A 512 kb 0.069 μm^2 Logic 3T GCRAM with 27 μs Retention Time at 85°C in 16 nm FinFET

A. Yigit[1], A. Avcioglu[1], R. Giterman[2], L. Johansson[1], I. Lavi[2], E. Leizerovitz[2], X. Li[1], C. Mueller[1], E. Rotem[2], Y. Shoshan[2], A. Teman[2], S. Yuzhaninov[2], A. Burg[1]
[1]*RAAAM Memory Technologies, Ecublens, Switzerland,* [2]*RAAAM Memory Technologies, Petah Tikva, Israel*

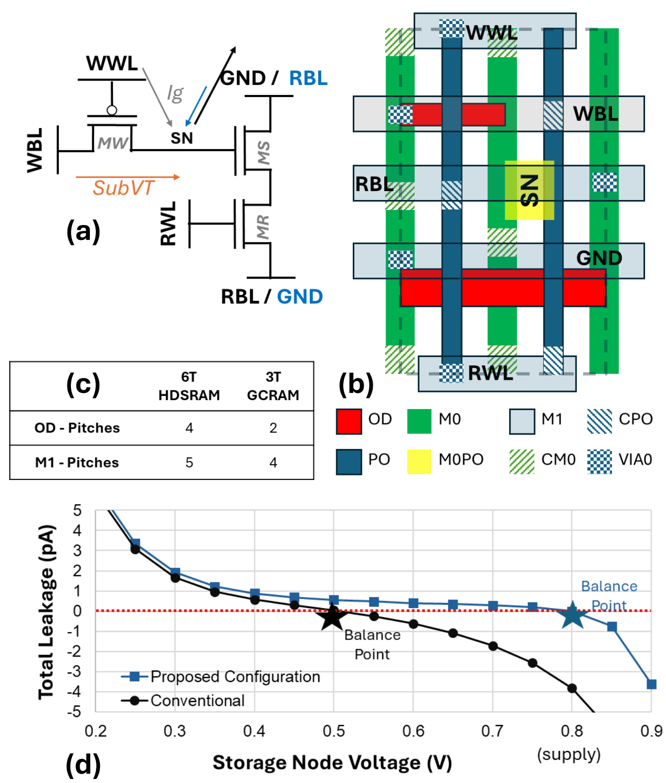
***Abstract*—We present the highest-density reported single-supply Gain-Cell (GCRAM) memory macro (0.069 μm^2 bitcell) incorporating low-overhead write and retention WL-assist and a PVT-compensating dynamic reference generation. The proposed bitcell introduces a novel flipped topology that improves steady-state storage node voltage and retention performance, enabling read currents that are largely independent of data retention time. The 512kb (1024 bit wide) logic-rule macro area in 16 nm FinFET is 0.044 mm^2. Measurements demonstrate a retention time of 147 μs at 25ºC, which is the longest GCRAM retention time reported in 16nm. The write and read currents of 16 nA/MHz/bit and 26 nA/MHz/bit achieve significant power reduction over SRAM.**

***Index Terms*—embedded memories, gain-cell, GCRAM, SRAM.**

I. Introduction

As the demand for memory capacity continues to grow and 6T SRAM density scaling stagnates, memory increasingly dominates die size and power consumption in modern Systems-on-Chip (SoCs). The 1T-1C eDRAM is denser and more power efficient; however, it requires additional process steps, creating challenges for SoC integration in advanced process nodes. As a middle ground between these two alternatives the Gain-Cell (GCRAM) Memory offers full CMOS compatibility with less area and better power when compared to the SRAM. The single drawback of the GCRAM is the data retention time (DRT), as the data is stored on the small gate capacitance of a single transistor and additional layout parasitics. As retention time decreases with technology scaling, boosted voltages are required to suppress leakage. However, the need for different supplies for a macro makes SoC integration challenging, and these supplies must track the main supply voltage changes for low power operations.

Previous works on GCRAM [1]–[3] have demonstrated its potential as low-power, area-efficient memories. However, these implementations have relied on external supplies and have limited retention times. Furthermore, previous GCRAM memories, implemented with logic design rules, have always exceeded the SRAM bitcell area despite the lower transistor count. This work advances the state of the art by demonstrating the first logic-rule 3T GCRAM with a bitcell area smaller than that of high-density 6T SRAM (based on foundry bitcell) in 16 nm FinFET. Moreover, it features integrated voltage generation and a novel topology that enhances both retention and read stability.



(c)	6T HDSRAM	3T GCRAM
OD - Pitches	4	2
M1 - Pitches	5	4

Fig. 1. (a) 3T GCRAM bitcell schematic with leakage currents and (b) bitcell layout up to metal layer-1. (c) Layout comparison with 6T-HDSRAM and (d) the storage node leakage currents according to the storage node voltage.

II. GCRAM Memory Array

This section describes the architecture and circuit techniques employed in the proposed GCRAM macro to achieve high density, low power, and robust retention. It is organized into three main parts: the 3T GCRAM bitcell and memory organization, the bit-line periphery and dynamic reference generation, and the word-line periphery along with assist voltage generation.

A. 3T GCRAM Bitcell Design

The proposed 3T GCRAM bitcell schematic and layout are shown in Fig. 1(a) and (b), respectively. They occupy 6% less area compared to a “pushed-rule” foundry High-Density 6T SRAM bitcell in the same technology. Applying “pushed-rule” design rules to the logic bitcell enables up to 30%

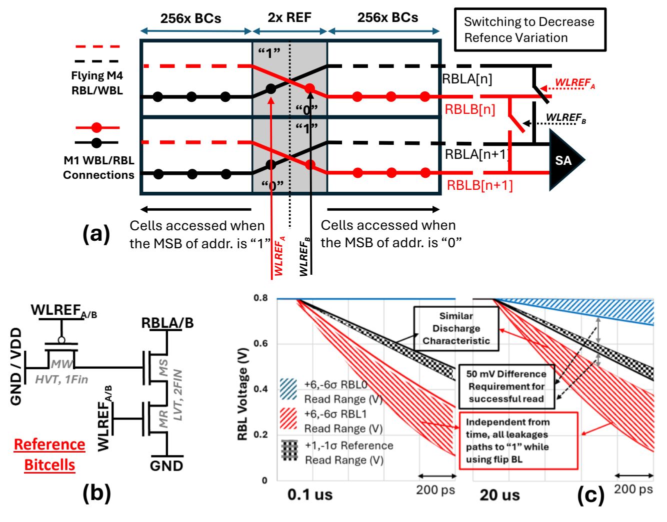


Fig. 2. Reference cell implementation in the macro architecture with schematic of the reference cell. Worst-case condition simulation (125°C and FF corner) results for the read operation at different retention times (0.1 μs and 20 μs).

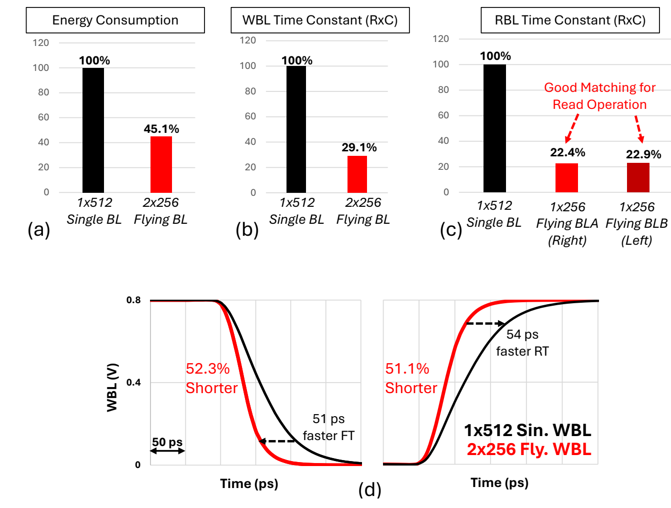


Fig. 3. Parasitic effect comparison of the single 512x BL and 256x2 flying BL (a) under energy consumption on *WBL*, (b) loading of the *WBL*, (c) loading of the *RBL*, and (d) *WBL* transient characteristic.

additional area reduction, as the 3T GCRAM requires fewer active layers (OD) and metal pitches (Fig. 1(c)).

Compared to previous 3T GCRAM bitcells, read bit-line (*RBL*) and common node (*GND*) connections to the storage transistor (MS) and the read transistor (MR) are reversed, as shown in Fig. 1(a). In the proposed bitcell, the steady-state voltage of storage node (*SN*) approaches the supply level (*VDD*), whereas the steady-state voltage of the conventional bitcell stabilizes at *VDD/2*, compromising both '0' and '1' retention times (Fig. 1(d)). The proposed configuration therefore helps to maintain a strong *SN* "1" level, also resulting in a read-1 current that is mostly independent of retention time. This independence is key to achieving high performance, high read margins, and thus high retention times.

B. Bitcell Array and Bit-line Periphery Design

The memory is divided into two sections based on the MSB of the address, with their read bit lines (*RBLs*) and write bit lines (*WBLs*) separated into sections A and B as shown in Fig. 2(a). This division reduces the bit line load and can be further exploited for reference bit-line generation as explained in the following. The bit-line capacitances are matched with flying bitlines which are drawn on the metal layer-4.

Two reference bitcells (Fig. 2(b)), one of them connected to the *GND* and the other one connected to the *VDD* are placed exactly in the middle of the row, with identical layouts to the data-storing bitcells, to achieve high regularity in the bitcell region. During a read, a write is first performed to the *SNs* of the associated reference bitcells, followed by a read to mimic both write and read coupling, controlled by the reference word-line (*WLREF*) signal.

The reference currents generated from the 512 reference bitcells per half-array (512x512) (holding 256x '1' and 256x '0') are averaged at the input to the sense amplifier using switches to achieve a low-variance reference current. Variations in read currents with generated reference current under different retention times (1 μs and 20 μs) are illustrated in Fig. 2(c) based on simulations. As mentioned, the read-1 current of the data does not change much with increasing retention time and maintains the read speed. Therefore, the generated reference current is placed closer to the read-1 curves by selecting smaller threshold (faster) transistors in the read port and the DRT is determined by the "0" retention.

In the bitline periphery, a differential sense amplifier determines the digital output by comparing the data *RBL* and generated reference bit-line. An offset-canceling differential sense amplifier is employed achieving an input offset of 50 mV, and its common mode range is selected according to the reference variation.

Fig. 3 shows the further advantages of the flying bitline technique from both time and energy perspectives for read and write operations, significantly improving read time with a 77.1% reduction in the time constant. An additional metal layer-4 strap enhances the match between the time constants of the *RBLA* and *RBLB* lines.

Since most of the write power is consumed on the *WBLs*, dividing *WBLs* into two sections can reduce energy consumption by up to 55%, translating to a decrease in write power of up to 45%. Fig. 3(d) shows the *WBL* signal transient characteristics, where the flying *WBL* significantly improves the slew rate, shortening the write time by 105 ps.

C. Word-line Periphery Design and Assist Voltage Generation

Fig. 4 presents the write configuration of the macro. The macro utilizes two boosted supplies: *VB* and *VSS*. The positive boosted supply *VB* is generated by a charge pump (CP) located at the top of the macro, occupying only 0.5% of the macro area.

The charge pump circuit generates a voltage nearly 1.25 times the supply voltage (a fixed gain determined by the

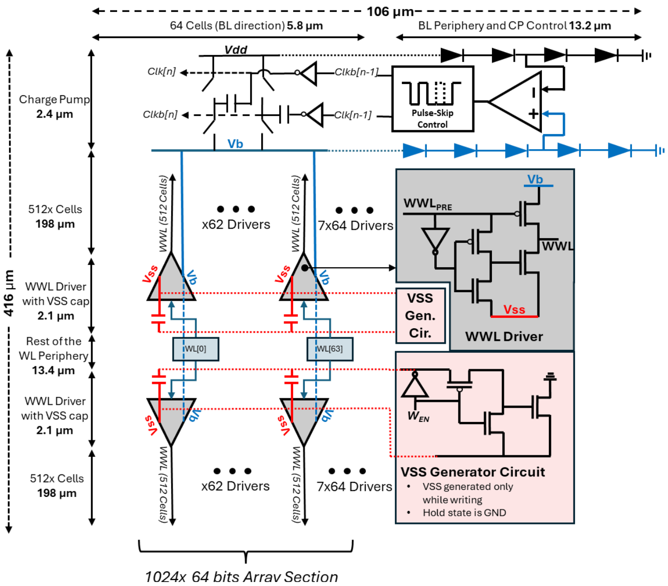


Fig. 4. The column representation of the macro with the schematic of the WWL driver, VSS generator circuit.

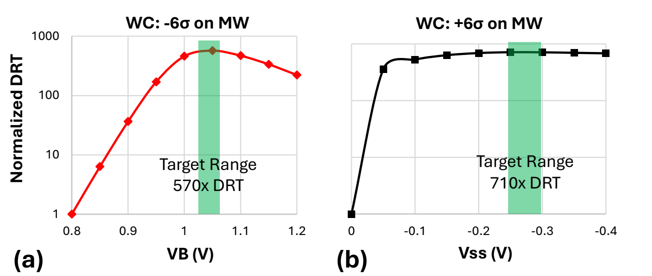
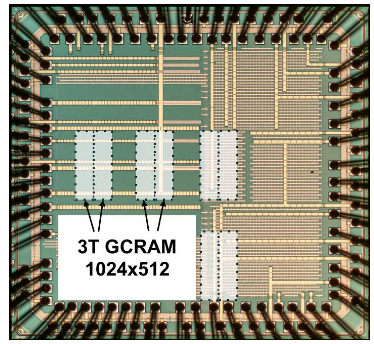


Fig. 5. Simulations of the assist voltage ((a)VB and (b)Vss) levels effect on the DRT with sigma shifted write transistor.

voltage divider diodes) and stabilizes the output voltage using a pulse-skip method [4]. The boosting capacitances are a combination of MOM-MOS caps for reduced variation, with each CP cell fitting within a 64-bitcell width. A total of 16 cells used in the CP circuit, utilizing 1024x *WWLs* as output capacitance, generate the *VB* level with a ripple of less than 10 mV. Due to its diode-based feedback, the CP output tracks supply changes, adapts to lower operating voltages, and suppresses subthreshold leakage with a lower power consumption.

The *VSS* voltage is also generated in the macro and is used for writing strong "0" levels. This voltage is produced by the *VSS* generator circuit [5], which employs a bootstrap *WWL* driver to reduce the leakage currents from *VSS* to *GND*. The *VSS* voltage is generated only during write to minimize power consumption and utilizes only MOS caps due to routing limitations. A *VSS* generator circuit is used for each half-array, with their capacitances (*VSS* caps) distributed along with *WWL* drivers, occupying only 1% of the total macro area.

The boosted voltage levels are determined based on worst-case simulations, as shown in Fig. 5. In these simulations,



Test Chip Summary	
Process	16 nm FinFET
Die Size	2 mm x 2 mm
Core Voltage	0.80 V
Memory Area	106 µm x 416 µm
Frequency Range	10 - 800 MHz
Temperature Range	0 - 85°C

Fig. 6. 3T GCRAM test chip micrograph with a summary table.

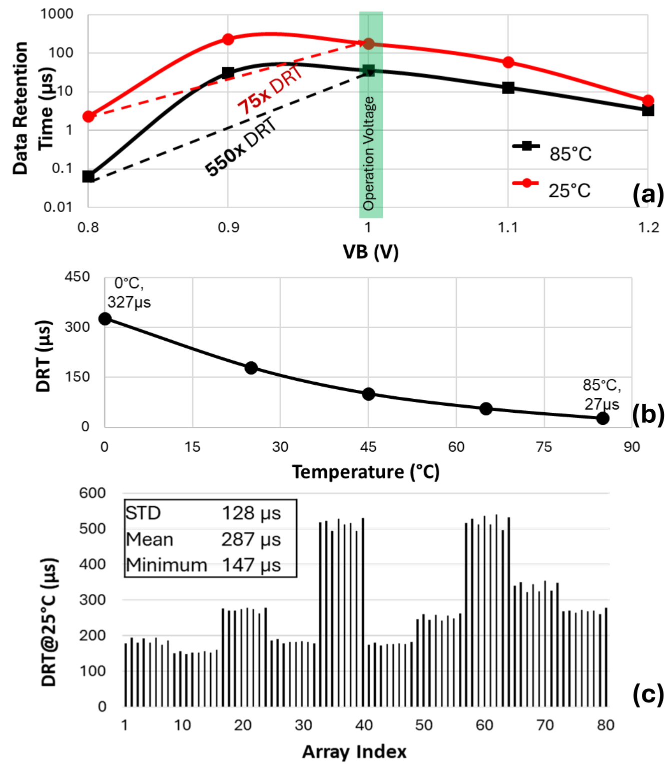


Fig. 7. (a)Boosted positive supply (*VB*) effect on retention time for 25ºC and 85ºC.(b)Measurement results of the DRT over temperature with internal boosted supplies. (c) DRT measurement of 80 arrays across 10 dies at 25ºC .

the write transistor (M_W) is shifted by -6σ to model a high leakage condition during retention for evaluating the effect of the VB assist. Similarly, a +6σ shift is applied to MW to emulate a high-threshold scenario during the write operation for analyzing the *VSS* assist. The simulation results show that the DRT can be improved by up to 600× under worst-case conditions with appropriate assist voltages.

III. Measurement Results

A test chip (Fig.6) containing the 3T 512-kb GCRAM macro was fabricated in 16 nm FinFET technology and integrated into an SoC with a RISC-V core and built-in self-test (BIST) logic. The memory was tested to characterize retention

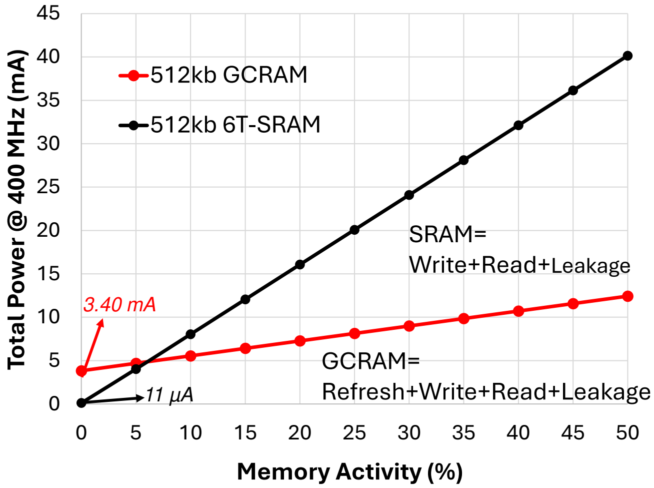


Fig. 8. Measured power results for same size GCRAM and 6T SRAM memories while operating at 25°C and 400 MHz on a single chip. Memory activity composed of %50 write and 50% read operations.

Table. 1. Measured power results for same size GCRAM and 6T-SRAM memories while operated at 25°C and 0.8 V.

	[6]	[7]	[7]	[1]	This work
Bit Cell Topology	6T SRAM	8T 2P-SRAM	6T SRAM	3T GCRAM	**3T GCRAM**
Number of Ports	1	2	2	2	**2**
Process	16 nm	16nm	16 nm	16 nm	**16 nm**
Cell Area (µm2)	0.073 (pushed)	0.138 (pushed)	Not reported	0.080 (logic)	**0.069 (logic)**
Memory Size	128 kb	64 kb	256 kb	1 Mb	**512 kb**
Macro Area / bit (μm^2/bit)	0.103* (pushed)	-	0.172 (logic)	0.12 (logic)	**0.084 (logic)**
External Assist Voltages	-	-	-	2	**-**
I/O Width	32	32	32	64	**1024**
DRT (µs)	-	-	-	77@25ºC 8@125ºC	**147@25ºC 27@85ºC**
Leakage Power (nA/bit)	Not reported	Not reported	Not reported	175	**39**
Write Power@ 25ºC (nA/MHz/bit)	Not reported	Not reported	Not reported	640	**16**
Read Power@ 25ºC (nA/MHz/bit)	Not reported	Not reported	Not reported	670	**26**

*Estimated from chip area

time under varying *VB* conditions—supplied externally during characterization and internally during functional and DRT tests—and across temperature (Fig.7(a)). The effect of the boosted *WWL* driver during retention on performance was measured to be 75× at 25ºC and 550× at 85ºC. This increase aligns with expectations, as subthreshold leakage effects become more dominant at higher temperatures. Furthermore, GCRAM macros tested under different temperature conditions (Fig.7(b)) showed DRT values of 327 µs at 0ºC and 27 µs at 85ºC. As expected, the DRT decreases with increasing temperature due to higher leakage currents, especially subthreshold leakage. Additionally, 80 different arrays from 10 different dies were tested with internal boost voltage at 25ºC, and the minimum DRT recorded was 147 µs across 40 Mbits.

The power consumption of the GCRAM array compared to foundry 6T-SRAM is plotted in Fig. 8. According to the measurements, the GCRAM macro consumes 13.1 µW/MHz and 21.3 µW/MHz for write and read operations, respectively, while operating at 400 MHz with a 0.8 V supply. Despite the additional refresh power, the power consumption of 6T-SRAM exceeds that of GCRAM when memory activity exceeds 5%, due to its significantly reduced access energy.

Table. I shows a comparison table where the proposed circuit demonstrates better performance and area compared to the state-of-the-art GCRAM macros [1]. The proposed logic-rule macro has a smaller area when compared to SRAMs [6]–[8], where [6] and [7] are high-density SRAMs based on the "pushed-rule" foundry bitcell.

Conclusion

This work presents a high-density, single-supply 512-kb 3T GCRAM macro implemented in 16 nm FinFET technology, achieving a record 0.069 μm^2 bitcell area and 147 μs retention time at 25°C. By integrating on-chip voltage generation and employing novel circuit techniques, the proposed design significantly reduces both area and power consumption compared to conventional SRAM. The proposed flipped bitcell topology improves steady-state storage voltage and enables read currents that are largely independent of retention time, contributing to both robustness and scalability. Measurement results demonstrate robust performance across process and temperature variations, with improved energy efficiency. These findings position GCRAM as a promising embedded memory solution for next-generation SoCs.

References

[1] R. Giterman, A. Shalom, A. Burg, A. Fish, and A. Teman, "A 1-Mbit Fully Logic-Compatible 3T Gain-Cell Embedded DRAM in 16-nm FinFET," *IEEE Solid-State Circuits Letters*, vol. 3, pp. 110–113, 2020.

[2] O. Harel *et al.*, "A 16-kB 65-nm GC-eDRAM Macro With Internal Bias Voltage Generation Providing Over 100-μs Retention Time," *IEEE Journal of Solid-State Circuits*, pp. 1–10, 2024.

[3] K. C. Chun, P. Jain, T.-H. Kim, and C. H. Kim, "A 667 MHz Logic-Compatible Embedded DRAM Featuring an Asymmetric 2T Gain Cell for High Speed On-Die Caches," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 2, pp. 547–559, 2012.

[4] D. Kilani, M. Alhawari, B. Mohammad, H. Saleh, and M. Ismail, "An Efficient Switched-Capacitor DC-DC Buck Converter for Self-Powered Wearable Electronics," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 10, pp. 1557–1566, 2016.

[5] A. Yigit, E. N. Casarrubias, R. Giterman, and A. Burg, "A 128-kbit GC-eDRAM With Negative Boosted Bootstrap Driver for 11.3× Lower-Refresh Frequency at a 2.5

[6] Y.-H. Chen *et al.*, "A 16 nm 128 Mb SRAM in High- κ Metal-Gate FinFET Technology With Write-Assist Circuitry for Low-VMIN Applications," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 170–177, 2015.

[7] H. Fujiwara *et al.*, "17.2 A 64kb 16nm asynchronous disturb current free 2-port SRAM with PMOS pass-gates for FinFET technologies," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3.

[8] M. Yabuuchi *et al.*, "16 nm FinFET High-k/Metal-gate 256-kbit 6T SRAM macros with wordline overdriven assist," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 3.3.1–3.3.3.