The Analysis of 3D NAND Flash Memory

1st Guangzong Chen
Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, USA
guangzong@pitt.edu

***Abstract*—Memory is a necessary part of modern computer technology. For memory devices, the thing we most care about is its read/write speed and the raw bit error rate. Compared with two-dimensional NAND flash memory, 3D NAND flash has a large capacity due to the increased number of layers made on silicon. But 3D NAND memory does not simply stack planar memories together, it has a lot of different.**

This paper is a review of *3D NAND Flash Memory Lifetime by Tolerating Early Retention Loss and Process Variation*. The paper analyzes three main differences between 3D and planar NAND flash memory. These differences include storage cell different, structure different and manufacturing process different. Then analyze the raw bit error source caused by these differences. Also, it discussed the structure of 3D NAND flash memory and analyzes the reasons for errors in 3D NAND flash memory. Finally, it provides four methods to mitigate errors and improve the lifetime of 3D NAND flash memory.

***Index Terms*—3D NAND flash memory, raw bit error rate(RBER), lifetime**

I. Introduction

Solid-state drives (SSD) composed of NAND and NAND are now the focus of memory research. In recent years, most products have begun to use SSD as storage instead of traditional hard drives. Truly, SSD is much faster than hard disk. However, the capacity of SSD is smaller than the disk due to their structure difference. The storage cell in traditional planar NAND flash memory is tiled on silicon, which has a low space utilization. In order to increase space utilization, 3D NAND flash memory was developed. Compared to planar memory, 3D memory is more space-saving but also has a higher raw bit error rate.

The paper mainly focuses on the reason for the error and the solution of 3D NAND flash memory. In order to describe the raw bit error rate of 3D NAND flash memory, the difference between 3D NAND flash memory and planar flash memory must be introduced.

There are three main differences between planar and 3D NAND flash memory.

- The flash cell architecture that 3D NAND uses is different from planar NAND flash memory. Planar NAND uses a floating gate transistor. 3D NAND flash memory uses charge trap transistor, which stores charge within an insulator.
- 3D NAND has multiple layers of silicon on one chip. Planar NAND only has one layer.
- The manufacturing process of planar NAND flash memory is usually 10-15nm, but 3D NAND flash memory does not need to be so dense, generally 30-50nm.

For manufacturers, it is necessary to continuously reduce the production cost and increase the capacity of SSD. For planar NAND, the only way to increase capacity is increasing the density of memory cells per unit area. But when the density reaches a certain level, it is difficult to increase. Therefore, it must develop 3D NAND flash memory to increase storage capacity and reduce the cost.

It has to point out that because the 3D memory uses a different structure, it is less reliable than planar NAND. The paper has identified three new error sources that were not found in planar NAND flash.

- 3D NAND flash memory has layer-to-layer differences. This is a unique phenomenon in 3D NAND flash. The average error rate of each layer of 3D NAND flash is significantly different. The original bit error rate of the middle layer is 6 times to the top layer, which means 3D NAND flash memory has a layer to layer variation.
- Early retention loss. 3D flash memory has an early loss. Within a few hours after programming, a bit error occurred in the memory due to charge leakage. Experiments show that within three hours after programming, the error rate will increase more than 10 times.
- Keep interference. This phenomenon only exists in 3D NAND flash. The rate of charge leakage depends on the charge state of adjacent bits.

The paper discussed the distribution of raw bit error rate (RBER) changed by these three errors, and analyzed the distribution of RBER to propose the following four methods to solve this problem.

- Layer Variation Aware Reading.(LVAR)
- Layer Interleaved Redundant Array of Independent Disks(LI-RAID)
- Retention Model Aware Reading (ReMAR)
- Retention Interference Aware Neighbor-Cell Assisted Correction (ReNAC).

The specific description will show in the section below. Due to page limitations, some detail will be overlooked.

II. NAND Flash Memory Basics

Before analysis, the difference between 3D and planar NAND flash memory, the detailed composition of NAND

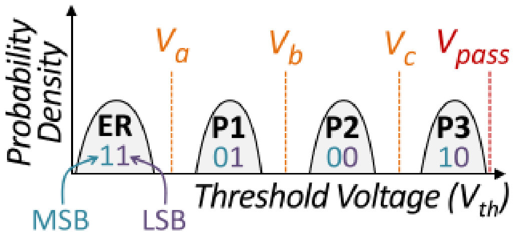

Fig. 1. Threshold voltage distribution and read reference voltages [4]

flash will be introduced. In flash memory, each memory cell contains a transistor to store charge. The different states of transistors represent different data stores in the transistor(0,1). In the read process, the different data stored in the transistor is read by different threshold voltages. Due to the variation of the memory during the production process, each state does not correspond to a single value, but the distribution of each status are following the Gaussian distribution.

The value stored in each cell depends on the threshold voltage. Fig.1 represents the relationship between the threshold voltage and stored value in 3D NAND flash memory. Since each memory cell in 3D NAND flash can store two bits, three threshold voltages are required to read data. At the same time, Fig.1 the graph shows that each state is not a fixed value. They are following normally distributed.

III. Difference between 3D NAND and planar NAND

According to the paper, there are three main differences. The first is the structure of the memory cell. The second is the way the memory cells arranged in the chip. The third is the process of manufacturing.

A. Flash Cell Design

As shown in Fig.2(a), planar NAND flash memory is using a floating-gate transistor. As shown in Fig.2(b), the charge trap transistor(CT) is used in 3D NAND flash memory In Fig.2, the position of several components is different, including the control gate, substrate, charge trap and the location of the charge stored. The thickness of the charged trap cell reduced a lot compared to the floating-gate cell. The biggest difference is that the charge trap transistor can store two bits, but the floating-gate transistor can only store one bit.

B. Flash Chip Organization

In planar NAND flash memory, all memory cells are aligned on one plane, and they are almost the same. The structure of 3D NAND flash memory not only stack multiple planar NAND flash memories but also use the etch technique to make multiple layers on a silicon chip. Because the charge trap transistor is used, it allows vertical distribution of bit-lines within a block. The control gates are connected together in a block to form a word-line. Then multiple blocks are stacked in the x-direction to form a 3D NAND flash memory.

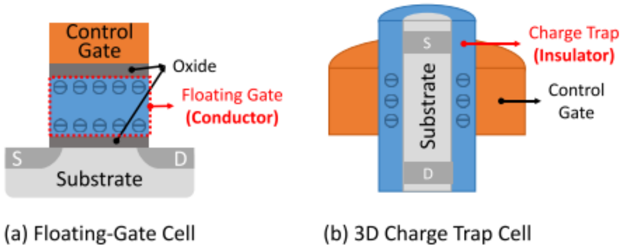

Fig. 2. The structure of floating gate cell and charge trap cell [4]

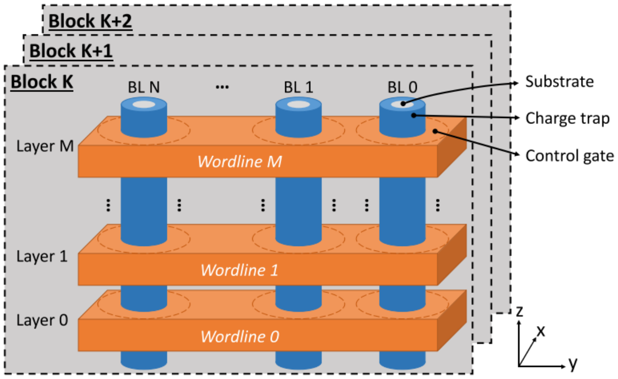

Fig. 3. 3D NAND flash memory organization. [4]

C. Manufacturing Process Technology

In the manufacturing process, planar NAND flash memory uses 10-15nm technology, but 3D NAND flash memory uses 30-50nm technology. Because the later one is more difficult to produce and has less limitation in space use.

IV. Characterization of 3D NAND Flash Memory Errors

Due to different structures, 3D NAND flash memory mainly has three different error sources.

- layer-to-layer process variation.
- Early retention loss.
- retention interference.

Some error is unique to 3D NAND FLASH, while others happen in both cases. Each of them will be discussed in the section.

A. optimal read reference voltage

Because after multiple read and write cycles, the distribution of each status will be changed. Fig.4 is the distribution of each status after the memory has been read and written 10^4 times. The distribution of each status gets closer. It can be found from the figure that after 10K P/E cycles, the gap of each status is very small.

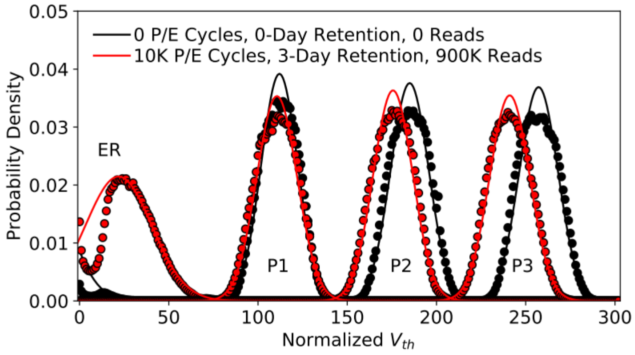

Fig. 4. 3D NAND threshold voltage distribution before(black) and after(red) [4]

From Fig.1, it can be seen that different thresholds can lead to different error rates. So, an optimal threshold voltage that has the least error rate should be found to read the data in the memory. Moreover, because the distribution of the storage state is changing all the time, the optimal threshold voltage should also change accordingly. It should not be a constant value.

B. layer to layer process variation

This error is caused by the different arrangements of memory cells between planar and 3D NAND flash memory.

Planar NAND flash memory has only one layer. But 3D NAND flash memory has multiple layers. Some work shows that the current etching technology cannot produce multiple layers of the same 3D NAND unit. As a result, each layer is very different. By reading and writing the data of different layers many times, and record the experimental data the effect of the layer to layer variation to error could be determined. The result is shown in Fig.5. In order to protect the information of manufacturers, the number of layers is normalized to 100 layers.

From Fig.5, the error rate of the most significant bit(MSB) is higher than the least significant bit(LSB). It is because determining the LSB requires only one threshold voltage, which is V_b (shows in Fig.1). But determining the MSB requires two threshold voltages, that is V_a and V_c.(shows in Fig.1). More threshold voltage, more errors.

It can also be found from the Fig.5 that the error rate of the middle layer is higher than that of other layers.

This is because 3D NAND flash memory is not simply stacked layers. Instead, it etches the silicon chip to form a 3D structure. The layers in the middle are the most difficult part to manufacture. Therefore, the error rate has also increased. [2], [6]

It can be seen from Fig. 6 that the optimal threshold voltage V_a, V_b, V_c between different layers are also different. In order to reduce errors, we need to select different read voltages for different layers.

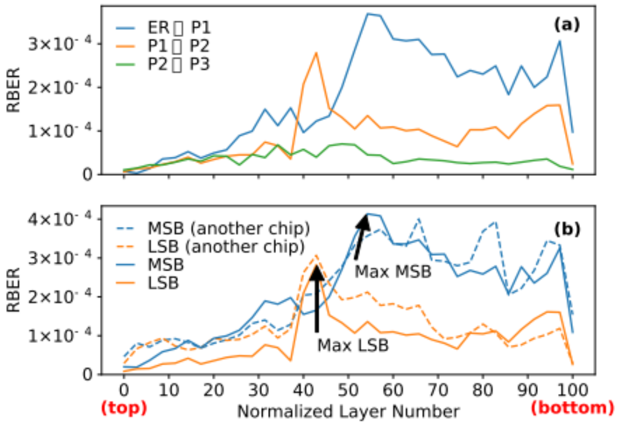

Fig. 5. variation of EBER across layers [4]

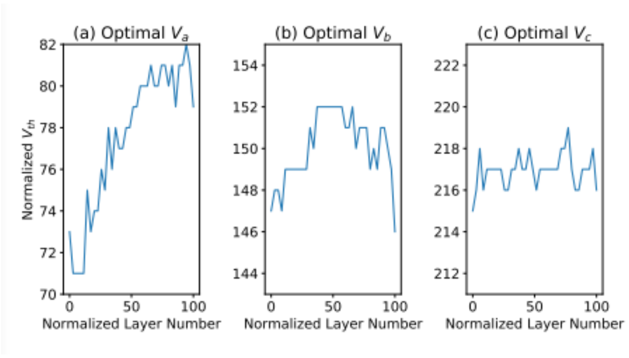

Fig. 6. variation of optimal read reference voltage across layers. [4]

C. Early Retention Loss

Retention errors means that the state stored in the flash store cell changed after programming. This error is caused by charge leakage at most time. Previous study indicates that especially within a few seconds after programming. Moreover, the charge trap flash cell is more possible to leakage charge compared to floating gate.

Fig.7 is a comparison between 3D and planar flash memory. From this Fig.7, we can see that the error rate of 3D NAND is much higher than planar flash memory.

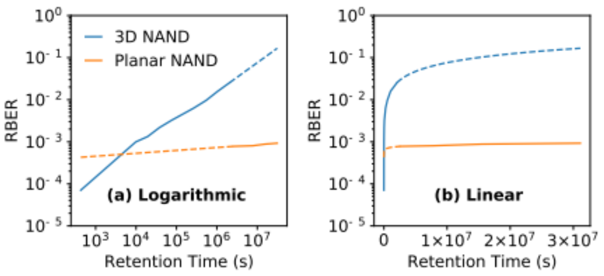

Fig. 7. Retention error rate comparison [4]

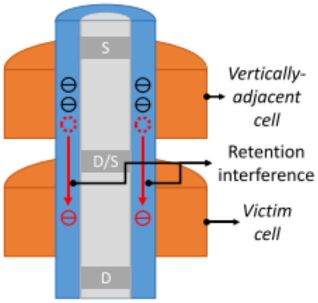

Fig. 8. Retention interference [4]

The early retention loss difference may be caused by two reasons. The first reason is that tunnel oxide layer is thinner in 3D flash memory. Due to the different storage structure, the tunnel oxide layer needs to be designed thinner in 3D to improve programming speed.

The second reason is that after programming, the charge may quickly leak to adjacent cells. As can be seen from Fig.3, the same bit line storage units are adjacent. Therefore, the leakage of charge is very easy to happen.

Through the study of early loss, a better strategy can be developed to avoid such errors . It can even improve the structure of memory cells to further reduce the impact of early wear and tear. [1], [3], [5]

D. retention interference

Retention interference only happens in 3D NAND flash memory storage. Retention interference refer to that if two adjacent storage cells stores different values, the charge will leak from the higher one to the lower one. As the Fig.8 shows, the top cell hold more charge than bottom cell. Some charge moves from the top cell to the bottom cell. So retention interference makes the voltage of top cell decreased and the voltage of the bottom cell increased. If the amount of transferred charge exceeds the threshold voltage, the data in the memory cell will be wrong.

V. 3D NAND ERROR MITIGATION TECHNIQUES

Based on the three errors listed above, the paper proposed four methods to solve these problems. In order to alleviate layer-to-layer process variation, LaVAR and LI-RAID are proposed. LaVAR predicts the optimal reading reference voltage by learning the layer-to-layer model in real time. Use learning model to tune the reference voltage for each layer. Thereby reduce the layer-to-layer variation loss.

The ReMAR is proposed to mitigate the retention loss. It is a new strategy that tracks the retention time information and stores it in the SSD controller. Then it uses this information to tune preset model.

In order to alleviate retention interference, ReNAC is proposed to help the correction of adjacent cells. Neighbor-Cell Assisted Correction(NAC) is a technology that already exists in planar NAND flash memory. The paper modified it and make it suit for 3D NAND flash memory.

A. LaVAR: Layer Variation Aware Reading

In the planar NAND technology, it is assumed that all memory storage cells are the same and RBER are the same. So it is reasonable to use a single threshold voltage. But in 3D NAND, this assumption is valid only inside one layer. Therefore, different threshold voltages should be use to reduce RBER. The key point to solving this problem is to determine the shifted value for each layer. What LaVAR has done is recording the number of P/E, and then predict the offset through the model. To achieve this effect, its SSD controller has a read-retry function. The optimal reading voltage is fixed by randomly reading and writing data of certain blocks of each layer. It is recorded in the SSD controller. This would fix errors caused by layer-layer variation and extend the lifetime of 3D NAND flash memory. Since this method is finally implemented on SSD controller firmware, there is almost no overhead.

B. LI-RAID: Layer-Interleaved RAID

In enterprise SSDs, RAID is used to help recover the error. But even the latest RAID does not take changes between 3D NAND layers int account. So is the difference in RBER between MSB and LSB. As a result, the new version of RAID may combine the two most unreliable components together, limiting SSD reliability.

Therefore, new RAID technology is designed to improve reliability. This design is based on two very simple ideas.

1. group unreliable layer and reliable layer
2. group MSB page with LSB page. Because of the reliability of LSB and MSB is different.

In order to solve Retention losses, the retention losses models should be constructed online. To achieve that, the controller randomly selects a flash block to test optimal read reference voltage, P/E cycle count and retention time. As soon as this information was obtained by the SSD controller, the optimal read and write voltage can be calculated. In other words, V_a, V_b, V_c can be evaluated based on different retention times.

By accurately predicting read reference voltage, ReMAR increases the accuracy of the data and decreases the raw bit error rate, which extends flash memory lifetime.

C. ReNAC:Retention Interference Aware Neighbor-Cell Assisted Correction

The basic idea of ReNAC is very simple. ReNAC is developed to solve retention interference. So ReNAC read the status in Neighbor-cell. Then it adjusts read voltage(V_a, V_b, V_c) in real-time to mitigate raw bit error rate and improve flash memory lifetime.

VI. SUMMARY

This paper reviews the works done by *Improving 3D NAND Flash Memory Lifetime by Tolerating Early Retention Loss and Process Variation*. The main differences between the 3D and 2D NAND flash memory are discussed in detail. Three different error sources in 3D flash memory are described.

Based on these problems, four practical solutions are proposed. These four methods can effectively reduce the errors in 3D NAND flash memory and extend the life of the memory.

References

[1] Bongsik Choi, Sang Hyun Jang, Jinsu Yoon, Juhee Lee, Minsu Jeon, Yongwoo Lee, Jungmin Han, Jieun Lee, Dong Myong Kim, Dae Hwan Kim, Chan Lim, Sungkye Park, and Sung-Jin Choi. Comprehensive evaluation of early retention (fast charge loss within a few seconds) characteristics in tube-type 3-d nand flash memory. In *2016 IEEE Symposium on VLSI Technology*, pages 1–2, 2016.

[2] Chun Hsiung Hung, Meng Fan Chang, Yih Shan Yang, Yao Jen Kuo, Tzu Neng Lai, Shin Jang Shen, Jo Yu Hsu, Shuo Nan Hung, Hang Ting Lue, Yen Hao Shih, Shih Lin Huang, Ti Wen Chen, Tzung Shen Chen, Chung Kuang Chen, Chi Yu Hung, and Chih Yuan Lu. Layer-Aware Program-and-Read Schemes for 3D Stackable Vertical-Gate BE-SONOS NAND Flash Against Cross-Layer Process Variations. *IEEE Journal of Solid-State Circuits*, 2015.

[3] Yixin Luo, Saugata Ghose, Yu Cai, Erich F. Haratsch, and Onur Mutlu. HeatWatch: Improving 3D NAND Flash Memory Device Reliability by Exploiting Self-Recovery and Temperature Awareness. In *Proceedings - International Symposium on High-Performance Computer Architecture*, 2018.

[4] Yixin Luo, Saugata Ghose, Yu Cai, Erich F. Haratsch, and Onur Mutlu. Improving 3D NAND Flash Memory Lifetime by Tolerating Early Retention Loss and Process Variation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(3):1–48, dec 2018.

[5] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. *PROC. OF THE 9th PYTHON IN SCIENCE CONF*, 2010.

[6] Yi Wang, Lisha Dong, and Rui Mao. P-alloc: Process-variation tolerant reliability management for 3d charge-trapping flash memory. *ACM Trans. Embed. Comput. Syst.*, 16(5s), September 2017.