A 1-Mbit Fully Logic-Compatible 3T Gain-Cell Embedded DRAM in 16-nm FinFET

Robert Giterman, Amir Shalom, Andreas Burg, *Member, IEEE*, Alexander Fish, *Member, IEEE*, and Adam Teman, *Member, IEEE*

***Abstract*—Gain-cell embedded DRAM (GC-eDRAM) is a logic-compatible embedded memory alternative to SRAM, offering higher density, lower leakage power consumption, and an inherent two-ported functionality. However, increased leakage currents and process variations under technology scaling lead to a reduced data retention time (DRT), resulting in increased refresh power and reduced memory availability, currently limiting its implementation to planar 28-nm technologies and above. This letter presents the first GC-eDRAM in 16-nm FinFET technology, featuring a mixed-V_T 3T gain-cell structure to minimize the storage node (SN) leakage. The implemented 1-Mbit 3T GC-eDRAM is fully logic-compatible and provides a 2× smaller bitcell size compared to a 6T SRAM with similar design rules, offering the highest density logic-compatible memory cell in 16-nm technology. Measurement results demonstrate a 77-μs DRT under a 600-mV V_{DD}, which is over 10× longer than previously reported GC-eDRAMs in 28-nm technologies. The memory was fully operational at temperatures spanning −40 °C to 125 °C and under a supply voltage as low as 450 mV, providing the lowest measured V_{DDmin} and widest temperature range reported in the literature for GC-eDRAM.**

***Index Terms*—Embedded DRAM, gain cell (GC), low voltage, retention time, SRAM.**

I. Introduction

With the increasing demand for memory capacity, the die size and power consumption of modern Systems-on-Chip (SoCs) are often dominated by memory. While 6T SRAM has been the topology of choice for most systems, its size and poor energy efficiency often limit the cost and performance of emerging applications such as machine learning accelerators [1], [2]. 1T-1C eDRAM offers a denser memory alternative to SRAM, however, its additional fabrication steps and limited process-availability hinders its usage in the state-of-the-art applications. Gain-cell embedded DRAM (GC-eDRAM) is a fully logic-compatible alternative to SRAM, offering a smaller bitcell size, nondestructive read operation, and inherent two-ported functionality [3]–[9]. However, GC-eDRAM implementation has become more challenging in advanced processes due to increased leakage currents and process variations and reduced storage node (SN) capacitance. These factors have led to a decreased data retention time (DRT), causing a higher refresh power consumption and reduced memory availability, with GC-eDRAM demonstrations only available down to the 28-nm node [7], [8]. In this letter, we present the first GC-eDRAM implementation in a FinFET process. The memory macro was implemented in a standard commercial 16-nm technology, making it the first demonstration of GC-eDRAM beyond 28 nm. The implemented array relies on a 3-transistor (3T) bitcell, which offers 2× better density compared to 6T SRAM with similar design rules. The measured DRT of a 1-Mbit array was found to be 77 μs, which is over 10× better than recently reported GC-eDRAMs implementations in 28-nm technologies [7], [8]. The manufactured arrays were fully functional at temperatures between −40 °C and 125 °C and at a V_{DDmin} as low as 450 mV. These are the lowest measured V_{DDmin} and the widest temperature range reported in the literature for GC-eDRAM.

Manuscript received May 3, 2020; revised June 23, 2020; accepted June 24, 2020. Date of publication July 2, 2020; date of current version July 23, 2020. This work was supported by the HiPer Consortium funded by the Israeli Innovation Authority. This article was approved by Associate Editor Stefan Rusu. *(Corresponding author: Robert Giterman.)*

Robert Giterman and Andreas Burg are with the Telecommunications Circuits Laboratory, Institute of Electrical Engineering, EPFL, 1015 Lausanne, Switzerland (e-mail: robert.giterman@epfl.ch).

Amir Shalom, Alexander Fish, and Adam Teman are with the Emerging Nanoscaled Integrated Circuits and Systems Labs, Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel.

Digital Object Identifier 10.1109/LSSC.2020.3006496

II. Implementation

FinFET processes offer better channel control and less variability compared to planar Bulk technologies. Therefore, carefull bitcell device choices and assist circuitry are critical to achieve a good compromise between the various leakage currents affecting the SN to maximize DRT. Fig. 1(a) depicts the 3T gain cell (GC) implemented in 16-nm FinFET technology with its main leakage paths which deteriorate the stored level. While pMOS devices were traditionally used to implement the GC write port due to their lower subthreshold (sub-V_T) leakage compared to nMOS, the 16-nm process used in this letter provides similar Ion/Ioff characteristics for both transistor types and, therefore, an nMOS write device with a single fin was used in order to better balance the leakage currents affecting the SN. The voltage of the write word line (WWL) is lowered below GND (VNEG) to further suppress the sub-V_T leakage, while a boosted supply (V_{DD} +300 mV) is needed to transfer a full 1 level during write. pMOS transistors were chosen to implement the read port of the 3T GC due to their lower gate leakage. Moreover, high-V_T devices were selected in order to suppress the current between the read bit line (RBL) and V_{DD} during standby, leading to less leakage and a better read margin. Fig. 1(b) shows the resulting SN degradations of both 1 and 0 as simulated around a TT corner, including transistor mismatch using 10K Monte-Carlo runs, with VNEG set to −150 mV. The selected 3T structure helps to compensate the various leakage currents to mitigate the average SN voltage deterioration, resulting in a balanced degradation of both 1 and 0, contributing to the enhancement of the DRT.

The implemented 1-Mbit GC-eDRAM, based on the 3T GC, is composed of four 256-kbit macros, each containing 16 256 × 64 (16 kbit) subarrays, as shown in Fig. 2. The size of the full 1 Mbit GC-eDRAM is 295 μs × 446 μm, while each 16-kbit subarray measures at 34 μs × 56 μm with the bit-cell array accounting for 76%. Each subarray contains local data-in (DI) latches, read word line (RWL) and WWL enable logic and drivers, and inverter-based sense amplifiers. The 3T GC layout was drawn according to standard logic design rules and measures at a height of two poly pitches (2PPs) and width of eight diffusion pitches (DPs), which is 2× denser compared to a 6T SRAM drawn with the same design rules.

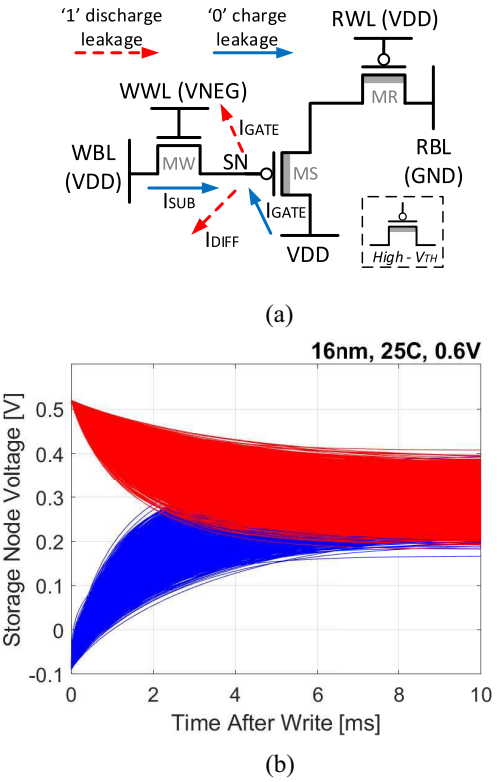


Fig. 1. (a) Implemented 3T GC. (b) Data degradation curves of "1" (red) and "0" (blue) across process variations after write.

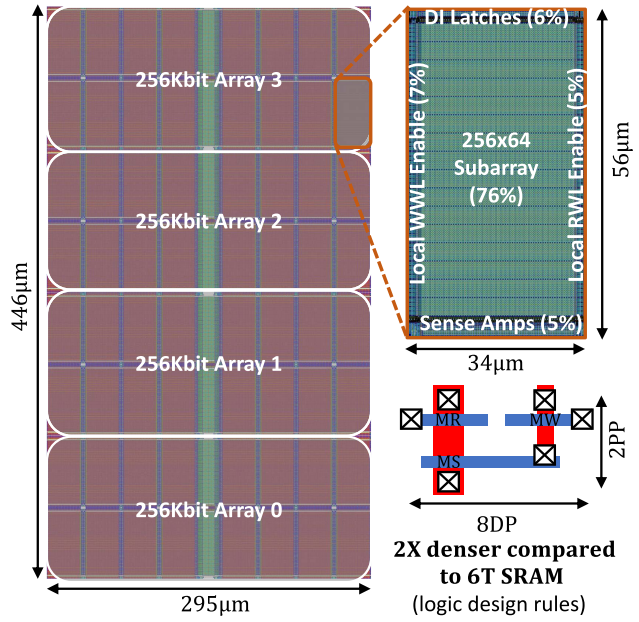


Fig. 2. Layout views of the 1 Mbit 3T GC memory macro.

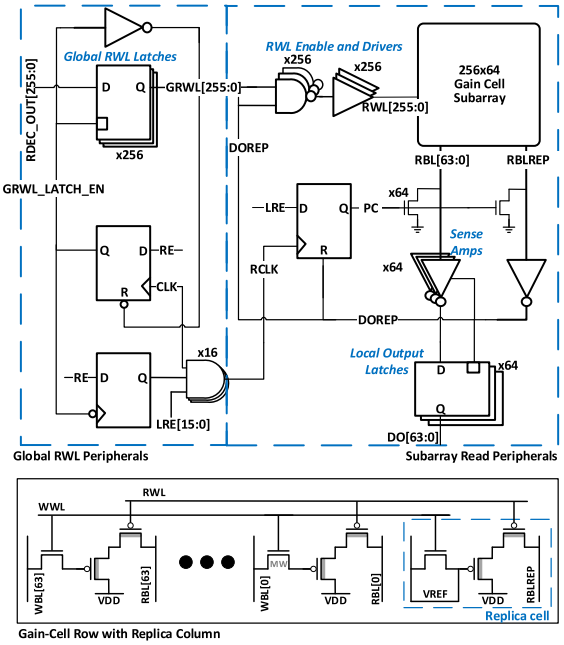
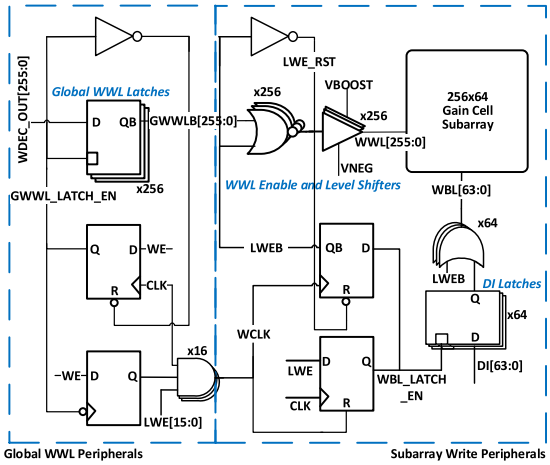
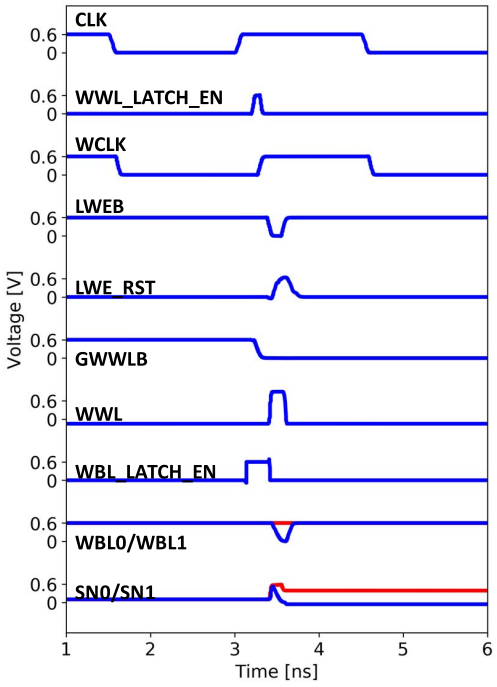


Fig. 3. (a) Schematics of the global (256-kbit macro) and local (16-kbit subarray) read peripherals. (b) Waveform illustration of the read operation obtained from simulations.

Figs. 3 and 4 depict the schematics of the write and read peripherals of the memory and key waveforms illustrating their operating principle. Each 256-kbit macro contains write and read decoders and global WWL and RWL pulsed latches to drive the GRWL and GWWLB signals to all subarrays. The global peripherals also generate local write and read clocks (WCLK and RCLK) which delay the evaluation of the local word-lines until the GRWL and GWWLB have been evaluated. The local (subarray) read peripherals, depicted in Fig. 3 with a corresponding waveform illustration, implement a self-timed read control using dedicated replica cells, which are integrated in a single column of each subarray. The replica cells are similar to the 3T GC with an external reference voltage (VREF) connected to their SN. The assertion of RCLK initiates a precharge (PC) pulse that discharges the RBLs. As soon as the replica RBL (RBLREP) is discharged, it resets the PC pulse using the DOREP signal, which is also used to enable the selected RWL. The RWL is enabled until RBLREP is charged above the threshold of the sense inverter, which de-asserts DOREP to latch the data outputs and to disable RWL. The write cycle, illustrated in Fig. 4, begins with the latching of the DI signals, which is enabled during the GWWL evaluation and disabled upon the assertion of the



(a)



(b)

Fig. 4. (a) Schematics of the global (256-kbit macro) and local (16-kbit subarray) write peripherals. (b) Waveform illustration of the write operation obtained from simulations.

WCLK. With the arrival of WCLK, the selected WWL is enabled to transfer the levels on the WBLs to the SNs of the selected row. Then, following a write delay set by the LWE_RST signal, WWL is de-asserted and all write bit lines (WBLs) are charged to V_{DD} to suppress the leakage of 1 s from the SNs.

III. Test-Chip and Measurement Results

An SoC with the GC-eDRAM was implemented in 16-nm FinFET technology. The GC-eDRAM subsystem enables several access modes to the 1-Mbit array, including L3 cache access using a CEVA-X2 DSP core, direct built-in-self-test (BIST) access for memory characterization, and access with EFLX eFPGA for additional tests. VNEG and VREF were supplied externally through dedicated I/O pins. The test-chip microphotography and key features of the GC-eDRAM are shown in Fig. 5. Measurement results of

Test-Chip Summary	
Process	16nm FinFET
Die Size	3.7mm X 3mm
Memory Capacity	1Mbit
Bitcell Type	Mixed 3T Gain-Cell
Bitcell Size	52% of 6T SRAM
Periphery Type	Mixed SVT/LVT
1Mbit Memory Size	295um X 446um
Voltage Range	0.45V – 0.8V
Frequency Range	Up-to 400MHz @ 0.8V
Temperature Range	-40C – 125C
Retention Time (99.9% bit yield)	77us @ 0.6V, 25C 8us@ 0.6V, 125C
Leakage Power	105uW/Mbit @ 0.6V, 25C
Refresh Power	276uW/Mbit @ 0.6V, 25C
Read Power	2.58uW/MHz @ 0.6V, 25C
Write Power	2.46uW/MHz @ 0.6V, 25C

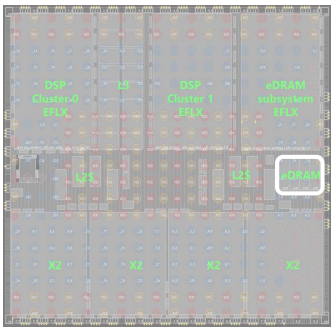


Fig. 5. Microphotography and key features of the implemented test-chip.

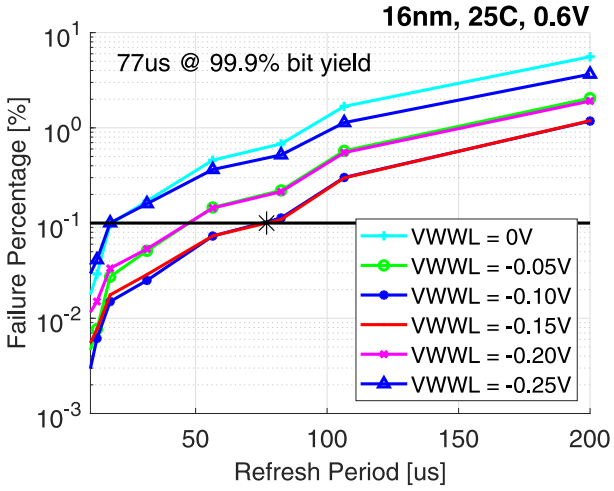


Fig. 6. DRT failure rate versus VWWL.

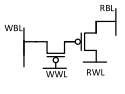
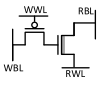
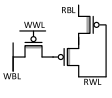
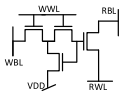
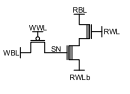
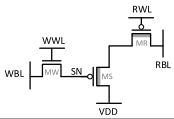
the developed 1-Mbit GC-eDRAM illustrate full functionality across −40 °C to 125 °C and down to 450 mV (with a 44-μs DRT). A 600-mV supply was selected for the remaining measurements as it provided sufficient frequency and DRT for the L3 cache used in this system. The targeted 99.9% bit-yield[1] DRT was found to be 105, 77, and 8 μs under worst-case WBL biasing, for −40 °C, 27 °C, and 125 °C, respectively. The DRT measurement was repeated on five additional chip samples, providing the same or better results. To put this into perspective, the measured 99.9% bit-yield of a commercial 1T1C eDRAM employing deep-trench capacitors in 14-nm FinFET technology was 90 μs at 85 °C under a −250 mV VNEG.

Fig. 6 illustrates the bit failure percentages under a WWL voltage sweep during standby. As the WWL bias voltage of unselected rows is lowered, the sub-V_T leakage is decreased, however, gate leakage compromising a stored 1 is increased. With a WWL bias of −150 mV, the DRT is increased by over 5× compared to zero bias. Therefore, −150 mV was selected as the optimal negative bias for the remainder of the measurements.

Fig. 7 depicts the measured DRT and the maximum frequency across varying memory supply voltages, with VREF = V_{DD}/3. The maximum frequency of the 1-Mbit array was found to be 400 MHz under 800 mV. The frequency can be further increased by integrating

[1]The bit-yield is defined as the percentage of bits which maintained both 1 and 0 under the given refresh period. is defined as the percentage of bits which maintained both 1 and 0 under the given refresh period.

TABLE I
COMPARISON BETWEEN THE PROPOSED DESIGN AND OTHER LOGIC-COMPATIBLE MEMORY OPTIONS

	2009'ISSCC [3]	2010'VLSI [4]	2012'VLSI [5]	2015'JSSC [6]	2018'JSSC [7]	2019'ASSCC [8]	Proposed
Cell Structure							
Technology Node	65nm Bulk	65nm Bulk	65nm Bulk	65nm Bulk	28nm Bulk	28nm FD-SOI	16nm FinFET
Redrawn Cell Size (μm^2)*	0.51μm^2	0.51μm^2	0.60μm^2	0.51μm^2	0.25μm^2	0.16μm^2	0.08μm^2
Cell Ratio to 6T SRAM*	0.44X	0.44X	0.51X	0.44X	0.71X	0.49X	0.52X
Memory Capacity	2Mbit	192kbit	34kbit	24kbit	8kbit	24kbit	1Mbit
Retention Time (@25C)	10us** @ 1.2V (100% bit yield)	110us** @ 1.1V (99.9% bit yield)	1us @ 1.2V (100% bit yield)	50us @ 1.2V (99% bit yield)	5us @ 0.9V (99% bit yield)	4.7us @ 0.9V (99.9% bit yield)	77us @ 0.6V (99.9% bit yield)
Min. Voltage	0.8V	0.8V	0.69V	0.9V	0.6V	0.7V	0.45V (44us data retention)
Min. Latency	4ns @ 1.1V	1.5ns @ 1.1V	1.8ns @ 1.3V	4.6ns @ 1.1V	1.25ns @ 0.9V	2.5ns @ 0.9V	2.5ns @ 0.8V
Temperature Range	NA	25C – 85C	25C – 125C	25C – 85C	0C – 85C	NA	-40C – 125C

*Logic design rule ** 85C

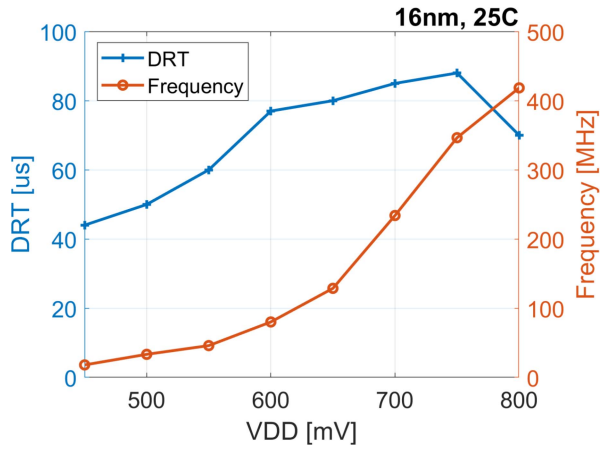


Fig. 7. DRT and maximum frequency versus V_{DD}.

differential sense amplifiers and through reduction of VREF, which reduces the RBL evaluation phase. The DRT reduces below 600 mV and above 750 mV due to increased impacts of process variations and sub-V_T leakage, respectively.

Table I compares the proposed memory to previously reported GC-eDRAMs. The proposed array is the first GC-eDRAM in 16-nm FinFET, providing 2× reduced cell size compared to a 6T SRAM with same design rules. The measured DRT is over 10× higher compared to 28-nm implementations and has the lowest operational V_{DDmin} and the largest temperature range among the reported GC-eDRAMs.

IV. CONCLUSION

This letter presented the first implementation of logic-compatible GC-eDRAM in a 16-nm FinFET technology. A 1-Mbit memory was integrated in an SoC with a 2× higher bitcell density compared to a 6T SRAM with similar design rules. Measurements results illustrate a 77-μs DRT under a 600-mV V_{DD}, which allows for a high memory availability with a parallel refresh and low refresh power consumption. The implemented array was fully functional under −40 °C to 125 °C and down to 450-mV V_{DD}, which shows that the leakage issue of GC-eDRAM can be kept under control even for high temperatures and that GC-eDRAM is also an option for low-voltage operation.

ACKNOWLEDGMENT

The authors would like to thank R. Golman, M. Goldzwig, T. Noy, and Y. Shoshan for their help in the design and measurements of the test-chip.

REFERENCES

[1] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[2] K. Ando *et al.*, "BRein memory: A single-chip binary/ternary reconfigurable in-memory deep neural network accelerator achieving 1.4 TOPS at 0.6 W," *IEEE J. Solid-State Circuits*, vol. 53, no. 4, pp. 983–994, Apr. 2018.

[3] D. Somasekhar *et al.*, "2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 174–185, Jan. 2009.

[4] 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 J. Solid-State Circuits*, vol. 47, no. 2, pp. 547–559, Feb. 2012.

[5] Y. S. Park, D. Blaauw, D. Sylvester, and Z. Zhang, "Low-power high-throughput LDPC decoder using non-refresh embedded DRAM," *IEEE J. Solid-State Circuits*, vol. 49, no. 3, pp. 783–794, Mar. 2014.

[6] W. Choi, G. Kang, and J. Park, "A refresh-less eDRAM macro with embedded voltage reference and selective read for an area and power efficient Viterbi decoder," *IEEE J. Solid-State Circuits*, vol. 50, no. 10, pp. 2451–2462, Oct. 2015.

[7] R. Giterman, A. Fish, N. Geuli, E. Mentovich, A. Burg, and A. Teman, "An 800-MHz mixed- V_T 4T IFGC embedded DRAM in 28-nm CMOS bulk process for approximate storage applications," *IEEE J. Solid-State Circuits*, vol. 53, no. 7, pp. 2136–2148, Jul. 2018.

[8] J. Narinx *et al.*, "A 24 kb single-well mixed 3T gain-cell eDRAM with body-bias in 28 nm FD-SOI for refresh-free DSP applications," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Macau, China, 2019, pp. 219–222.

[9] A. Shalom, R. Giterman, and A. Teman, "High density GC-eDRAM design in 16nm FinFET," in *Proc. 25th IEEE Int. Conf. Electron. Circuits Syst. (ICECS)*, Bordeaux, France, 2018, pp. 585–588.