Characterizing the Reliability and Threshold Voltage Shifting of 3D Charge Trap NAND Flash

Weihua Liu[†], Fei Wu[†*], Meng Zhang[†], Yifei Wang[†], Zhonghai Lu[‡], Xiangfeng Lu[§], Changsheng Xie[†]
[†] Huazhong University of Science and Technology, China
[‡] KTH Royal Institute of Technology, Sweden
[§] Beijing Memblaze Technology Co., Ltd., China
[*]Corresponding author: {Fei Wu, wufei@hust.edu.cn}
{liuweihua, zgmeng, yifeiwang, cs_xie}@hust.edu.cn, zhonghai@kth.se, xiangfeng.lu@memblaze.com

***Abstract*—3D charge trap (CT) triple-level cell (TLC) NAND flash gradually becomes a mainstream storage component due to high storage capacity and performance, but introducing a concern about reliability. Fault tolerance and data management schemes are capable of improving reliability. Designing a more efficient solution, however, needs to understand the reliability characteristics of 3D CT TLC NAND flash. To facilitate such understanding, by exploiting a real-world testing platform, we investigate the reliability characteristics including the raw bit error rate (RBER) and the threshold voltage (Vth) shifting features after suffering from variable disturbances. We give analyses of why these characteristics exist in 3D CT TLC NAND flash. We hope these observations can guide the designers to propose high efficient solutions to the reliability problem.**

***Index Terms*—3D NAND Flash, Charge Trap, Reliability, Threshold Voltage Shifting**

I. Introduction

Recent years, 3D TLC NAND flash is gradually widely used in data center, smart devices, and personal computers due to higher storage capacity than planar NAND flash. However, reliability becomes a problem because of high raw bit error rate (RBER) induced by P/E cycles and retention periods [1]. To improve reliability, fault tolerance algorithms and data management schemes are proposed in flash-based storage systems, but developing more efficient solutions requires to clearly understand the reliability characteristics of 3D TLC NAND flash.

Many previous works analyzed the NAND flash based on physical characteristics and materials. Liebault et al. [2] investigated the trapping properties of a thin oxide layer deposited on a non-conductive substrate, to characterize the role of the interface in charge trap (CT) component. Kim et al. [3] evaluated the data retention characteristics from activation energy, temperature and threshold voltage (Vth) shifting, to predict the retention lifetime of CT NAND flash. Some works are based on the test. Mauroux et al. [4] made an earlier preliminary study on actual defects in NAND flash with the test. Cai et al. [5] researched the characteristics of MLC NAND flash, such as data retention and Vth shifting. Recently, they characterized the read disturb errors in MLC NAND flash through the Vth shifting [6]. Xiong et al. [7] had a more comprehensive characteristic analysis of 3D floating gate (FG) NAND flash. In conclusion, the previous works lack characterizing the reliability and Vth distribution and shifting (VDS) of 3D CT NAND flash.

In this paper, by exploiting a real-world hardware platform to conduct comprehensive experiments on the real 3D CT TLC chips. We investigate the reliability characteristics and give many observations, such as RBER and VDS features after suffering from various disturbances. We give analyses of causing the reliability characteristics.

To our knowledge, this is the first paper to characterize the reliability of 3D CT NAND flash with real VDS. The contributions of this paper are as follows:

- We make a comprehensive analysis of the reliability of 3DCT NAND flash through the test. We design the test experiment from a practical point of view, to research the RBER changes and the VDS with variations of disturbance.
- We have several new observations in the paper, such as an unexpected downgrade of RBER with the growth of P/E cycles, different Vth shifting speed. They can provide more methodologies to research the fault tolerance algorithms and error correcting schemes [8] to solve the reliability issues in 3D CT NAND flash. This work makes the reliability predictable to some extent.
- As the basic research of 3D CT NAND flash, we give a complete methodology from test and analysis to application. It is an integral part of designing flash-based systems.

The rest of the paper is organized as follow. Section II introduces the basic about 3D CT TLC NAND flash. Section III presents the experiment schemes. Many observations of the reliability and VDS characteristics are given in Section IV. Finally, the conclusions are made in Section V.

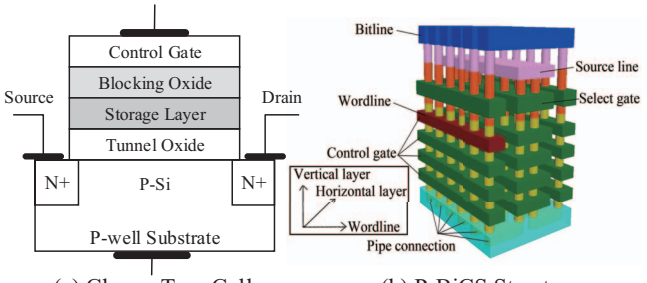
II. Background

NAND flash cell stores data by saving electronics. The structure diagram of CT NAND flash cell is shown in Fig. 1(a). The CT NAND flash cell uses a storage layer instead of FG, the storage layer is a silicon nitride film and far thinner than FG in FG NAND flash. Since the differences of materials and tunneling mechanism [9], the CT cell requires a lower programming voltage than FG, and it can reduce the stress and wear on the tunnel oxide. In addition, it consumes less energy

during program and erase operations. In particular, it is also faster to program than conventional FG flash. In summation, the CT cell can improve the endurance, reliability and reduce the latencies of NAND flash, compared with FG [10]. So the NAND flash manufacturers trend to use the CT structure in the era of 3D NAND flash [7].

Toshiba adopts the Pipe BiCS (P-BiCS) structure of 3D CT NAND flash, as illustrated in Fig. 1(b). A wordline (WL), which contains several cells, is adjacent to other WLs with both horizontal and vertical layers. On the microscopic level, the reliability of 3D CT NAND flash is affected by various aspects, such as the vertical charge loss through top and bottom oxides, the lateral migration towards spacers, the transient Vth shifting, and the limitations of vertical hole design [9]. When executing operations on one WL, the data stored in adjacent WLs which connected with the CG or U-Pipe connector will be affected by those mechanisms. Macroscopically, the program/erase (P/E) cycles, the program and read disturb, and retention time have significant impacts on the reliability of 3D NAND flash.

The number of electronics stored in cells will be changed after various disturbances, leading to the homologous Vth shifting and broadening. Fig. 2 depicts the ideal Vth distribution and possible real distribution after disturbances. X-axis is the Vth of TLC cell, and the states of {'111','011','001','000','010','110','100','101'} represent the logic data {'0','1','2','3','4','5','6','7'} respectively. Each state has 3 bits, denoted as the least significant bit (LSB), the central significant bit (CSB) and the most significant bit (MSB), respectively. And a WL contains 3 logic pages of LSB, CSB, and MSB. When reading data from cells, NAND flash first senses the Vth to judge the data state. If the Vth shift far from the original reference voltage, the data may be misread.



(a) Charge Trap Cell (b) P-BiCS Structure

Fig. 1: The structure CT NAND flash cell P-BiCS NAND array.

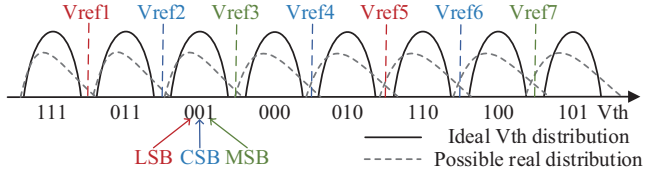


Fig. 2: The ideal Vth distribution vs. a possible real distribution.

III. Experiment Setup

We implement comprehensive experiments to investigate the reliability characteristics by testing 3D CT TLC NAND flash chips. The flash we test are TH58TFT1T22BA8H and TH58TFT1V23BA8H. The former has 48 horizontal layers and the latter has 64, which belong to BiCS2 and BiCS3 respectively. They have similar performance. The experiment can be divided into the reliability with P/E cycles, read disturb, and retention time experiments. We use RBER and Vth to characterize the reliability of NAND flash. The Vth distribution information can be obtained by adjusting the reference voltage and implementing the read try operation as [5] [6]. In all experiments, the randomly generated data are written to all healthy blocks selected with one-shot program operation. We had executed and recorded P/E operation in earlier times, so we can select different P/E cycles as we want from the chips.

1 The experimental process of reliability with retention time.

Definitions:

R_t: the number of retention time, be initialized to 0;
I_{R_t}: a vector of the interval of equivalent retention time;
MAX_{R_t}: the maximum value of retention time;

Process:

Select blocks from chips evenly with different P/E cycles;
Execute erase/program operation;
while $R_t <= MAX_{R_t}$ **do**
    Execute read operations;
    Compare and collect RBER and Vth;
    Sleep(I_{R_t});
    R_t += I_{R_t};
end while

2 The experimental process of reliability with read disturb.

Definitions:

Rd_{count}: the read counts, be initialized to 0;
$I_{Rd_{count}}$: the interval of equivalent read counts;
$MAX_{Rd_{count}}$: the maximum value of read counts;

Process:

Select blocks from chips evenly with different P/E cycles;
Execute erase/program operations;
while $Rd_{count} <= MAX_{Rd_{count}}$ **do**
    Execute sequential read operation in a block;
    Rd_{count}+=1;
    Sleep(1s);
    if (Rd_{count} mod $I_{Rd_{count}}$) is 1 **then**
        Compare and collect RBER and Vth;
    end if
end while

The algorithm 1 and 2 are the processes of reliability with retention time and read disturb. We bake the chips to accelerate the speed of retention at 85°C, and we calculate the acceleration factor (AF) by the Arrhenius equation. Previous work [7] tried to read a single page with tremendous millions of times to observe the read disturb in the block. However, in the real using of the flash-based system, the research result is hard to be applied, because the design of the read-times counter is space-consuming and a single page rarely happens to read millions of times. In addition, this test way may be limited by the spatial locality in a block and lead to unbalanced Vth distribution and RBER. Hence, to optimize the test method, we adopt a full-

block read sequentially. This method is more matchable to the design concept of flash-based systems. And it is easier to be applied with the unit block.

IV. Observations and Analyses

A. Reliability with P/E cycles and read disturb

As the increase of P/E cycles, NAND flash wears down to a great degree. The P/E cycles have a physical damage on NAND flash. So many previous works use the endurance represent the lifetime of NAND flash. But the enterprise flash-based systems commonly require the data can be correctly read out under a condition of 3 months' placement at 40 °C. Hence we invested test limited with the ability of ECC correct errors.

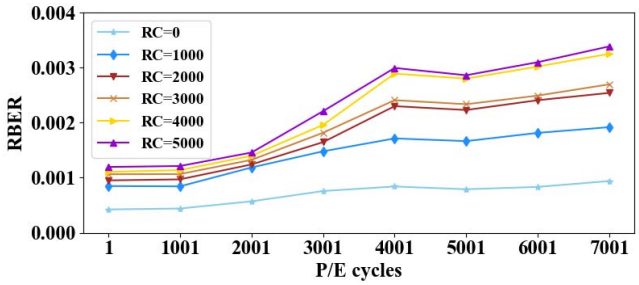


Fig. 3: The reliability of P/E cycles and read counts.

Fig. 3 illustrates this research result. X-axis is the P/E cycles and Y-axis is RBER. All the lines are figured by average values. Different lines represent different read counts. RBER increase as the growth of P/E cycles. It is interesting that the average RBER of 5000 P/E cycles even lower than 4000 P/E cycles'. The same phenomenon appears in different batches and models of the 3D CT NAND flash. We infer that the manufacturer optimizes the chips in different stages. Such as changing the read retry strategy and calibrate the reference voltage according to the latencies or other feedbacks. This phenomenon hasn't been observed in previous works, to our knowledge. In an experiment with a fine granularity of P/E cycles, RBER jitter and rise with the increase of P/E cycles. It is predictable. So like [11] [12], we can predict RBER and design effective algorithm with 3D CT NAND flash.

Moreover, RBER goes up with the read disturb. The full-block read function is executed, and each page evenly disturbed by the maximum 384000 (5000 multiplies 768, the page counts in a block we test) times' read. As our preliminary observation, RBER has an unevenly increase with read disturb.

B. Reliability with retention time

Another critical disturbance is retention time. In Fig. 4, X-axis is the retention time and Y-axis is RBER. For a better comparison of the disturbance weight of retention time and P/E cycles, we figured different P/E cycles' data with different lines in Fig. 4. RBER rise sharply in the early stage of retention time, and the rise speed slows down with the growth of retention time. In the early stage, the high Vth makes the main contribution of fast shifting. As the Vth becomes lower, the shifting speed and RBER growth slow down. This observation can be used to choose the best time for implementing refresh operations [13].

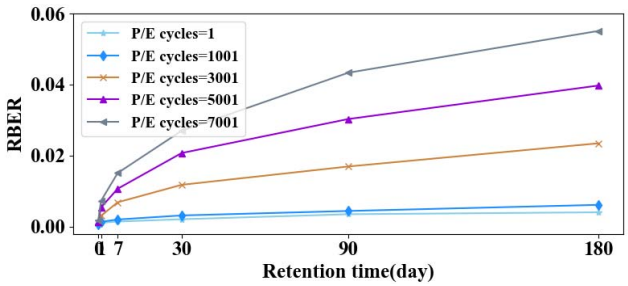


Fig. 4: The reliability with retention time.

C. Reliability variations of WLs

Fig. 5 depicts the reliability variations of WLs. We count the average RBERs of LSB, CSB, and MSB with different P/E cycles. The results may not match with some other works because of the differences of Gray Code design. In our results, the MSB has the highest RBER as shown in Fig. 5(a). RBERs have a significant variation with different WLs. In Fig. 5(b), the RBERs of all pages soar after 5000 full-block read operations. And the RBER of CSB has the most obvious growth. Finally, after 180 days' retention operation, RBERs experience a crazy increase. It is worth researching that RBERs tremendously increase with the WLs growth. We believe it is introduced by the vertical technology with stack layers. With the growth of layers, the retention time will affect the reliability more seriously. The new finding can be used to design diverse algorithms with different horizontal layers.

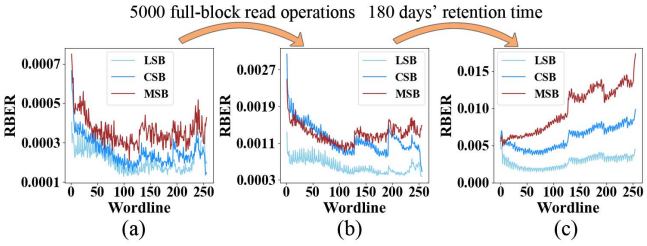


Fig. 5: The reliability variations of WLs.

D. Vth shifting with disturbances

We take a comprehensive analysis of VDS with disturbances. From Fig. 6 to Fig. 8, X-axis is the relative Vth, and the minimum step of X-axis is 1 with 10mv. Because the manufacturers keep the absolute value as a confidence, we can only calculate the relative value of Vth. The erase state (S0) is default state and it takes the relative Vth from 0 to 130. S1 to S7 represent the state 1 to 7, and Y-axis is probability density we test and calculated as [5] [6].

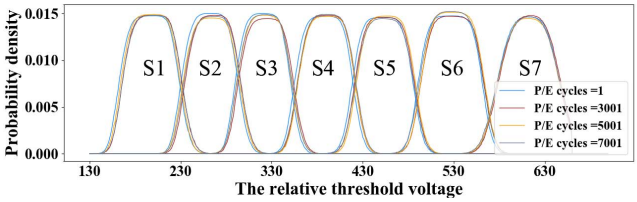


Fig. 6: The Vth shifting with P/E cycles.

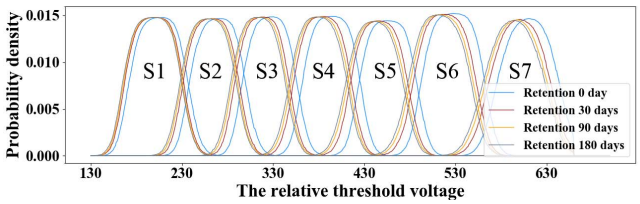


Fig. 7: The Vth shifting with retention time.

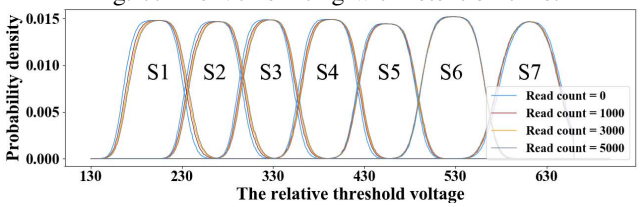


Fig. 8: The Vth shifting with read disturb.

As P/E cycles increase, the Vth has a right shifting. The Vth distribution in lower states shifts more obvious than higher states when evaluating the shifting with P/E cycles. But this analysis result is reversed with retention time, in Fig. 7. The shifting introduced by retention operation has a heavier impact on the higher states. And the shifting direction is also inverse, the Vth shifts to the left with retention operation but shifts to the right when the P/E cycles increase. With left shifting, the cross range is also broadened. It can be observed that as the retention time growth, the left shifting speed is gradually slowing down. Generally, the higher states with high voltage are easier to suffer from the left shifting, and the right shifting behaviors affect the lower states more.

The read disturb will also lead the Vth shifts to right, so it has a little impact on the higher states. As shown in Fig. 8, the right shifting speed of read disturb even lower than retention time in state 6 and 7, because the full-block read is really a time-consuming job. We conclude the preliminary results of Vth shifting in Table I. The Vth shifting will lead to RBER changes directly, it is critical for research the reliability of NAND flash. The VDS can be used to improve the read performance of SSD via the asymmetric cross scope [14]. And the different state behaviors of VDS can also help to improve the failure mode analysis [15].

TABLE I: The Vth shifting behaviors with disturbances.

Disturbance	Shifting direction	Heavier impact
P/E cycles	Right	Lower states
Retention time	Left	Higher states
Read disturb	Right	Lower states

V. Conclusion

In this paper, we analyze the differences between CT and FG NAND flash. Then we illustrate the experimental methodologies. With comprehensive experiments and analyses of data, we finally characterize the reliability and VDS of 3D charge trap NAND flash in depth. All the data are collected from a credible platform, and many new phenomena have been observed. The observations of reliability and VDS are very helpful to analyze the error pattern, fault modeling, and further researches. It has a practical and wider significance for optimizing the flash-based systems.

VI. Acknowledgment

The authors are grateful to the reviewers for their insightful comments that have greatly improved the article. This research is sponsored by the National Natural Science Foundation of China under Grant No 61872413, No. U1709220, No. 61821003, and Wuhan Science and Technology Project No. 2017010201010108, and the Fundamental Research Funds for the Central Universities No.2016YXMS019, and the 111 Project (No. B07038). This work is also supported by Key Laboratory of Data Storage System, Ministry of Education. We thank Memblaze Technology for providing the experimental platforms and technical support.

References

[1] Yajuan Du, Deqing Zou, Qiao Li, Liang Shi, Hai Jin, and Chun Jason Xue. Laldpc: Latency-aware ldpc for read performance improvement of solid state drives. In *International Conference on Massive Storage Systems and Technology*, 2017.

[2] J Liebault, D MoyaSiesse, J Bernardini, and G Moya. Charge trapping characterization in the thin oxide layer/nonconductive substrate system. *Surface Interface Analysis*, 34(1):668–671, 2010.

[3] Juhyung Kim, Changseok Kang, Sung Il Chang, Jongyeon Kim, Younseok Jeong, Park Chan, Joo Heon Kang, Sang Hoon Kim, Sunkyu Hwang, and Byeong In Choe. New phenomena for the lifetime prediction of tanos-based charge trap nand flash memory. pages 99–100, 2010.

[4] P. D Mauroux, A Virazel, A Bosio, and L Dilillo. Nand flash testing: A preliminary study on actual defects. In *Test Conference, 2009. ITC 2009. International*, pages 1–1, 2009.

[5] Yu Cai, Yixin Luo, Erich F. Haratsch, Ken Mai, and Onur Mutlu. Data retention in mlc nand flash memory: Characterization, optimization, and recovery. In *IEEE International Symposium on High PERFORMANCE Computer Architecture*, pages 551–563, 2015.

[6] Yu Cai, Yixin Luo, Saugata Ghose, Erich F. Haratsch, Ken Mai, and Onur Mutlu. Read disturb errors in mlc nand flash memory. pages 438–449, 2018.

[7] Qin Xiong, Fei Wu, Zhonghai Lu, Yue Zhu, You Zhou, Yibing Chu, Changsheng Xie, and Ping Huang. Characterizing 3d floating gate nand flash: Observations, analyses, and implications. *Acm Transactions on Storage*, 14(2):1–31, 2018.

[8] Meng Zhang, Fei Wu, Yajuan Du, Chengmo Yang, Changsheng Xie, and Jiguang Wan. Cooecc: A cooperative error correction scheme to reduce ldpc decoding latency in nand flash. In *IEEE International Conference on Computer Design*, pages 657–664, 2017.

[9] Rino Micheloni. *3D Flash Memories*. Springer Netherlands, 2016.

[10] Jim Handy. 3d nand: Benefits of charge traps over floating gates. http://thememoryguy.com/3d-nand-benefits-of-charge-traps-over-floating-gates/, November 22, 2013.

[11] Meng Zhang, Fei Wu, Xubin He, Ping Huang, Shunzhuo Wang, and Changsheng Xie. Real: A retention error aware ldpc decoding scheme to improve nand flash read performance. In *MASS Storage Systems and Technologies*, 2017.

[12] Debao Wei, Liyan Qiao, Shiyuan Wang, Ning Fu, and Xiyuan Peng. Research on prediction model for nand flash bit errors. In *IEEE International Conference on Electronic Measurement and Instruments*, pages 233–238, 2016.

[13] Yejia Di, Liang Shi, Kaijie Wu, and Chun Jason Xue. Exploiting process variation for retention induced refresh minimization on flash memory. In *Design, Automation and Test in Europe Conference and Exhibition*, pages 391–396, 2016.

[14] Qiao Li, Liang Shi, Chun Jason Xue, Qingfeng Zhuge, and H. M. Sha. Improving ldpc performance via asymmetric sensing level placement on flash memory. In *Design Automation Conference*, pages 560–565, 2017.

[15] Meng Zhang, Fei Wu, He Huang, Qian Xia, Jian Zhou, and Changsheng Xie. Fpga-based failure mode testing and analysis for mlc nand flash memory. In *Design, Automation Test in Europe Conference Exhibition*, pages 434–439, 2017.