes depending on the application's retention and reliability requirements.

V. Conclusion

A comprehensive experimental and modeling study is conducted on 2T-nC FeRAM architecture to push towards hundreds of layer stacking. We validated the architecture's feasibility and sensing behavior on 64 capacitors. And we addressed the issue of read-sequence dependence in multi-capacitor by introducing a simple V_{FN} discharge that ensures reliable sensing.

Acknowledgement: This work was partially supported by SUPREME and PRSIM, two of the SRC/DARPA JUMP2.0 centers, NSF 2346953, and Singapore MOE Tier 1 A-8001168-00-00 and A-8002027-00-00. TCAD and SPICE modeling was supported by U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Energy Frontier Research Centers program under Award Number DESC0021118. **References:** [1] J. Duan et al., arXiv 2025; [2] Y. Xiao, et al., IEEE EDL 2023; [3] S. Deng, et al., IEDM 2023.

2T-nC FeRAM Hybrid Cell for High Density, Performance, and Reliability Memory

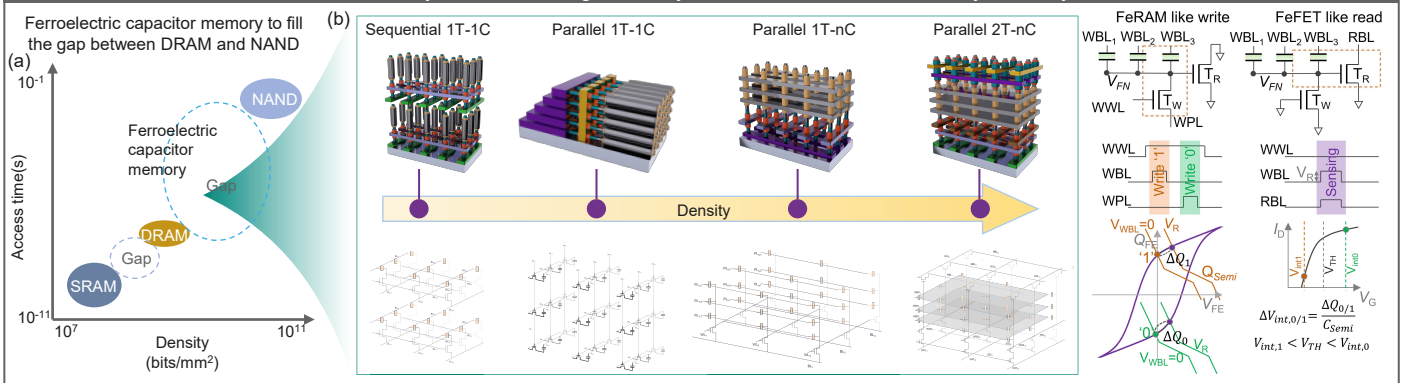

Fig.1. (a) Ferroelectric capacitor memory can fill the gap between DRAM and NAND, by offering fast read/write and nonvolatility. (b) Ferroelectric capacitors can be integrated into memory cells in various configurations, among which the vertically stacked 2T-nC structure offers the highest density. (c) The separate write & read paths of 2T-nC cell ensure V_{FN} =0 after write, leading to 1T-1C FeRAM like reliability. During the read, different memory states causes difference in ΔQ_0 and ΔQ_1, which is amplified by T_R action, thus ensuring excellent scalability of the cell for high density.

Question 1: What is the Limit of the Number of Capacitors (i.e., *n*) in a 2T-nC FeRAM Cell?

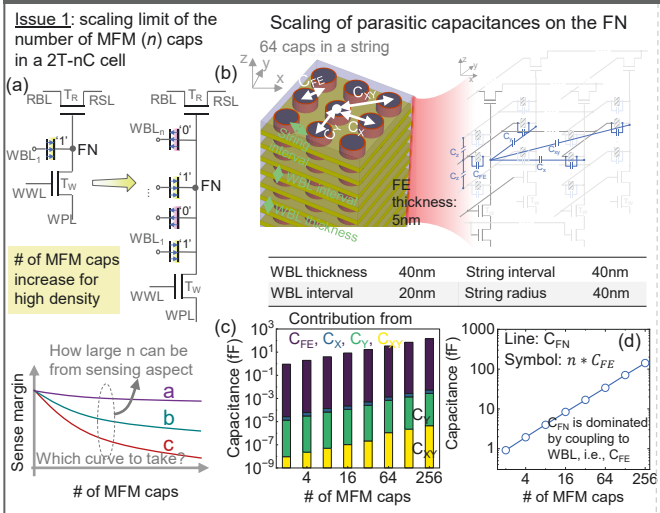

Fig.2. (a) A critical issue is the limit of *n* in a 2T-nC FeRAM cell. (b) To answer that, TCAD models of varying n are built to extract (c) parasitic caps. on the FN, which can degrade sense margin and limit scalability. (d) The C_{FN} is dominated by coupling to WBL (i.e., C_{FE}).

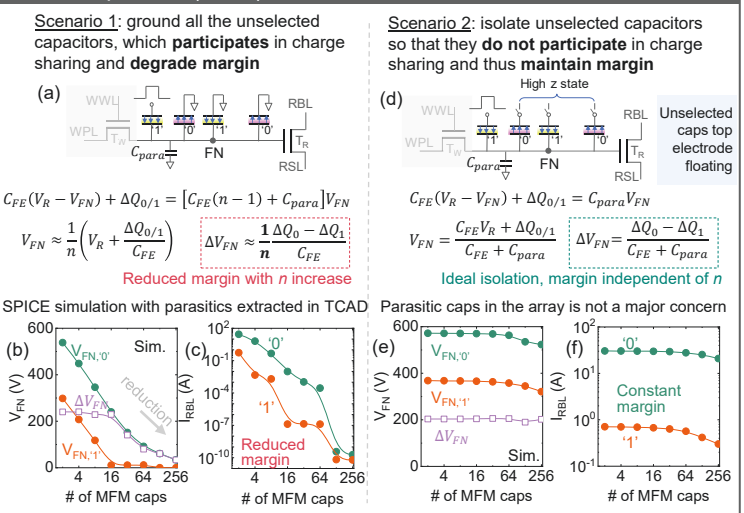

Fig.3. Comparison of sensing margin in a 2T-nC cell with unselected MFM capacitors either grounded or floating. (a)-(c) the sense margin degrades with increasing *n* due to charge sharing from unselected caps in grounded case, (d)-(f) while the margin remains constant regardless of *n* in ideal floating case.

How the Isolation Impacts the Sensing Margin?

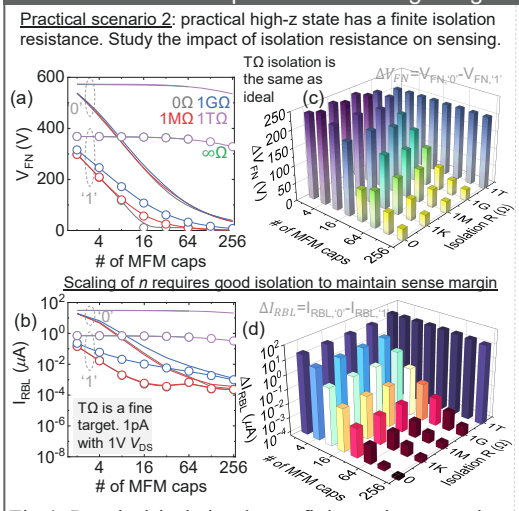

Fig.4. Practical isolation has a finite resistance, when small can cause reduction in (a) V_{FN}, (b) sensed I_{RBL}, (c) ΔV_{FN} margin, and (d) ΔI_{RBL} margin between '0' and '1' with increasing *n*. A good enough isolation (e.g., TΩ) is needed to maintain sense margin with *n*.

Experimental Demonstration of 2T-nC FeRAM Operation (n from 3 to 64)

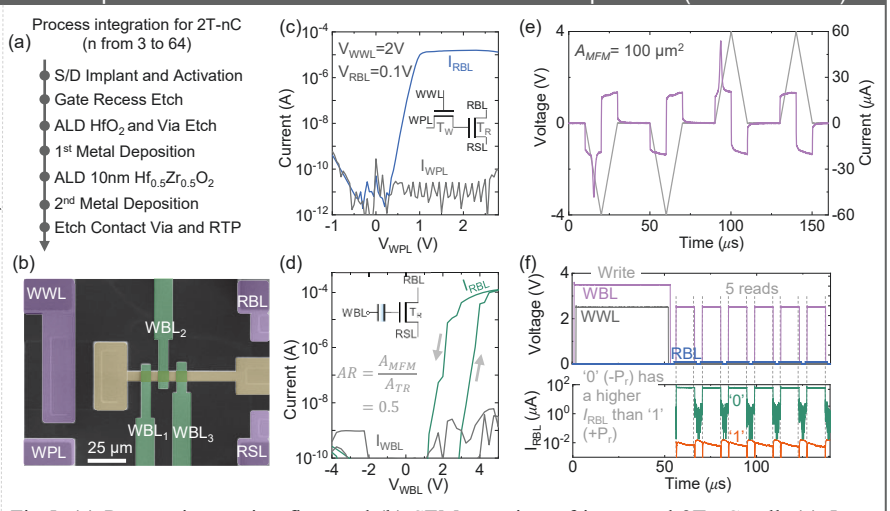

Fig.5. (a) Process integration flow and (b) SEM top view of integrated 2T-nC cell. (c) I_{RBL}-V_{WPL} curve shows correct function of the T_W+T_R structure. (d) DC I_D-V_G of the MFM cap + T_R structure, showing expected hysteresis window. (e) PUND responds of the MFM cap confirms the ferroelectricity of the MFM cap. (f) Experimental waveforms of write and read operation of a 2T-3C cell and the RBL current for data '1' and '0', which shows a large margin.

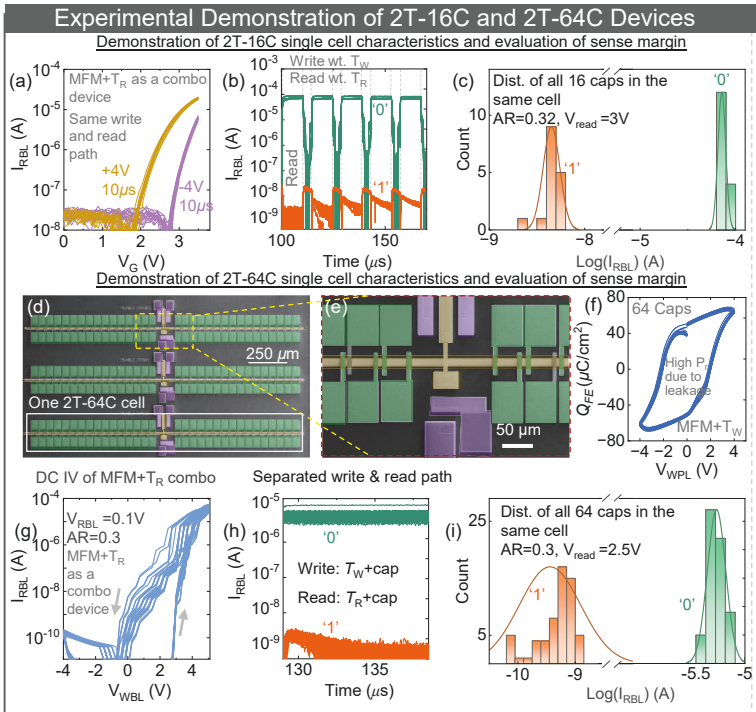

Fig.6. (a) Pulse I–V of the 2T-16C with cap+T_R combo shows uniform FeFET switching. (b) the separated paths one. (d)&(e) SEM images and (f) Q_{FE}-V_{WPL} of the 2T-64C cells. (g) I_D–V_G curves of 64 cap+T_R combo. (h) Distinguishable '0' and '1' states in 64 devices. I_{RBL} dist. of the (c) 2T-16C and (i) 2T-64C cell.

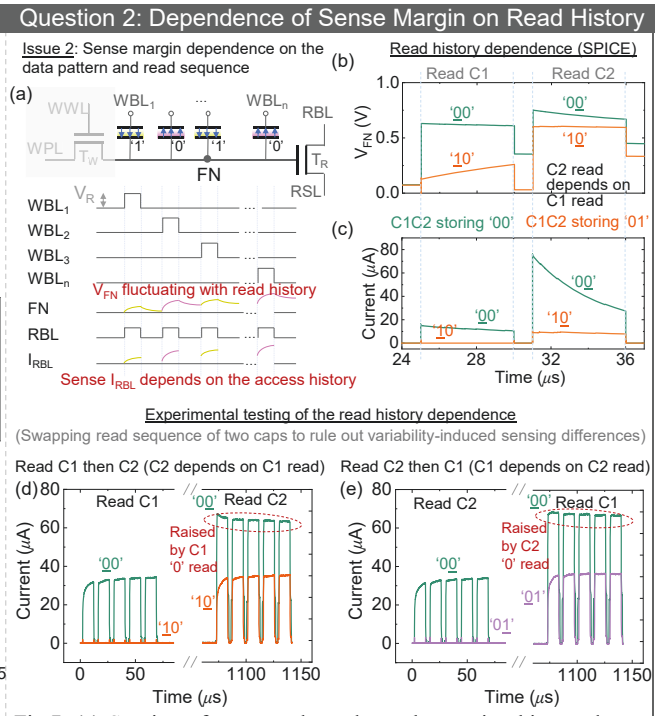

Fig.7. (a) Sensing of one cap depends on the sensing history due to V_{FN}. Simulated (b) V_{FN} and (c) I_{RBL} of 2 caps storing '00' and '10' show that sensing C2 depends on C1. (d) Experiments verifies such behavior and (e) confirms its only dependence on read sequence.

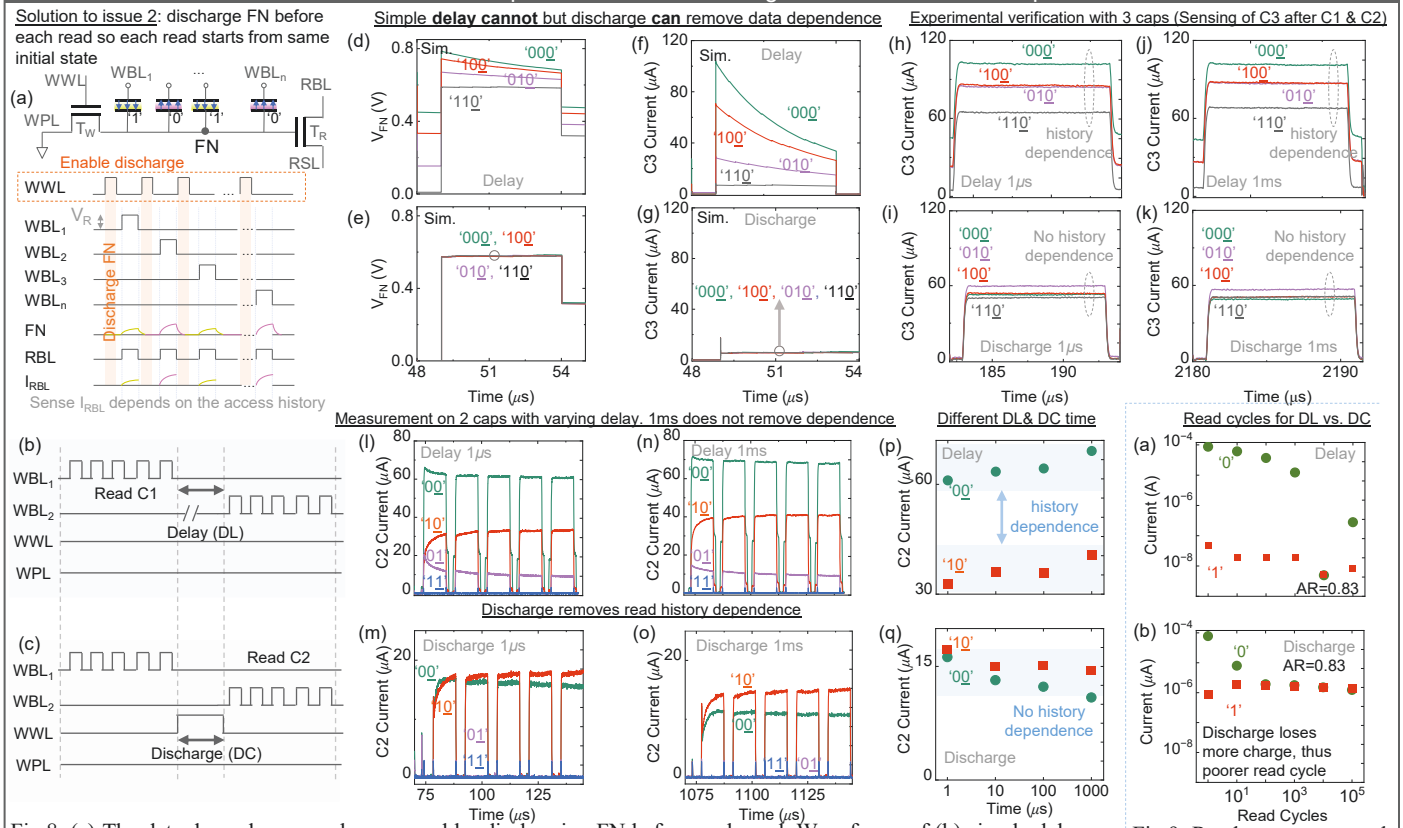

Fig.8. (a) The data dependence can be removed by discharging FN before each read. Waveforms of (b) simple delay vs. (c) discharge are used to compare their effect. Simulated (d)/(e) V_{FN} and (f)/(g) C3 read current after C1 & C2 read and measured C3 read current after (h)/(i) 1μs and (j)/(k) 1ms for simple delay/discharge, respectively. It clearly shows with simple delay, up to 1ms, C3 read current still strongly depends on the data pattern in C1 and C2, while 1μs discharge can remove this dependence. Measurement on 2 caps show that C2 read current after C1 read at (l)/(m) 1μs and (n)/(o) 1ms for delay/discharge, respectively are aligned with the 3 cap data. (p)/(q) Summary of C2 current vs. delay/discharge time.

Fig.9. Read current vs. cycle for (a) delay and (b) discharge cases. As discharge removes the V_{FN}, causing more charge loss and hence less read cycles.