

FPGA-based reliability testing and analysis for 3D NAND flash memory

Debao Wei*, Zhelong Piao, Hua Feng, Liyan Qiao, Xiyuan Peng

School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, PR China

ABSTRACT

In this article, we described a FPGA-based NAND flash memory hardware and software collaborative design experimental platform. Our novel experimental platform implements timing control and error feature extraction of NAND flash memory, equipped with a computer software based on LabVIEW. The test results show that our test platform can process the experimental data with high speed, stability and accuracy so that users can effectively evaluate the storage system indicators based on NAND flash memory. In addition, we conducted research on a newly discovered 3D NAND flash memory reliability threat which exists in reading an idle block.

1. Introduction

Due to the inherent characteristics of 2D NAND flash, the miniaturization process has enhanced some reliability threats such as cell-to-cell interference and the hot-carrier disturbance, which eventually leads to a paradigm shift in the NAND flash technology from the planar cell to the vertical NAND cell [1]. The reliability of 3D NAND flash memory depends on many factors, most of which are caused by the inter-layer differences brought by the stacking process [2,3], causing some new reliability challenges and complicating the work of ECC in maintaining a high level of reliability. Such as data resident capability and cross-temperature immunity are all affected indicators [4].

Many efforts in the reliability community are devoted to investigate the reliability-loss of this storage medium. It's essential to add some flash signal processing technologies (such as machine learning) to the flash controller, called as NAND flash management algorithms, realizing the reliability improvement of system based on NAND flash memory. At present, NAND flash memory management algorithms are very extensive, but there is no unified method to verify the performance of these algorithms in the world. Intel developed a variety of NAND flash controller IP Cores based on ALTERA FPGA products, such as Cyclone, MAX and Stratix Series [5,6]. Seoul National University and Massachusetts Institute of Technology have development an experiment platform called BlueSSD. Chuo university has developed a 3D flash memory reliability test platform compatible with multiple packaging forms and studied the extension of data-retention lifetime through Adaptive Artificial Neural Network coupled (ANN) LDPC ECC (ANN-LDPC ECC) [7]. Huazhong University of Science and Technology designed a flash memory test platform based on Xilinx Spartan3E FPGA and conducted research on program-disturb error [8]. Many validation methods based on homemade hardware in the world are ideal flash reliability test platforms, because the actual hardware verification

algorithms and performance evaluation have a high degree of credibility. However, these methods require the R&D teams have a high ability of hardware development and the designer not only must grasp more EDA design tools, but also have the ability to debug the hardware. In addition, these designs are generally closed, which means that users cannot obtain sufficient reliability indicators according to their experimental needs. These factors limit the widespread use of these methods.

In order that users can flexibly change the configuration of experiment apparatus and obtain lower-level reliability indicators, this research designs an experimental platform that supports multiple packages and various types of NAND flash memory chips, conducts reliability index collection, and studies a newly discovered reliability problem of 3D NAND flash memory.

2. Experimental platform design

2.1. NAND flash basic

The memory cell of NAND flash device is floating gate (FG) transistor [9], in which a polysilicon island (or floating gate) is sandwiched between an inter-poly oxide layer and a tunnel oxide layer, both of which have insulating effect. Since the charge is not easy to leak from the isolated floating gate, a nonvolatile memory element is realized. On top of a flash cell is the control gate (CG). The voltage applied on the CG controls the conductivity of the conductive channel between the source and the drain electrodes [10]. The minimum voltage required is called as the threshold voltage (V_T) of the flash memory cell. The threshold voltage can be changed by injecting charge into FG. The injected charge Q (charge programmed on the FG, variable with negative value) and the threshold voltage V_T can be formulated as [11]:

* Corresponding author.

E-mail address: weidebao@hit.edu.cn (D. Wei).

<https://doi.org/10.1016/j.microrel.2020.113846>

Received 26 May 2020; Received in revised form 16 July 2020; Accepted 19 July 2020

0026-2714/ © 2020 Elsevier Ltd. All rights reserved.

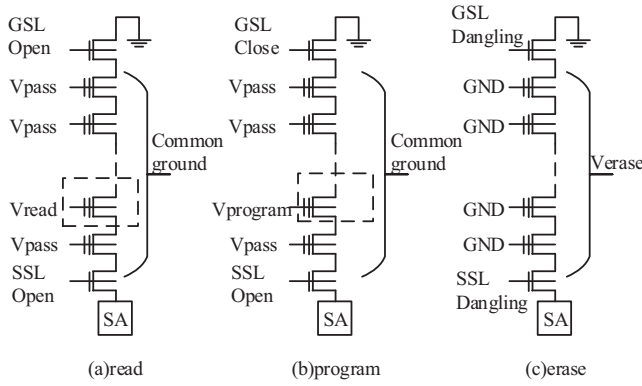


Fig. 1. NAND flash access control: (a) read (b) program (c) erase.

$$V_T = V_{T0} - \frac{Q}{C_{pp}} \quad (1)$$

In formula 1, V_{T0} is the so-called neutral threshold voltage, which corresponds to zero charge on the floating gate and C_{pp} is the dielectric capacitance of polysilicon. During the program and erase operation of the cell, the injection and ejection of electrons through the tunnel oxide is achieved by the Fowler-Nordheim (FN) tunneling effect [12]. During the read operation, the internal circuit of the flash memory determines the amount of programmed charge by the threshold voltage.

Here is a brief introduction to the flash memory internal circuit that implement read, program, and erase operations. As shown in Fig. 1(a), when a cell is read on a bit line, other cells are placed in the on state, that is, a conduction voltage (V_{pass}) is applied to the cell control gate, and the cell after the turn-on voltage can be regarded as a wire, and the conduction current generated by the read source and drain of the cell flows to the sense amplifier. By reading the conduction current, the data stored in the cell can be obtained.

As shown in Fig. 1(b), the programming operation is implemented by applying a programming voltage to the control gate to inject charge into the flash memory cells, and the cells that are not in the programming state are in the on state. The programming voltage is an incremental step voltage pulse (Incremental Step-Pulse Programming, ISPP for short), and the programming voltage is increased step by step until the threshold voltage of the flash memory cell reaches a predetermined value [13,14].

As shown in Fig. 1(c), the erase operation of the flash memory chip is implemented by applying an erase voltage to the base of the flash memory cell, and the charges in the erased flash memory cell will be removed from the floating gate. Since the bases of all flash memory cells in a block are connected together, the erasing of the flash memory chip takes the block as the basic unit.

2.2. Test platform design goal

Our design goal is that the test platform can accurately and quickly obtain the error characteristics of the flash memory, so as to study the reliability of different stages in the lifetime cycle of the flash memory chip and verify various NAND flash management algorithms effect to achieve the healthy management of flash memory chips.

2.3. Test platform architecture

To provide a more efficient and flexible hardware/software design environment to evaluate the performance of NAND flash management algorithms, by adopting a modular idea, we designed the overall architecture as shown in Fig. 2. The test platform consists of a 3D NAND flash experiment hardware platform and a software based on LabVIEW. The hardware platform is composed of four modules: FPGA controller,

3D flash memory chips, 100 M Ethernet chip and E²PROM. FPGA controller consists of flash controller, E²PROM controller and 100 M Ethernet controller. Flash controller compatibles with the ONFI 4.0 interface of NAND flash. It can implement all basic operations (e.g. Read, program, erase) on NAND flash and obtain reliability-related indicators. E²PROM is used to record a small amount of experiment-related information (such as temporarily written or read data). 100 M Ethernet is used for communication between the 3D flash experiment platform and the computer. Software is built based on LabVIEW, used for communication, processing and storage of experimental data.

2.4. FPGA controller design

As the core module of 3D NAND flash experiment platform, the FPGA controller consists of flash controller, E²PROM controller and 100 M Ethernet controller. It completes the analysis of the commands or data sent by the host software, operates the storage device, obtains the experimental results, and uploads them to the computer software. The following is a specific introduction to the FPGA controller.

The NAND flash memory controller implements the following basic operations on flash memory and through further orderly combination of these operations to obtain the reliability index of 3D NAND flash. Here are five specific function descriptions of the flash memory controller.

- 1) The initialization operation should be performed only after the flash memory is powered on. It is used to reset the NAND flash memory and select a timing mode. The flash memory adopts different time modes to match the speed of the external device. Our test platform makes the flash work in mode5.
- 2) The read-status function is used to judge the operation progress of write and erase operations. This function can not only judge whether the operation is over, but also whether the operation is successful. It is recommended to use this function instead of reading the R/B pin to obtain more accurate results when performing P/E cycles.
- 3) As described in Section 2.1, reading pages, programming pages, and erasing blocks implement data input, output, and delete operations on NAND flash memory chips.
- 4) Setting parameters is used to configure the specific register parameters of the NAND flash memory chip. For example, the read-retry mechanism is used to reduce the error bits.
- 5) Commonly used operations for reliability studies are Program/Erase operations and retrieval of error bits. Our test platform can perform a specified number of P/E operations on any flash memory block, with a desired delay between P/E operations. In order to comprehensively analyze the data stored in the 3D NAND flash, our test platform can upload the data read from the 3D NAND flash to the host software for processing. Our test platform can upload about 32,000 bytes of data from 3D NAND flash every second to the computer. Another common solution is to directly retrieve the number of error bits in the target area (page or block) through the FPGA controller and upload the results to the host, which can reduce the deviation of the flash memory reliability measurement due to the delay of large data transmission. It takes about 1 s to retrieve the error bits number of each block.

The E²PROM controller is used to realize serial data reading and writing, and is used to assist the experiment of 3D NAND flash memory.

The 100 M Ethernet chip of the 3D flash experiment platform is responsible for the client in the TCP/IP communication protocol. The 100 M Ethernet controller program consists of an initialization thread and a data transmission thread. The initialization thread is used to initialize the TCP/IP protocol stack, including setting the local IP address, destination IP address, port number and so on. After establishing the communication connection, the data transmission thread first starts the data receiving program in response to the instructions or data sent by

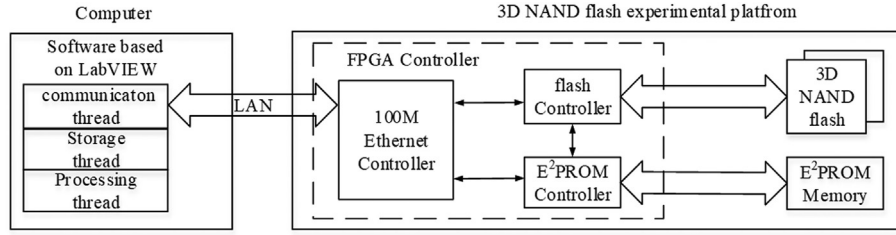


Fig. 2. System architecture of experiments platform for NAND flash.

the host software. After that, the data transmission thread starts the data sending process and uploads the data generated by the experiment to the host software.

3. Implementation methodology

In this section, we introduced experimental environment and the relevant implementation scheme.

3.1. Experimental environment

Fig. 3 shows a photograph of our 3D NAND flash experimental platform. We adopt the test platform to take the reliability index of 3D NAND flash. The FPGA device used in this hardware platform is EP4-CE15E22C8, which is an Altera FPGA. The crystal on the test platform makes FPGA work at 50 MHz. The selected NAND flash memory chip model is MT29F64G08CBCGB, which is a 3D MLC NAND flash from Micron Inc. The model used by E²PROM is 93LC56B for temporary recording of experimental information. All experiments were conducted at a constant temperature of 25 degrees Celsius. To make the charge leakage effect more obvious, all cells were programmed with the data pattern of “0 × 00”, due to the fact that biasing flash memory cells state to this pattern requires more electrons to be injected into the floating gate.

3.2. Experimental scheme

Our recent experiments on various types of 3D NAND flash and the latest international research present a new reliability threat that leads to read error [15] (called as Temporary Read Errors (TRE) in this reference paper). This threat exists in reading an idle 3D NAND flash

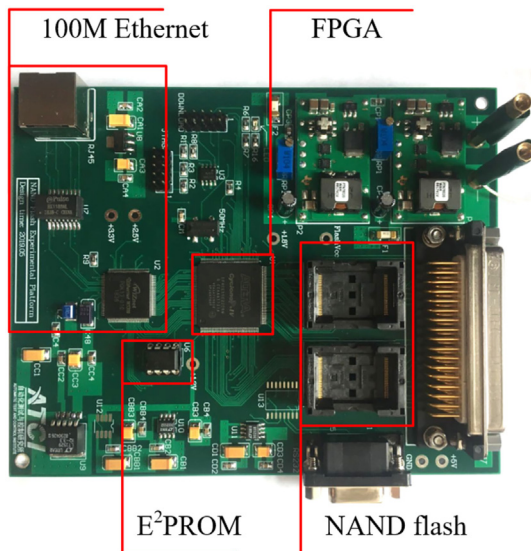


Fig. 3. Photograph of 3D NAND flash experimental platform.

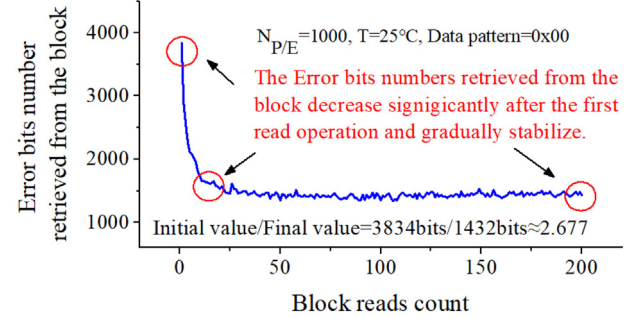


Fig. 4. Error bits numbers retrieved from 200 continuous block reads after 1000 P/E cycles.

memory. A temporary high error bits number retrieved from a block will be reported in the first read and the error bits number will gradually stabilize in the following reads. Fig. 4 shows a typical curve of this phenomenon.

Fig. 4 shows the error bits number retrieved from 200 continuous block reads after 1000 P/E cycles. For 2D NAND flash, several continuous block error bits numbers retrieval results are usually stable. This characteristic of 3D flash memory shows a big difference from 2D flash memory, which means that the error bits number retrieved from a 3D NAND flash block only through one read is inappropriate to be regarded as valid. To study the characteristics of this phenomenon and reduce the reliability loss caused by this phenomenon through proper operation, we conducted the following experiments.

One thing needs special emphasis. we found that the error bits number retrieved by performing a continuous block read operations immediately after performing P/E cycles operation on a block cannot exhibit the above phenomenon. A delay of a few minutes between P/E and read operations is usually appropriate. Fig. 5 shows the curve obtained after performing 500 P/E cycles on a block and waiting for 10 s to read continuously. It can be seen that the result obtained is very unstable. In the following experiments, we took the scheme of delaying 10 min after the P/E cycles and then performing the first read operation.

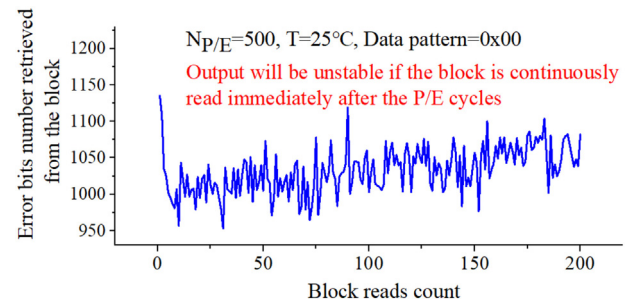


Fig. 5. Error bits numbers retrieved from 200 continuous block reads after 10 s of 500 P/E cycles.

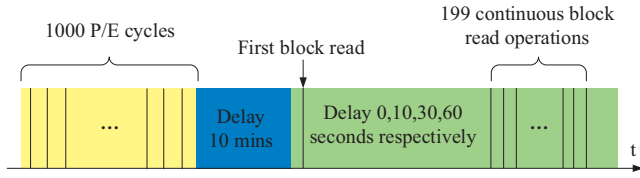


Fig. 6. Experimental scheme of adding different delays after the first read operation.

3.2.1. Research on the factors that lead to fewer error bits retrieved after the first read operation

First, we want to know the factor that leads to a decrease in the error bits numbers retrieved after the first read operation. Two factors are taken into consideration. One is enough block read operations and the other is that error bits number drops automatically after a period of the first read operation. In view of the fact that it takes about 1 s to retrieve the error bits number from a block, our experimental method was to wait for a period of time after the first read and then perform continuous block reads. The experimental steps are shown in Fig. 6. In the first step, we performed 1000 P/E cycles without interval for a block. In the second step, we separately delayed 0 s, 10 s, 30 s, and 60 s after the first read and then performed 199 continuous read operations to obtain the numbers of error bits retrieved from the block.

3.2.2. Changes in the number of error bits retrieved from each page in continuous block read operations

Second, we studied the difference between the error bits numbers retrieved from each page in the first block read and the final block read. The test platform supports page-by-page retrieval of the error bits number. The experimental process is concretely described as Fig. 7.

3.2.3. Reliability index measurement throughout the 3D NAND flash whole life cycle

In addition, we studied the error bits number retrieved from continuous block reads during the entire lifetime cycle of the 3D NAND flash. The experimental steps are shown in Fig. 8. We tested a total of 40 sets of data with P/E cycles varying from 100 to 10,000. After the P/E operation was completed, we delayed for 10 min, performed 200 continuous block read operations, retrieved and uploaded the number of error bits from the block after each read.

4. Experimental results

In this section, we analysed the reliability threat mentioned above that will cause 3D NAND flash read errors and reliability indicators related to endurance.

4.1. Research on the factors that lead to fewer error bits retrieved after the first read operation

Fig. 9 shows the error bits number retrieved from 200 block read operations on the NAND flash block with a delay of 0 s, 10 s, 30 s, and 60 s after the first read respectively. It can be seen that adding delay

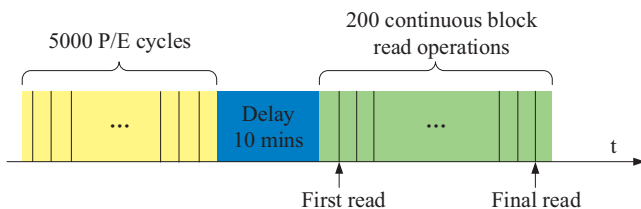


Fig. 7. Upload the number of error bits retrieved from each page at the beginning and end of the continuous read operations.

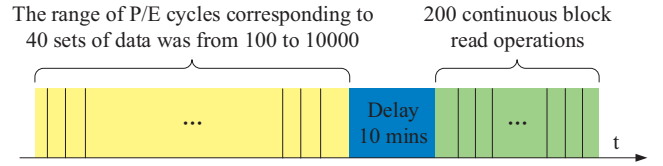


Fig. 8. Continuous block read test throughout the 3D NAND flash memory life cycle.

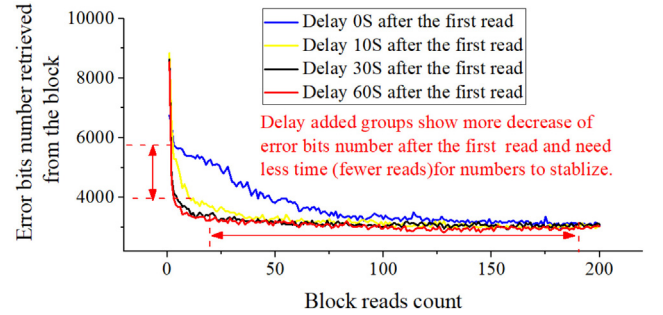


Fig. 9. Error bits number retrieved from 200 block reads in the experiment of adding different delay after the first read.

after the first read and then performing continuous read operations lead to more decrease of error bits numbers retrieved than the control group without delay. In addition, in the case of a delay after the first read, the error bits numbers retrieved from the block will automatically decrease and tend to stabilize. Based on this feature, users can add a delay after the first read and perform a second read, after which the data read can be considered as valid. Another solution is to perform enough block read operations and add an error bits numbers discrimination mechanism to the system. When the number of error bits retrieved fluctuates less than a given threshold, the system can stop the read operation and consider the output data to be valid.

4.2. Changes in the number of error bits retrieved from per page in continuous block read operations

Fig. 10 shows the error bits number retrieved from each page in the last time of 200 continuous read operations. The read operations were performed after the block had experienced 5000 P/E operations and another 10 min for retention. The 3D NAND flash memory selected for this test has 16 SLC pages at each end of the 512 pages in each block. It can be seen that there are few error bits retrieved from the SLC pages. This phenomenon is attributed to the wider threshold voltage window for the cells on the SLC pages, by which these pages have greater fault tolerance for charge leakage. In addition, most error bits are retrieved from the first MLC pages layer (page number from 16 to 47) in all layers

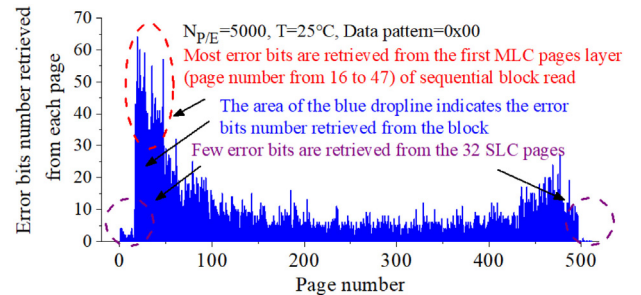


Fig. 10. Number of error bits retrieved from per page after 5000 P/E cycles for an ordinary block.

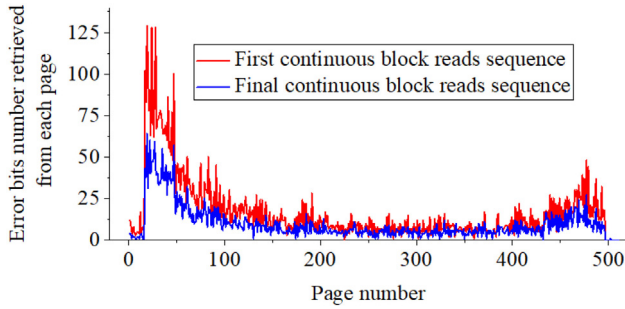


Fig. 11. Number of error bits retrieved from each page at the beginning and end of continuous block read operations.

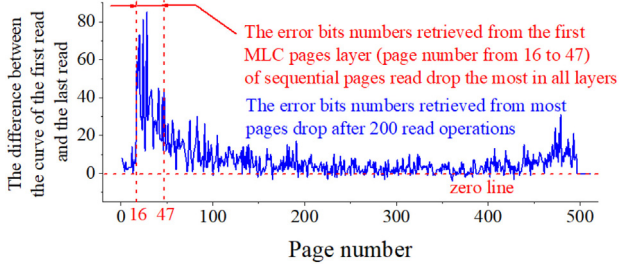


Fig. 12. Difference between the two curves in Fig. 11.

of sequential pages reads.

Fig. 11 shows the error bits number retrieved from each page in the first read and the last read of 200 continuous block reads. Fig. 12 shows the difference between error bits number retrieved from each page in the two sequences. It can be seen that the error bits number retrieved from most pages decrease after the continuous read operations. Only few pages report more error bits and the numbers are fewer than 10 bits. In addition, it can be seen that if each continuous block read operations start sequentially from the page address of 0, the number of error bits number retrieved from the first MLC pages layer (page number from 16 to 47) decrease the most. This is consistent with the conclusion about the study of TRE phenomenon [15].

4.3. Reliability index measurement throughout the 3D NAND flash whole life cycle

Fig. 13 shows the raw bit error rate (RBER) retrieved after P/E operations with a range from 100 to 10,000 P/E cycles in 200 continuous block reads. It can be seen that the error bits numbers of all curves first decrease and then gradually stabilize, which means that the above-mentioned read error exists throughout the lifetime cycle of the

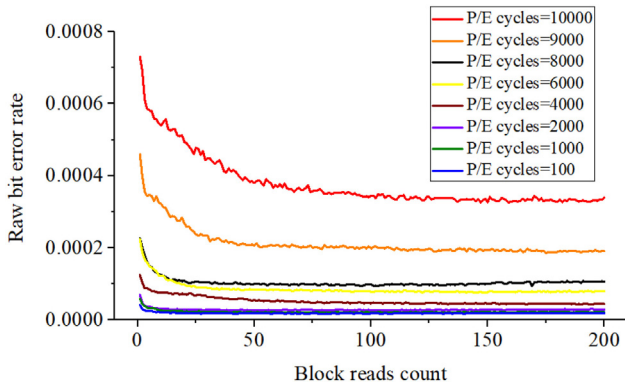


Fig. 13. RBER retrieved after P/E Operation with a range from 100 to 10,000 P/E cycles in 200 continuous block reads.

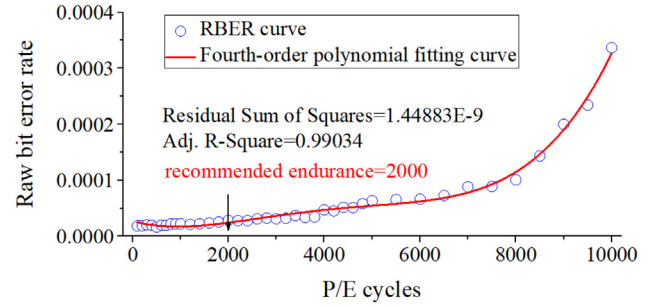


Fig. 14. RBER curve with the increasing P/E cycles.

3D NAND flash. In addition, it can be seen that the more P/E cycles the flash block experiences, the longer it takes for the number of error bits retrieved after the first read to stabilize. Further research is needed on the relationship between two variables. This helps to add an appropriate delay after the first read and reduce the read times before the data obtained is considered valid.

Through the above experiments, we obtained the changes of RBER retrieved from the block over the whole lifetime cycle (as shown in Fig. 14). It can be seen that the endurance of the selected 3D NAND flash is about 10,000 P/E cycles, which is 5 times the endurance mentioned in the datasheet. The variation range of RBER is from the level of 10^{-5} to 10^{-4} , and there is a significant improvement at the end of the flash memory lifetime. we took the scheme of fourth-order polynomial fitting commonly used in 2D NAND flash to model the RBER curve. The result is that Residual Sum of Squares result equals 1.44883E-9, and the R-Square equals 0.99034, which means the mathematical model fits the experimental data well.

5. Conclusion

In this research, we designed a test platform that supports testing multiple packages 3D NAND flash memory chips complying ONFI 4.0 interface. We used the test platform to conduct reliability tests on 3D MLC NAND flash and obtained some valuable observations. For example, after performing Continuous block read operations on an idle 3D NAND flash block and retrieving the number of error bits, we find that the numbers decrease significantly after the first read and gradually stabilize. In addition, we have studied the characteristics of this phenomenon and given countermeasures. The number of error bits retrieved from the block will automatically stabilize after a period of time of the first read, and the decrease in the number mainly comes from the first MLC pages layer of sequential pages read. In order to reduce the reliability loss caused by this phenomenon, users can delay a period of time after the first read and regard the data obtained from the second time at as valid, or perform enough read operations until the fluctuation range of the retrieved error bits number is less than the given value before consider the data read to be valid. In addition, we modelled the changes in RBER over the whole lifetime cycle of the selected 3D NAND flash block. The model fits well with the experimental data, and it can be seen that the endurance of the selected 3D NAND flash is about 5 times that recommended by the datasheet.

CRediT authorship contribution statement

Debao Wei: Conceptualization, Methodology, Writing - review & editing. **Zhelong Piao:** Software, Data curation, Writing - original draft. **Hua Feng:** Software, Investigation, Formal analysis. **Liyan Qiao:** Project administration, Writing - review & editing. **Xiyan Peng:** Supervision, Funding acquisition, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61802088, in part by the Fundamental Research Funds for the Central Universities under Grant HIT. NSRIF. 201814, in part by the Heilongjiang Postdoctoral Fund under Grant LBH-Z18084, and in part by the Guangxi Key Laboratory of Automatic Detecting Technology and Instruments under Grant YQ20205.

References

- [1] Y. Park, et al., Scaling and reliability of NAND flash devices, *Proc. IEEE IRPS*, Jun. 2014 pp. 2E.1.1–2E.1.4.
- [2] W.J. Tsai, W. Lin, C.C. Cheng, et al., Polycrystalline-silicon channel trap induced transient read instability in a 3D NAND flash cell string, *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2016, pp. 11.3.1–11.3.4, <https://doi.org/10.1109/IEDM.2016.7838395>.
- [3] Y. Kim, M. Kang, Down-coupling phenomenon of floating channel in 3D NAND flash memory[J], *IEEE Electron Device Letters* 37 (12) (2016) 1566–1569.
- [4] C. Zambelli, R. Micheloni, P. Olivo, et al., Reliability challenges in 3D NAND flash memories[C], *International Memory Workshop*, 2019.
- [5] <https://www.intel.com/content/www/us/en/programmable/solutions/partners/partner-profile/eureka-technology-inc-/ip/synchronous-onfi-nand-flash-controller-ep502-.html#ip-quality-metrics>.
- [6] <https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/an/an500.pdf>.
- [7] T. Nakamura, Y. Deguchi, K. Takeuchi, et al., 9.1x error acceptable adaptive artificial neural network coupled LDPC ECC for charge-trap and floating-gate 3D-NAND flash memories[C], *custom integrated circuits conference* (2018) 1–4.
- [8] M. Zhang, F. Wu, H. Huang, et al., FPGA-based failure mode testing and analysis for MLC NAND flash memory[C], *Design, Automation, and Test in Europe*, 2017, pp. 434–439.
- [9] P. Pavan, R. Bez, P. Olivo, et al., Flash memory cells-an overview, *Proc. IEEE* 85 (8) (1997) 1248–1271.
- [10] Y. Cai, Y. Luo, E.F. Haratsch, et al., Data retention in MLC NAND flash memory: Characterization, optimization, and recovery[C], *High-Performance Computer Architecture*, 2015, pp. 551–563.
- [11] J.E. Brewer, M. Gill, *Nonvolatile Memory Technologies with Emphasis on Flash: A Comprehensive Guide to Understanding and Using NVM Devices*, IEEE Press Series on Microelectronic Systems, WILEY Press, 2007.
- [12] R. H. Fowler and L. Nordheim, “Electron emission in intense electric fields”, *Proceedings of the Royal Society A* 119 (781): 173–181.
- [13] K.D. Suh, B.H. Suh, Y.H. Lim, et al., A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits, IEEE Journal of* 30 (11) (1995) 128–129.
- [14] J. Brewer, M. Gill, *Nonvolatile Memory Technologies with Emphasis on Flash: A Comprehensive Guide to Understanding and Using Flash Memory Devices*. Upper Saddle River, Wiley-IEEE Press, 2008.
- [15] C. Zambelli, R. Micheloni, S. Scommegna, et al., First evidence of temporary read errors in TLC 3D-NAND flash memories exiting from an idle state[J], *IEEE Journal of the Electron Devices Society* (2020) 99–104.