ARTICLES

Journal of Semiconductors
(2023) 44, 044102
doi: 10.1088/1674-4926/44/4/044102

Temperature-insensitive reading of a flash memory cell

Weiyan Zhang[1, 2], Tao Yu[2], Zhifeng Zhu[1], and Binghan Li[2, †]

[1]School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

[2]Shanghai Huahong Grace Semiconductor Manufacturing Corporation, Shanghai 200125, China

Abstract: The temperature characteristics of the read current of the NOR embedded flash memory with a 1.5T-per-cell structure are theoretically analyzed and experimentally verified. We verify that for a cell programmed with a "10" state, the read current is either increasing, decreasing, or invariable with the temperature, essentially depending on the reading overdrive voltage of the selected bitcell, or its programming strength. By precisely controlling the programming strength and thus manipulating its temperature coefficient, we propose a new setting method for the reference cells that programs each of reference cells to a charge state with a temperature coefficient closely tracking tail data cells, thereby solving the current coefficient mismatch and improving the read window.

Key words: flash memory; temperature coefficient; reference cell; flash array

Citation: W Y Zhang, T Yu, Z F Zhu, and B H Li, Temperature-insensitive reading of a flash memory cell[J]. *J. Semicond.*, 2023, 44(4), 044102. https://doi.org/10.1088/1674-4926/44/4/044102

1. Introduction

Flash is the most widely-used nonvolatile memory technology and is used in electronic devices, smart cards, and microcontrollers to store data with high efficiency and reliability. Compared to EEPROM, which has a two-transistor-per-cell structure, the split gate flash memory with a 1.5-transistor-per-cell structure can store more data with less space and can erase data in blocks instead of in bytes, thus improving storage density and also operation efficiency, and enhancing integration.

The basic cell structure of the flash memory is just a MOSFET with an extra gate trapping and de-trapping electrons, thus manipulating its threshold voltage to store information. Like MOSFET, the flash cell working performance is susceptible to temperature variation, which results from the dependence of the drain–source current on temperature. This thermal-induced effect can sometimes be useful, such as in designing temperature sensors or temperature memories[1], but in most cases of flash applications, this thermal effect is considered undesirable because it gives bad working performances and various reliability issues[2–4]. Many methods have been proposed to avoid the temperature effect, such as adding extra peripheral circuits providing temperature bias control signals to improve voltage distribution[5, 6], using dummy cells as references to compensate for the temperature effect of current[7, 8], or resorting to complicated algorithms[9]. However, these methods consume much power and area and reduce storage capacity. We can make use of the adjustable threshold voltage of the stacked gate memory itself to avoid this temperature effect, and this temperature effect can be very useful in memory arrays under proper suppression and precise control.

In this paper, we propose a new preprogramming compensation method to suppress and control the temperature drift of the read current in flash memory. The new method is based on the compensation of the dependence of carrier mobility and threshold voltage on temperature. Moreover, we also briefly introduce its applications in a reference cell. This paper is organized as follows. We first give a brief introduction of the structure of the flash cell that is used in the experiment. The physics theory and the results of the experiment are then discussed and analyzed, which is followed by the introduction of the proposed reference cells. Finally, we summarize and draw a conclusion.

2. Cell structure and operations

The experiment is conducted on a split gate flash memory that is fabricated on the 90 nm self-aligned process platform of Huahong Grace Semiconductor Manufacturing Corporation (HHGrace), as shown in Fig. 1. It has a 2-bit per cell symmetric structure with two floating gates (FGs) to store the information, two bitlines (BLs) and one wordline (WL) to locate a single cell in a flash array, and two control gates (CGs) to assist programming. Essentially, the structure of this cell can be viewed as three sub-transistors (two bitcell transistors and one WL transistor) that are connected in series, which respectively control the channel beneath each of them. This new cell inherits the high program efficiency character of the conventional split-gate device and fast erase is achieved by poly-to-poly FN tunneling[10]. Compared to the conventional structure, it has a smaller unit area, and higher operation efficiency and reliability.

The process of injecting the electrons into FG to increase the threshold voltage is called programming, and repelling electrons from FG is called erasing. For this cell, the programming operation is done by applying higher bias (~9 V) on CGs to boost source side hot electrons injecting into the FG. Erasing is done by biasing CGs with a negative voltage (~–7.8 V) and WL with a much higher voltage (~8.3 V) to

Correspondence to: B H Li, Gordon.Li@hhgrace.com

Received 29 AUGUST 2022; Revised 31 OCTOBER 2022.

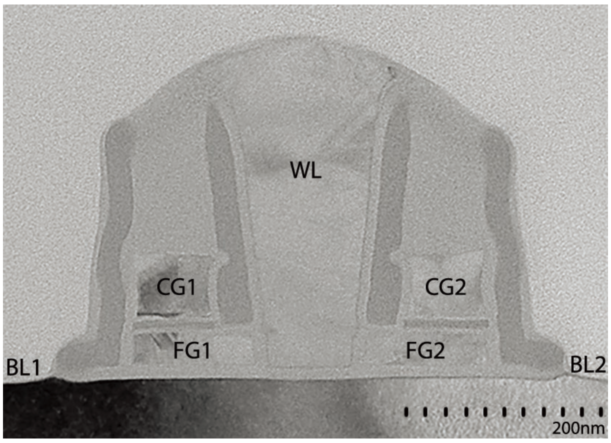

Fig. 1. The cross-section of the flash cell structure.

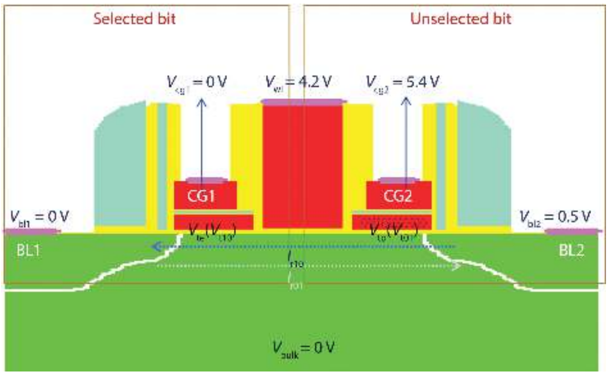

Fig. 2. (Color online) The cell schematic of reading and subscript denotation.

boost electrons tunneling from FG to WL. We assign the state of “1” to the erased cells, and “0” to the programmed cells. Due to the injection of the electrons, a programmed cell has a higher threshold voltage, and thus a lower drain-source current under the same gate–source voltage bias than that of those unprogrammed ones.

When reading a selected cell, a high voltage bias should be applied on its WL to open the channel beneath it. Thus, the drain–source current of a memory cell is dominated by the minimum difference between CG voltage and threshold voltage. Since the threshold voltage of a memory cell depends on the programming state, we can read the information (“0” or “1”) stored in the cell by comparing its drain-source current to a reference value. As is shown in Fig. 2, the device utilizes a reverse-read scheme[11].

When analyzing the properties of a certain bitcell of two, we add a binary subscript on this property to indicate which one was referenced, such as “V_{t01}”, which is the threshold voltage of the right-hand bitcell; and “I_{r10}”, which is the read current of the left-hand bitcell. In addition, the threshold voltage of a programmed bitcell is denoted by “V_{tp}”, and the threshold voltage for an erased bitcell is denoted by “V_{te}”. For the same bit pair, there is a slight difference between the voltage and current that it represents. Take “10” as an example, “V_{t10}” denotes exactly the threshold voltage of the left-hand bitcell transistor, which is irrelevant to the WL transistor and the right-hand bitcell transistor. However, “I_{r10}” represents the current that we can measure when reading the left-hand bitcell, and its value may vary with one of these three sub-transistors which has the minimum overdrive voltage. No matter which bitcell is read, since there is no junction implanted under WL, the current that we measure is the drain–source current of the entire device, which flows from one BL to the other. The difference is the bias voltage applied on the CGs of different bitcells, which is shown in Table 1.

Table 1. Reading operations of this split gate flash memory cell.

Bitcell	CG1 (V)	CG2 (V)	WL (V)		BL1 (V)	BL2 (V)
			Sel	Unsel		
Bit1(I_{r10})	0	5.4	4.2	0	0	0.5
Bit2(I_{r01})	5.4	0			0.5	0

3. Theory and experiment

3.1. Physical theory

The read current of a flash cell is given by:

$$I_{\rm D} = \mu C_{\rm ox}\frac{W}{L}\left[\left(V_{\rm GS} - V_{\rm TH}\right)V_{\rm DS} - \frac{1}{2}V_{\rm DS}^2\right], \tag{1}$$

where μ, $C_{\rm ox}$, and W/L are the parameters of the intrinsic characteristics of the device and are irrelevant to reading. $V_{\rm GS}$ and $V_{\rm DS}$ are the voltage put on CG and same-side bitline, respectively. $V_{\rm TH}$ is the threshold voltage of the floating-gate transistor, and has the negative coefficient with the temperature as[12–14]:

$$V_{\rm TH} = V_{\rm TH0} + \alpha_{\rm vt}\left(T - T_0\right), \quad \alpha_{\rm vt} = \frac{\partial V_{\rm TH}}{\partial T} < 0, \tag{2}$$

where μ is the mobility of the carriers, and is also negatively related to temperature：

$$\mu = \mu_0\left(\frac{T}{T_0}\right)^{\alpha_\mu}, \quad \alpha_\mu = \frac{\partial \mu}{\partial T} < 0. \tag{3}$$

Therefore, the drain current has a zero-temperature coefficient (ZTC), where the current variation induced by mobility and threshold voltage with respect to temperature can be totally compensated, thus maintaining the current stable at a certain value.

For the cell used in this experiment, $V_{\rm DS} \ll 2\left(V_{\rm GS} - V_{\rm TH}\right)$, so

$$\frac{\partial I_{\rm D}}{\partial T} = \frac{\partial \mu}{\partial T}C_{\rm ox}\frac{W}{L}\left[\left(V_{\rm GS} - V_{\rm TH}\right)V_{\rm DS}\right] - \mu C_{\rm ox}\frac{W}{L}\left[\frac{\partial V_{\rm TH}}{\partial T}V_{\rm DS}\right] \tag{4}$$

$$= C_{\rm ox}\frac{W}{L}V_{\rm DS}\left[\alpha_\mu\left(V_{\rm GS} - V_{\rm TH}\right) - \mu\alpha_{\rm vt}\right]. \tag{5}$$

Moreover, the mobility also decreases with increasing overdrive voltage as:

$$\mu_{\rm eff} = \frac{\mu_0}{\left[1 + \dfrac{\theta C_{\rm ox}\left(V_{\rm GS} - V_{\rm TH}\right)}{\kappa_{\rm sj}\varepsilon_0}\right]}, \tag{6}$$

where θ is an empirical positive parameter found to be technologically and substrate-bias dependent. Eqs. (5) and (6) indicate that the temperature coefficient is not a constant but varies with the mobility and threshold voltage. If the mobility variation dominates this trend, then the drain–source current will

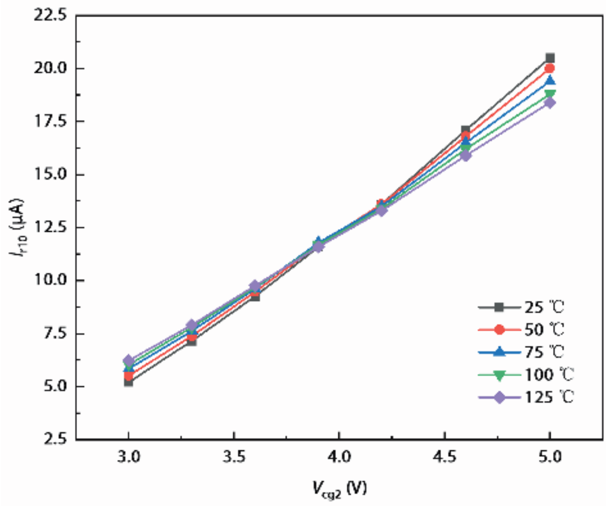

Fig. 3. (Color online) Read current in different temperatures with proposed programming condition (V_{tp} = 2.30 V, V_{te} = −8 V at 25 °C) .

decrease with increasing temperature; if the threshold variation dominates, then the current will be positively related to the temperature. Therefore, by controlling dopants or adjusting bias voltage, the effect of temperature on mobility and threshold can be partly or even completely canceled, which can be used to adjust the temperature coefficient of the current as required.

Specifically, by setting $\frac{\partial I_D}{\partial T}$ = 0, we can derive that

$$\frac{\alpha_\mu}{\alpha_{vt}} = \frac{\mu}{V_{GS} - V_{TH}}. \tag{7}$$

This implies that with a proper value of overdrive voltage $V_{GS} - V_{TH}$, the temperature coefficient can be close to zero. In other words, the temperature coefficient can be set precisely by controlling the overdrive voltage.

3.2. Experiment and results

In the experiment, all bitcells are first initialized with the threshold voltage V_{tp} = 2.30 V and V_{te} = −8 V at room temperature, and V_{cg1} = 0 V. By first adjusting the voltage of CG2 and then programming different cell samples with different conditions (e.g., time, bias voltage, temperature, etc.), we successfully set up cells with different temperature coefficients, thus verifying the variation of current with temperature. We determine the overdrive voltage of reference cells in three steps. First, we measure the read current of the tail cells under normal stressed voltage and we then calculate the temperature coefficients by analyzing the data obtained by the variable temperature test. Second, we select the memory cells located on different dies and group them, each group corresponding to the different read voltage. We then put them through the variable temperature test to find those test cells with the same temperature coefficients as the tail cells and measure their thresholds. Finally, we can get the overdrive voltage by subtracting the threshold voltage from the applied gate voltage. Furthermore, we verified the existence of the ZTC point and controllability of the temperature coefficients in the flash memory cells. The results are shown in Fig. 3.

The current I_{r10} for a constant V_{tp} and different V_{cg2} in different temperatures are tested, as illustrated in Fig. 3. As the control gate voltage increases from 3 to 5 V, the read current first increases and then decreases with the increasing temperature, indicating two different variation intervals dominated by mobility and threshold variation respectively. When V_{cg} = 3 V, the current increases with temperature, i.e., $-\alpha_{vt}$ dominates this trend more than α_μ in formula (5). As V_{cg2} increases, the coefficient of item α_μ (i.e., $V_{GS} - V_{TH}$) increases, and at the time, the increased V_{cg2} decreases the mobility (μ), the coefficient of α_{vt}. Both effects weaken the role of α_{vt} and strengthen the role of α_μ, and thus the current decreases with the increasing temperature as V_{cg} increases to 5 V. Specifically, at about V_{cg2} = 3.9 V, the variation of threshold and mobility to the temperature can be compensated, as formula (7) indicates. And the read current shows an insensitive characteristic to the temperature, exhibiting a ZTC point, which is illustrated in Fig. 4.

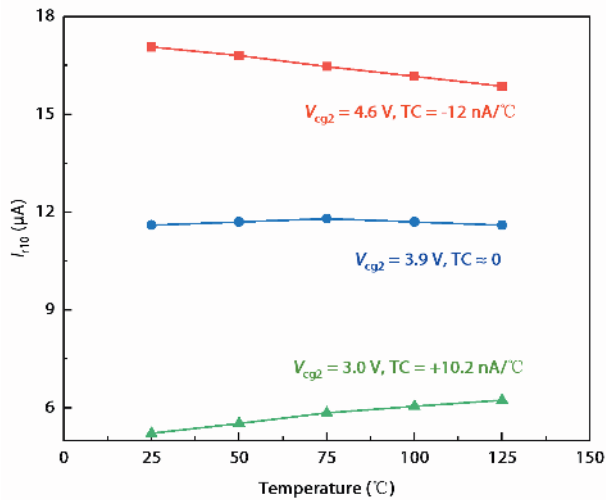

Fig. 4. (Color online) Zero temperature coefficient point of the read current.

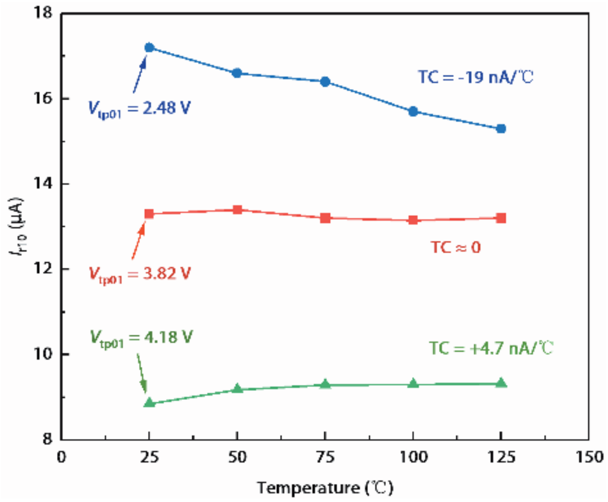

Fig. 5. (Color online) Current in different temperature with fixed CG bias (V_{cg2} = 5.4 V) and different threshold voltage (V_{tp01}).

Since the threshold voltage is initialized to 2.30 V, we now find that the overdrive voltage (i.e., the voltage difference between CG2 and threshold) is approximately 1.6 V. According to formula (6), we reset V_{cg2} = 5.4 V and change V_{th01} by pre-programming with different times. The longer time a cell is pre-programmed, the higher the threshold voltage (V_{tp}) is. Thus, we got another set of curves of current, as shown in Fig. 5. It can be clearly found that at the point

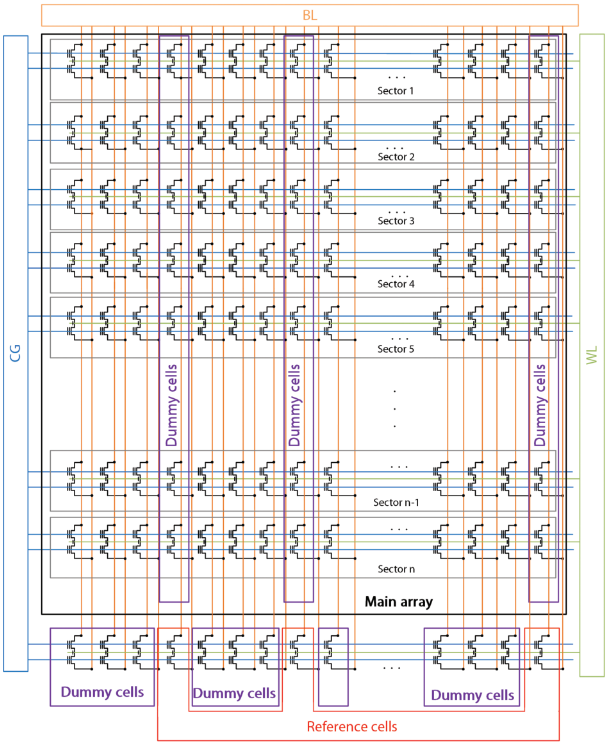

Fig. 6. (Color online) Memory cells in the NOR flash memory array.

where V_{tp} = 3.82 V at 25 °C, the threshold voltage and mobility decrease proportionally and compensate for each other. Thus, the read current exhibits insensitivity to the temperature, which is consistent with our earlier hypothesis. However, the current exhibits a little fluctuation because the threshold itself varies with the temperature.

4. Applications

Our experiment and analysis have suggested that the temperature characteristic can be used in setting the reference cell in flash memory. As shown in Fig. 6, a flash array consists of three types of cells with the same structure: memory cells to store information, reference cells to generate current reference, and dummy cells which share the same WLs of BLs with reference cells. Dummy cells cannot be read, and there are no contact formed in the process loop to connect them. They can create the same layout environment for cells on both sides as other cells, and thus avoiding mismatches due to stress and other reasons that may affect the accuracy of the results. The current generated by reference cells will be read separately, replicated and allocated proportionally by current mirrors to the memory cells as a comparison criterion and tolerance margin to differ “0” from “1”. Unlike normal memory cells, reference cells have fixed threshold voltage between the ‘0’ state and the ‘1’ state that are pre-programmed once they are produced, thereby exhibiting fixed drain–source current under certain bias condition. In addition, to make sure they can provide an absolutely standard current to assist reading operations, they don’t suffer from the P/E cycles test, which is mandatory for normal memory cells.

P/E cycles in endurance tests will bring some bitcells into a small current region due to oxide degradation and trapped charge. These bits (hereinafter called tail bits) have much smaller temperature coefficients and are less sensitive to the temperature variation than normal cells. A cell with a “10” state is taken as an example and depicted in Fig. 7. Although they are less sensitive compared with the normal cells, the tail bitcells exhibit a slightly left shift with the increasing temperature, induced by the aforementioned temperature coefficient variation of a “0” bitcell. However, the reference cells free of endurance tests don’t exhibit this shift. Thus, for these tail bits near the margin boundary, since their temperature coefficient is much smaller than that of reference cells, the subsequent temperature variation test will expel them from the correct reading identification region, which gives us the wrong reading data. In the manufacturing process, these tail bits can not maintain accuracy in the whole temperature interval and will fail the temperature test, eventually leading to yield loss.

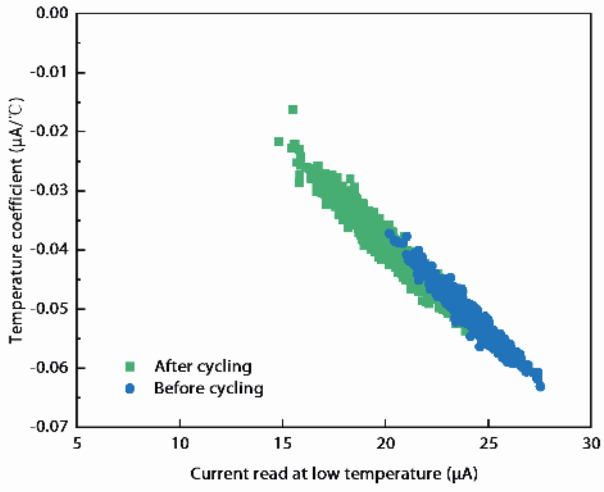

Fig. 7. (Color online) Sample current distribution and current shift due to endurance cycling.

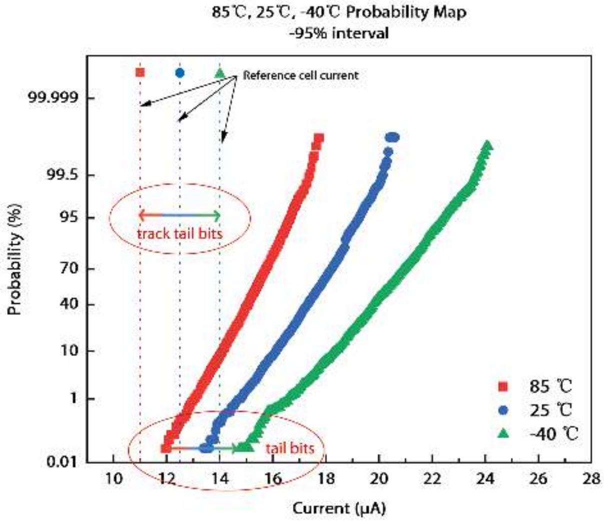

Fig. 8. (Color online) A tail-bit tracking reference cell current distribution.

This problem can be solved by pre-setting the reference cells with same temperature coefficients as those of low current cells, which is used to adjust the pre-programmed strength of each reference cell (as mentioned earlier). In this way, the reference level and margin generated by reference cells can track tail bits synchronously, as shown in Fig. 8, where the dotted straight lines represent the margin boundary of the reference at three temperature conditions. This approach effectively avoids misjudgments and significantly improves yield and endurance performance.

5. Conclusion

In this paper, a new cell structure with temperature-insensitive reading is introduced. It is designed based on the compensation of mobility and threshold voltage variation to the temperature when the read current is low. It can offer an adjustable temperature coefficient without the complex algorithm or extra temperature sensors, and is thus capable of acting as the reference cell of flash memory. By increasing programming time to increase or tune control gate voltage, the temperature coefficient of reference cells can be set to be equal to that of those temperature-insensitive tail bits with small currents, thus compensating for the endurance test shift and tracking tail bits in temperature tests to improve precision and yield.

References

[1] Han S T, Zhou Y, Roy V A. Towards the development of flexible non-volatile memories. Adv Mater, 2013, 25(38), 5425

[2] Chen F, Chen B, Lin H, et al. Temperature impacts on endurance and read disturbs in charge-trap 3D NAND flash memories. Micromachines, 2021, 12(10), 1152

[3] Resnati D, Goda A, Nicosia G, et al. Temperature effects in NAND flash memories: A comparison between 2-D and 3-D arrays. IEEE Electron Device Lett, 2017, 38(4), 461

[4] Zambelli C, Koebernik G, Ullmann R, et al. Modeling erratic bits temperature dependence for Monte Carlo simulation of flash arrays. IEEE Electron Device Lett, 2013, 34(3), 390

[5] Dilello A, Andryzcik S, Kelly B M, et al. Temperature compensation of floating-gate transistors in field-programmable analog arrays. 2017 IEEE International Symposium on Circuits and Systems (ISCAS), 2017, 1

[6] Shin H, Oh M, Choi J, et al. A 28nm embedded flash memory with 100MHz read operation and 7.42Mb/mm^2 at 0.85V featuring for automotive application. 2021 Symposium on VLSI Circuits, 2021, 1

[7] Dong Q, Wang Z, Lim J, et al. A 1Mb 28nm STT-MRAM with 2.8ns read access time at 1.2V VDD using single-cap offset-cancelled sense amplifier and in-situ self-write-termination. 2018 IEEE International Solid-State Circuits Conference (ISSCC), 2018, 480

[8] Guo X, Bayat F M, Prezioso M, et al. Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells. 2017 IEEE Custom Integrated Circuits Conference (CICC), 2017, 1

[9] Jin D H, Kwon J W, Seo M J, et al. A reference-free temperature-dependency-compensating readout scheme for phase-change memory using flash-ADC-configured sense amplifiers. IEEE J Solid-State Circuits, 2019, 54(6), 1812

[10] Fang L, Kong W, Gu J, et al. A novel symmetrical split-gate structure for 2-bit per cell flash memory. J Semicond, 2014, 35(7), 074008

[11] Lue H T, Hsu T H, Wu M T, et al. Studies of the reverse read method and second-bit effect of 2-bit/cell nitride-trapping device by quasi-two-dimensional model. IEEE Trans Electron Devices, 2006, 53(1), 119

[12] Tsividis Y, McAndrew C. Operation and modeling of the MOS transistor. 3rd ed. Oxford University Press, 2011

[13] Dwivedi A K, Tyagi S, Islam A. Threshold voltage extraction and its reliance on device parameters @ 16-nm process technology. Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT), 2015, 1

[14] Tao C, Vega R A, Alptekin E, et al. Understanding short channel mobility degradation by accurate external resistance decomposition and intrinsic mobility extraction. J Appl Phys, 2015, 117(6), 64507

Weiyan Zhang got his BS degree from Hunan Normal University in 2020. He is currently a Master student at ShanghaiTech University. His research interests include the structure, process and reliability of the embedded flash memory.

Tao Yu got his Master's Degree in Physics from Peking University in 2009. He Joined technology development department of Shanghai Huahong Grace Semiconductor Manufacturing Co. Ltd in 2010. His research focuses on the non-volatile memory, especially on the embedded flash.

Zhifeng Zhu received the BSc degree from University of Electronic Science and Technology of China in 2014 and the PhD degree from National University of Singapore in 2019. He joined ShanghaiTech University as an assistant professor in 2020. His research focuses on the theoretical and numerical study of spintronic devices for the application of MRAM and neuromorphic computing.

Binghan Li got his PhD from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences in 2004. After then he joined Shanghai Huahong Grace Semiconductor Manufacturing Co. Ltd and mainly engaged in the R & D of non-volatile memory. He has published more than ten papers, and applied for more than 50 patents.