Received 2 December 2021; accepted 3 January 2022. Date of publication 6 January 2022; date of current version 16 February 2022.
The review of this article was arranged by Editor S. Menzel.

*Digital Object Identifier 10.1109/JEDS.2022.3140949*

# A Novel Program Suspend Scheme for Improving the Reliability of 3D NAND Flash Memory

**ZHICHAO DU[1,2,3], ZHIPENG DONG[3], KAIKAI YOU[1,2,3], XINLEI JIA[1,2,3], YE TIAN[1,2,3], YU WANG[3], ZHAOCHUN YANG[3], XIANG FU[3], FEI LIU[1], QI WANG[1,2,3], LEI JIN[1,2,3], AND ZONGLIANG HUO[1,2,3]**

1 Institute of Microelectronics, The Chinese Academy of Sciences, Beijing 100029, China
2 University of Chinese Academy of Sciences, Beijing 100049, China
3 Yangtze Memory Technologies Company, Ltd., Wuhan 430205, China

CORRESPONDING AUTHOR: Z. HUO (e-mail: huozongliang@ime.ac.cn)

This work was supported in part by the National Science and Technology Major Project of China under Grant 21-02.

**ABSTRACT** Experimental results indicate that the conventional program suspend scheme in 3D NAND flash memory chip can generate unexpected additional read fail bits and reduce the reliability of 3D NAND flash memory. These extra read fail bits are observed when the program suspend command is issued during the program stage, and particularly, they become more obvious as the delay time between program suspend operation and other following operations exceeds tens of milliseconds. By analyzing the waveform of conventional program suspend scheme, it is suggested that the unexpected extra read fail bits are caused by the different occupancy of grain boundary traps (GBTs) in the polycrystalline silicon (poly-Si) channel during the idle time after the program suspend operation. Accordingly, a novel program suspend scheme is proposed by adding a "stabilizing" pulse immediately after the program stage. Silicon experimental data show that the proposed scheme can effectively limit the read fail bit count (FBC) to a normal range, thus improving the reliability of 3D NAND flash memory significantly.

**INDEX TERMS** 3D NAND flash memory, program suspend operation, fail bit count, grain boundary trap.

## I. INTRODUCTION

Over the past decade, the demand for NAND flash memory in mobile devices and solid state drives (SSDs) has grown rapidly [1], owing to its unprecedented advantages of low cost and large capacity. In addition, the development of word line (WL) stacking technology for three-dimensional (3D) NAND flash memory devices has further reduced the bit cost, and greatly increased the capacity [2], [3]. Nowadays, 3-bit/cell (TLC) and 4-bit/cell (QLC) memories have become the mainstream for 3D NAND flash memory [4], [5]. For TLC and QLC NAND flash memories, a typical program operation may take thousands of microseconds to complete [6], while the average read operation time per page of TLC and QLC only requires tens to hundreds of microseconds [3], [6]. To reduce the read latency, program suspend scheme is developed in NAND flash chip, so that it can support the host to interrupt the ongoing program operation and then perform multiple read operations on other pages [7]. Subsequently, as these read operations are completed, the host sends a program resume command to memory chip for continuing the interrupted program operation. Previously reported works [8], [9] studied the timing and sequence of the program suspend and resume scheme in NAND flash memory chip, where methods to reduce the required response time between program suspend command and read command were proposed. Progressively, one recent work [10] assessed the role of program suspend scheme in 3D NAND flash memory in SSD systems level, and suggested that the program suspend operation must be characterized carefully prior to its usage, to achieve better performance. In fact, in order to ensure the reliability of NAND device, the program suspend scheme in NAND chip must be carefully optimized when program suspend command is issued. However, there exist only a few studies that investigate the impact of program suspend scheme on the reliability of 3D NAND flash memory.

Correspondingly, this work first studies the conventional program suspend scheme [9] in Section II. Next, two series

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

of experiments are designed to measure the read fail bit count (FBC) [10]–[12] when program suspend and resume operations execute, since read FBC is one of the key parameters to characterize the reliability of the memory. It is found that read FBC increases significantly if program suspend command is issued in the program stage when the delay time between the program suspend operation and other operations exceeds tens of milliseconds. Furthermore, based on the analysis of conventional scheme, a novel program suspend scheme is proposed by adding a "stabilizing" pulse after the program pulse when the program suspend command is issued in the program stage. It is validated that with the proposed scheme the read FBC can be reduced to normal range successfully.

## II. PROGRAM SUSPEND AND RESUME INTRODUCTION

Before introducing conventional program suspend scheme, an example for a block structure of 3D NAND flash is shown in Fig. 1(a), where M strings are shared by one bitline. WLs are shared by different strings and one string is selected by applying appropriate voltage bias on the top select gate transistors (TSGs) [13]. The bottom select gate transistors (BSGs) connect strings to the source line [14]. Fig. 1(b) shows an example of TLC cells' threshold voltage (Vth) distribution and the measured read FBC is obtained by averaging the FBC of lower page (LP), middle page (MP) and upper page (UP) [12]. As the 3D NAND flash memory uses incremental step pulse programming (ISPP) scheme to program cells' threshold voltage to target different levels [15], a program operation can be divided into program stage, verify stage and idle stage. Since the program suspend command is issued randomly, it may execute at any stage. Meanwhile, when a program suspend command is issued, read operations during the program suspend period can occur on the same WL of other strings.

The sequences of conventional program suspend scheme [6] are shown in Fig. 2. Generally, if the program suspend command is issued during the program stage, the program stage is not interrupted, and instead, it continues as expected until the voltage bias of each WL ramps down to a common voltage (illustrated in Fig. 2(a)). Hence, a certain amount of time ($T_{RRL}$) is required before read operations can begin. After multiple read operations are finished, the program resume command can be issued, where the NAND chip continues the program operation from the verify stage. On the other hand, if the program suspend command is issued during the verify stage (Fig. 2(b)), the controller in the NAND chip terminates the verify stage and discharges WLs' voltages to a common bias to reduce the response time. Correspondingly, the NAND chip continues the program operation from the verify stage, as the program resume command is issued. Fig. 2(c) presents a detailed waveform example, where the program suspend command is issued during the program stage. In fact, during the program suspend period, if no read operations occur in the same block, the WLs of the block float and stay in an idle state.

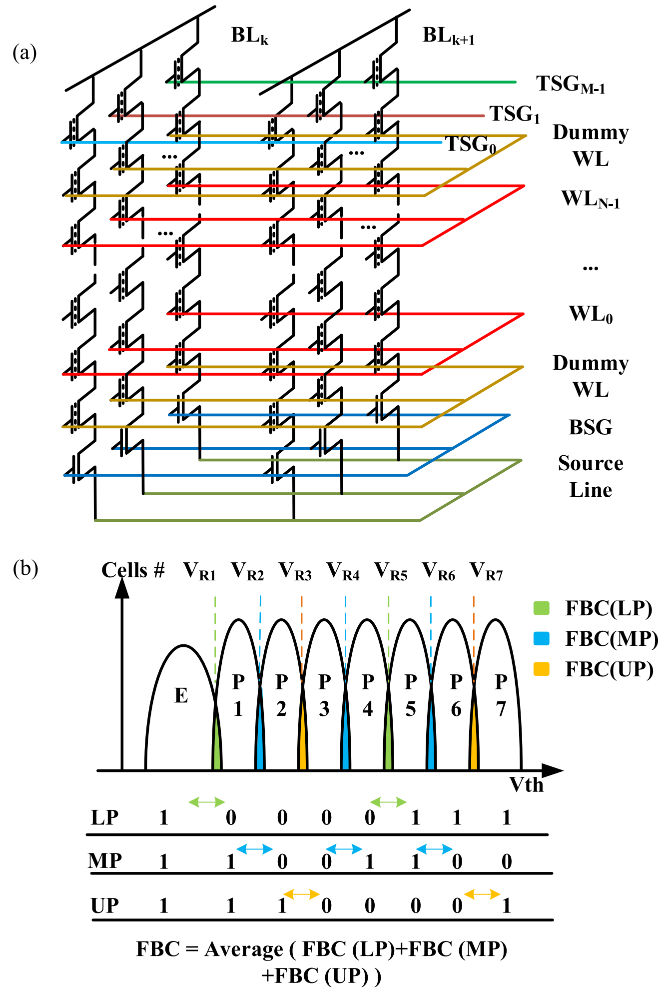


**FIGURE 1. (a) Block structure example of 3D NAND flash memory with M strings. In real silicon, the number of layers for TSG and dummy WL may be more. The program sequence in this work is from WL0 (locating at the bottom of the channel) to upper WLs. (b) Example of cells Vth distribution of a TLC NAND, the measured FBC is obtained by averaging the FBC of each page.**

As mentioned above, the program suspend command can be issued at any instant, and subsequent operations may arrive after a long idle time. In addition, studies [12], [16]–[18] have shown that read FBC increases when read operations are issued after a long idle time. Therefore, the reliability of devices undergoing the program suspend operation is concerned. Considering this, series of experiments for program suspend and resume operations are conducted on a raw 64-layered vertical 3D charge-trap TLC NAND flash test chip [19] by Teradyne$^{TM}$ Magnum V systems at room temperature. In this test chip, one page contains 16K bytes of data, and there exists no error correction function.

## III. EXPERIMENTS AND DISCUSSION

### A. EXPERIMENT OF READ DURING PROGRAM SUSPEND OPERATION

To test the read reliability during the program suspend operation, the following experiments are designed to execute

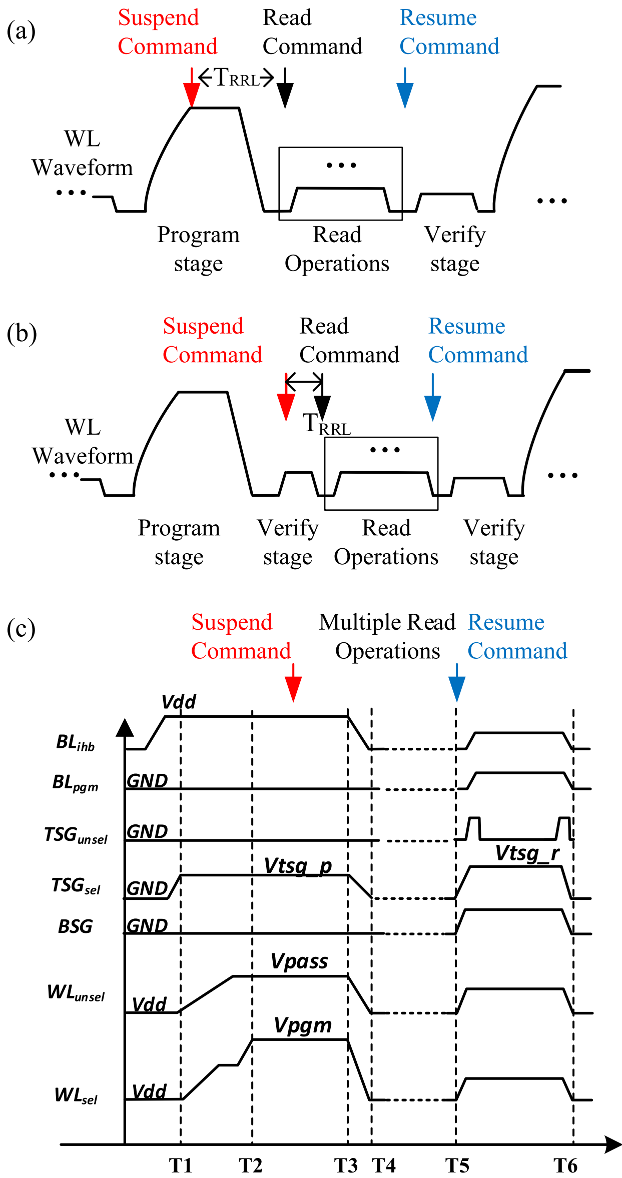


**FIGURE 2. (a) Sequence example of the program suspend scheme, when the program suspend command is issued during the program stage. (b) Sequence example of the program suspend scheme, when the program suspend command is issued during the verify stage. (c) WLs' waveform for conventional program suspend scheme, when suspend command is issued during the program stage.**

program suspend and read commands at different time instants. The experimental flow is displayed in Fig. 3(a). Initially, the selected block is erased and the pages all before the selected WL (WLn String1) address are sequentially programmed with random input data (The program sequence is from WL0 String0 to WL0 StringM-1, then another WL from String0 to StringM-1 until WLn String0). Next, program command is issued on WLn String1 with random input data. Following the program command, as time Tdly1 is elapsed, a program suspend command is issued. With another delay time of Tdly2 after the program suspend command, read commands are issued on WLn String0 to see if the read result of another string is influenced. At this stage, the average FBC of three pages is obtained. This erase-program-suspend-read sub-test is repeated many times with various Tdly1 and Tdly2 settings. Different Tdly1 times have been selected so that program suspend command can come in either during the program stage or the verify stage, while the Tdly2 ranges from 100$\mu$s to 1000ms. In this way, the dependence of the read FBC on the Tdly1 and Tdly2 is obtained. In order to present the results more effectively, the ratio between the read FBC of this experiment and the average read FBC without program suspend operation (FBC_ref) is utilized.

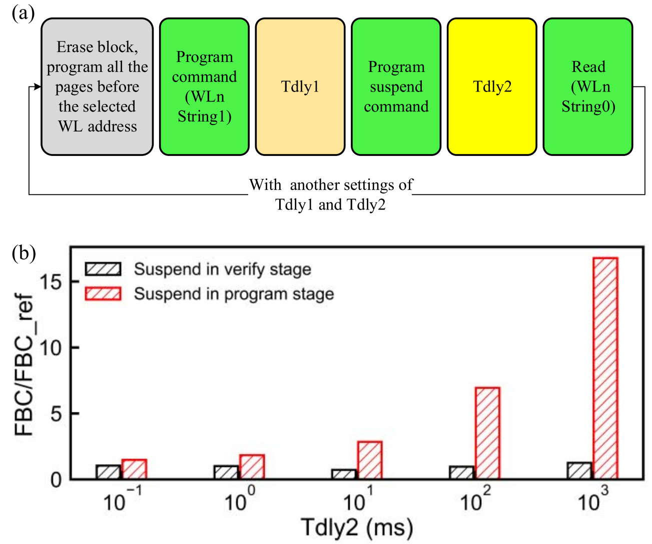


**FIGURE 3. (a) Example test flow of experiment A, where read operations are issued during the program suspend operation on another string. (b) The read FBC of the experiment A as a function of Tdly2, when program suspend command is issued in different stages.**

As shown in Fig. 3(b), with an increase in Tdly2 (the time between program suspend command and read command), an obvious increase in read FBC of same WL String0 is observed, if the program suspend command is issued during the program stage. The FBC increases non-linearly with the Tdly2, and it becomes noticeable as Tdly2 exceeds 10ms. Moreover, when Tdly2 increases to 1000ms, the read FBC reaches 17x of the reference read FBC (FBC_ref). In addition, it has been observed that if we read WLn String0 for second, third or more times, the read FBC of the subsequent read operations will gradually decrease from the first read operation. On the contrary, when the program suspend command is issued during the verify stage, variation in the read FBC with increasing Tdly2 is insignificant.

### *B. EXPERIMENT OF READ AFTER PROGRAM SUSPEND AND RESUME OPERATIONS*

In addition to above mentioned experiment, the impact of different delay times on the read FBC of the pages that are suspended and then fully resumed is also investigated. The relevant experimental flow is presented in Fig. 4(a).

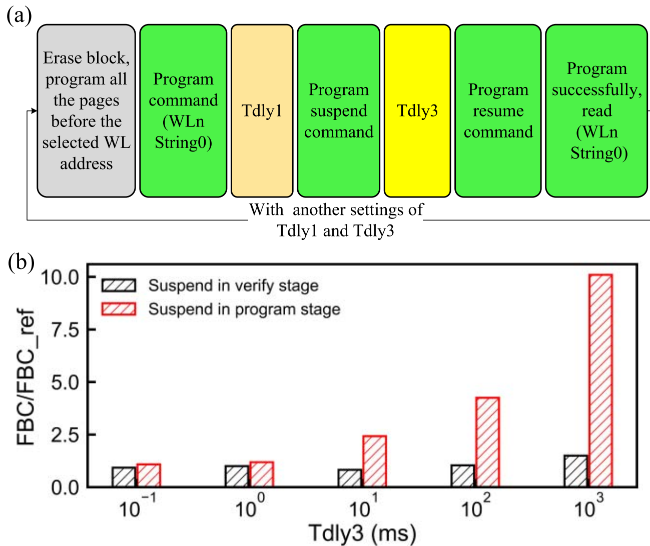


**FIGURE 4. (a) Example test flow of the experiment B, where read commands of same pages are issued after the program suspend and resume operations are completed. (b) The read FBC of experiment B of resumed pages as a function of Tdly3, when program suspend command is issued in different stages.**

First, the selected block is erased and the pages from WL0 to WLn-1 of all strings are sequentially programmed with random input data. Then, the program command is issued on WLn String0 with random input data. After a delay of Tdly1 from the program command, a program suspend command is issued. Next, with another delay time of Tdly3 after the program suspend command, a program resume command is issued to complete the whole program operation. Following the completion of program operation, we read these resumed pages and record the average read FBC. This erase-program-suspend-resume-read sub-test is repeated with different Tdly1 and Tdly3 settings. In our experiments, Tdly1 settings follow experiment A and the range for Tdly3 also spans from 100us to 1000ms. Similar to experiment A, the ratio of read FBC of this experiment to FBC_ref is used as an indicator to characterize the results.

Fig. 4(b) shows the extracted averaged read FBC ratio of this experiment. Again, abnormal read FBC appears when the program suspend command is issued during the program stage. The read FBC becomes noticeable as the Tdly3 exceeds 10ms, and when Tdly3 increases to 1000ms, the read FBC reaches to about 10x of the reference read FBC. However, when the program suspend command is issued during the verify stage, the average read FBC is comparable to the reference read FBC regardless of the increase in delay time Tdly3, i.e., consistent with experiment A.

### *C. MECHANISM ANALYSIS AND PROPOSED PROGRAM SUSPEND SCHEME*

As discussed in Section II, when program suspend command is issued during the program stage, the program pulse continues, and the voltages of WLs are discharged to a common voltage from program pulse. WL then remains in a floating idle state. If program suspend command is issued in the verify stage, though the verify stage time is reduced, the recovery of WL voltage is same as normal verify stage, and then it stays in an idle state. In addition, a normal program operation is always completed with a verify stage before WL shifts into an idle state.

Previously reported works [12], [16]–[18], the mechanism behind the temporary high read FBC of the early read operations of 3D NAND flash memory after hours long idle time has been investigated and elucidated. The main reason is the occupancy of grain boundary traps (GBTs) in poly-Si influences the sensing current. When a cell is programmed, the high gate voltage (Vpgm) makes most of the GBTs be filled. Then those trapped electrons at different trap energy levels in GBTs discharge partially and gradually. After a long idle time, the GBTs occupancy is reduced and the effective cell Vth is lowered, resulting in higher read FBC during the early read operations. More read operations following can help to charge the detrapped GBTs again, but it requires time. Particularly, it is suggested that a higher potential difference between the WL and the channel during the idle time can help to suppress the GBTs detrapping substantially [12]. In addition, in work [20], it has been reported that the channel potential after the verify stage is more negative, which means when the program suspend command is issued during the program stage, the GBTs detrapping influence should be more than during the verify stage, in line with our experimental results.

With above experimental results and theoretical analysis, a novel program suspend scheme is proposed by adding a “stabilizing” pulse to mimic the verify pulse after the program stage, if the program suspend command is issued during the program stage, as elaborated in Fig. 5 (a). All TSGs will be applied with a Vtsg_r voltage, all BSGs and WLs will be applied with a Vpass_rd bias to turn on all cells and all BLs will be biased to ground to save power. After a short period of time, all voltage biases will drop to a common voltage (e.g., Vdd).

Accordingly, a TCAD simulation of the proposed scheme is performed to check the channel potential of an inhibited cell that has already been in target state, as given in Fig. 5(b) and (c). A multi-layered 3D NAND devices are simulated as an example. In the simulation, the voltage parameters are set as follows: the Vpgm voltage is 20V, the Vpass voltage is 9V, Vtsg_p voltage is 3V, Vdd voltage is 2.4V, Vtsg_r and Vpass_rd voltages are both 6V. The position of selected WL is marked in gray. At time instant T2, the potential of the selected WL is boosted according to “Natural Local Self-Boosting Effect” in 3D NAND flash memory [21]. The potential drops a little between T2 to T3 due to the leakage. From instant T3 to instant T4, owing to a capacitive coupling between WLs and channel, the channel potential drops dramatically as the voltages of WLs ramp down, but at instant T4, the potential is still larger than zero

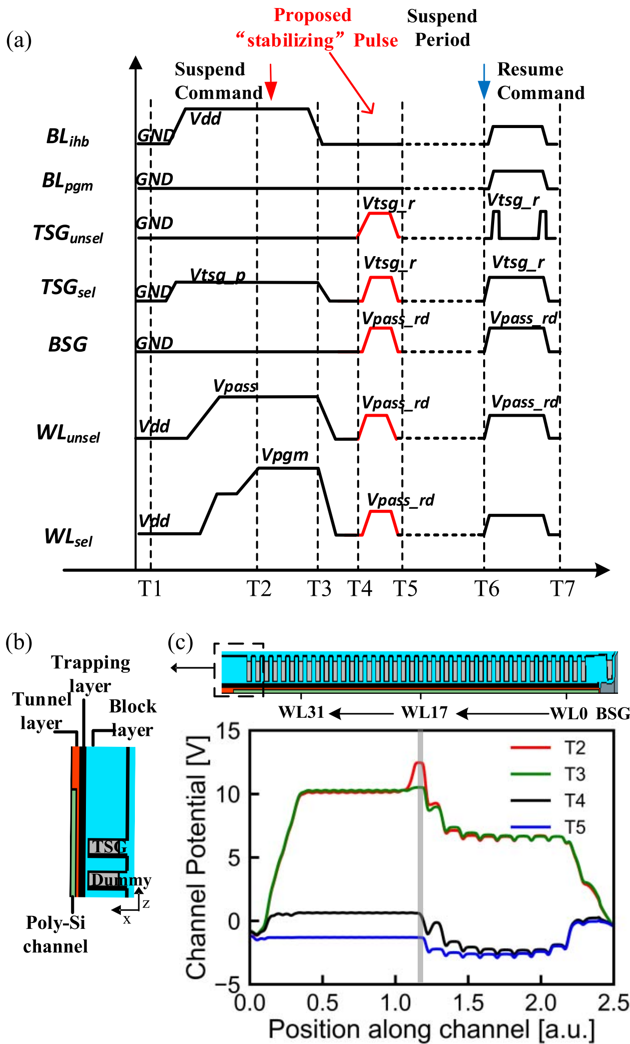


**FIGURE 5.** **(a) Proposed suspend scheme with a "stabilizing" pulse when suspend command is issued in the program stage. (b) Schematic of NAND structure used in TCAD simulation, the O/N/O stack consists of tunnel layer, trapping layer and block layer, respectively. (c) the simulated channel potential of a multi-layered 3D NAND devices of different time instants of the proposed scheme. The dashed line is the selected WL position. From WL18 to WL31, cells are in erased state as they have not been programmed yet. From WL0 to WL16, cells' Vth pattern is set to "P5/E/P5/E/P5/E/P5/E/P5/E/P5/E/P5/E/P5/E/P5" as an example. The selected WL during the simulation is in inhibited situation.**

at selected WL. However, at instant T5, after the proposed "stabilizing" pulse, the channel potential at selected WL drops to a negative value. Therefore, the potential difference between the gate and the channel after "stabilizing" pulse is more than that without "stabilizing" pulse, which complements our expectation. Since other strings' boosting scheme is similar to the inhibited channel, the channel potential of the same WL of other strings will be similar. Therefore, in the proposed scheme, lower read FBC is expected in both experiments mentioned in Section III when program suspend command is issued in the program stage.

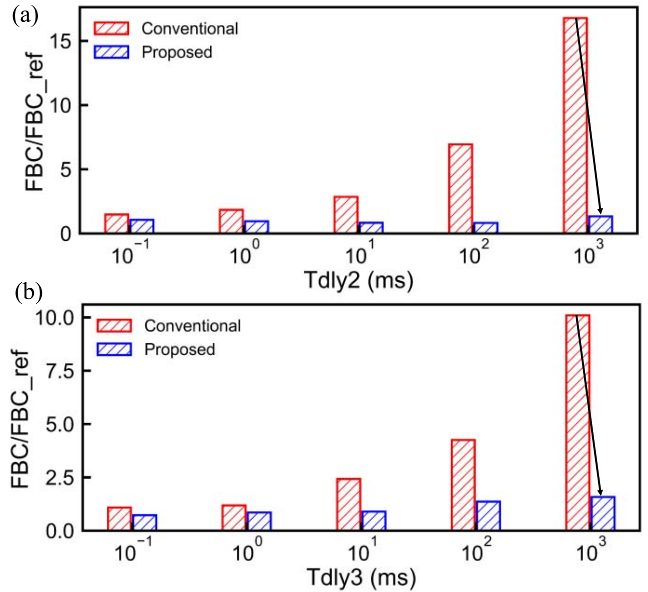


**FIGURE 6.** **(a) The average read FBC of experiment A, with or without "stabilizing" pulse when program suspend command is issued in the program stage. (b) The average read FBC of experiment B, with or without "stabilizing" pulse when program suspend command is issued in the program stage.**

### D. EXPERIMENTS RESULT WITH PROPOSED PROGRAM SUSPEND SCHEME

With this proposed scheme, the same experiments described in Section III are repeated on the same 3D NAND flash test chip. Fig. 6(a) and (b) presents the read FBC ratio with and without "stabilizing" pulse when program suspend command is issued in the program stage. The results from both experiments indicate that even when the delay times (Tdly2 and Tdly3) increase to 1000ms, with "stabilizing" pulse, the ratio FBC/FBC_ref reduces significantly from above 10 to around 1.3, which is small enough to be handled by ECC or other soft-correction mechanism.

Fundamentally, the "stabilizing" pulse corresponds to a short verify-alike pulse, it can be easily implemented by reusing the existing logic circuits in 3D NNAD flash chip. Based on silicon data, this "stabilizing" pulse requires ~10 microseconds. Certainly, this will increase the average read time (~3.3µs/page), however, this impact can be reduced further if more read operations are performed during the program suspend period. In addition, an average current with ~50mA and ~10µs width is measured by Keysight$^{TM}$ current probe tester during the "stabilizing" pulse. Considering the total program operation time (e.g., ~1.8ms), the increase of the average program current is ~0.27mA, which is also negligible. In the future work, the "stabilizing" pulse can be carefully merged with the program stage waveform to reduce the time and power consumption.

## IV. CONCLUSION

In this work, the reliability issue of conventional program suspend scheme in NAND flash memory chip is investigated.

It is found that there exist unexpected extra read fail bits when the program suspend command is issued during the program stage, and the longer the delay time between the program suspend command and the following commands, the more the read FBC. It is mainly because the GBTs detrapping is more in the idle time after the program suspend command is issued during the program stage. Accordingly, based on waveform analysis and TCAD simulation, a novel scheme is proposed by adding a "stabilizing" pulse immediately after the program stage for the program suspend operation. With proposed scheme, experiments have validated the reduction in read FBC to a normal range, and improvement in the reliability of 3D NAND flash memory. In addition, the proposal validated in this work can also be applied to QLC or other NAND flash memories.

## REFERENCES


[1] L. Zuolo, C. Zambelli, R. Micheloni, and P. Olivo, "Solid-state drives: Memory driven design methodologies for optimal performance," *Proc. IEEE*, vol. 105, no. 9, pp. 1589–1608, Sep. 2017, doi: 10.1109/JPROC.2017.2733621.

[2] D. Kang *et al.*, "7.1 256Gb 3b/cell V-NAND flash memory with 48 stacked WL layers," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2016, pp. 130–131, doi: 10.1109/ISSCC.2016.7417941.

[3] H. Maejima *et al.*, "A 512Gb 3b/cell 3D flash memory on a 96-word-line-layer technology," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2018, pp. 336–338, doi: 10.1109/ISSCC.2018.8310321.

[4] S. Lee *et al.*, "A 1Tb 4b/cell 64-stacked-WL 3D NAND flash memory with 12MB/s program throughput," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2018, pp. 340–342, doi: 10.1109/ISSCC.2018.8310323.

[5] N. Shibata *et al.*, "13.1 A 1.33Tb 4-bit/Cell 3D-flash memory on a 96-word-line-layer technology," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2019, pp. 210–212, doi: 10.1109/ISSCC.2019.8662443.

[6] D. Kim *et al.*, "13.1 A 1Tb 4b/cell NAND flash memory with tPROG=2ms, tR=110$\mu$s and 1.2Gb/s high-speed IO rate," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2020, pp. 218–220, doi: 10.1109/ISSCC19947.2020.9063053.

[7] G. Wu and X. He, "Reducing SSD read latency via NAND flash program and erase suspension," in *Proc. USENIX Conf. File Storage Technol.*, San Jose, CA, USA, Feb. 2012, pp. 1–10.

[8] N. Shibata *et al.*, "A 19nm 112.8mm$^2$ 64Gb multi-level flash memory with 400Mb/s/pin 1.8V toggle mode interface," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, 2012, pp. 422–424, doi: 10.1109/ISSCC.2012.6177073.

[9] T. Kouchi *et al.*, "13.5 A 128Gb 1b/cell 96-word-line-layer 3D flash memory to improve random read latency with tPROG=75$\mu$s and tR=4$\mu$s," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2020, pp. 226–228, doi: 10.1109/ISSCC19947.2020.9063154.

[10] C. Zambelli, L. Zuolo, A. Aldarese, S. Scommegna, R. Micheloni, and P. Olivo, "Assessing the role of program suspend operation in 3D NAND flash based solid state drives," *Electronics*, vol. 10, p. 1394, Jun. 2021. [Online]. Available: https://doi.org/10.3390/electronics10121394

[11] M. Abe, T. Nakamura, and K. Takeuchi, "Pre-shipment data-retention/read-disturb lifetime prediction & aftermarket cell error detection & correction by neural network for 3D-TLC NAND flash memory," in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, 2019, pp. T216–T217, doi: 10.23919/VLSIT.2019.8776480.

[12] S. Xia *et al.*, "Analysis and optimization of temporary read errors in 3D NAND flash memories," *IEEE Electron Device Lett.*, vol. 42, no. 6, pp. 820–823, Jun. 2021, doi: 10.1109/LED.2021.3073604.

[13] Y. Zhang *et al.*, "Leakage characterization of top select transistor for program disturbance optimization in 3D NAND flash," *Solid-State Electron.*, vol. 141, pp. 18–22, Mar. 2018, doi: 10.1016/J.SSE.2017.11.006.

[14] Y. Fukuzumi *et al.*, "Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory," in *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 2007, pp. 449–452, doi: 10.1109/IEDM.2007.4418970.

[15] K.-D. Suh *et al.*, "A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995, doi: 10.1109/4.475701.

[16] W.-J. Tsai *et al.*, "Polycrystalline-silicon channel trap induced transient read instability in a 3D NAND flash cell string," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2016, p. 11, doi: 10.1109/IEDM.2016.7838395.

[17] W.-L. Lin *et al.*, "Grain boundary trap-induced current transient in a 3-D NAND flash cell string," *IEEE Trans. Electron Devices*, vol. 66, no. 4, pp. 1734–1740, Apr. 2019, doi: 10.1109/TED.2019.2900736.

[18] C. Zambelli, R. Micheloni, S. Scommegna, and P. Olivo, "First evidence of temporary read errors in TLC 3D-NAND flash memories exiting from an idle state," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 99–104, 2020, doi: 10.1109/JEDS.2020.2965648.

[19] Y. Simon, "Unleashing 3D NAND's potential with an innovative architecture," Presented at the Flash Memory Summit, Aug. 2018. [Online]. Available: http://www.ymtc.com/index.php?s=/cms/cate/69.html

[20] Y. Kim and M. Kang, "Down-coupling phenomenon of floating channel in 3D NAND flash memory," *IEEE Electron Device Lett.*, vol. 37, no. 12, pp. 1566–1569, Dec. 2016, doi: 10.1109/led.2016.2619903.

[21] M. Kang and Y. Kim, "Natural local self-boosting effect in 3D NAND flash memory," *IEEE Electron Device Lett.*, vol. 38, no. 9, pp. 1236–1239, Sep. 2017, doi: 10.1109/LED.2017.2736541.