Characterizing 3D Charge Trap NAND Flash: Observations, Analyses and Applications

ABSTRACT

With the growing demand for storage capacity, 3D NAND flash memory is attracting increasing attention from both academia and industry. In 3D era, the Charge Trap (CT) NAND flash is employed by mainstream products, thus having a deep understanding of its characteristics is becoming increasingly crucial for designing flash-based systems. In this paper, we implement comprehensive experiments on advanced 3D CT NAND flash chips by developing an ARM- and FPGA-based evaluation platform. Based on the experimental results, we first make distinct observations on the characteristics of 3D CT NAND flash, including its reliability and performance features. Then we give analyses of the observations from physical and circuit aspects. Finally, based on the unique characteristics of 3D CT NAND flash, approaches to optimize the flash management algorithms in real applications are presented.

KEYWORDS

3D charge trap NAND flash, reliability, performance

1 INTRODUCTION

Over the past years, NAND flash has become the mainstream storage medium for various systems, ranging from embedded systems, desktops to high performance computing (HPC) servers, since it provides non-volatility and higher density compared with DRAM and much better performance than hard disk drives (HDDs). In order to satisfy the sustained demand growth in the storage market, vendors continuously shrink the feature size according to the Moore's Law, improving the density and meanwhile narrowing the gap of price-per-bit between NAND flash-based storage and HDDs. However, due to the technical complexity, the scaling down becomes increasingly challenging and costly, especially when the feature size is below 30nm. Moreover, the reliability of NAND flash decreases dramatically with the scaling down, as a result of thinner tunneling oxide, fewer electrons retained by a cell [14] and more interference among cells [18], etc. Hence, it is obvious that the contradictions of the capacity and reliability in traditional planar NAND flash technology make it no longer able to adequately meet the requirements of both, which poses greater urgency of innovative flash architecture.

3D NAND flash represents a very promising solution to conquer the contradictions in planar devices. By enabling scaling in Z direction rather than X and Y directions, tremendous improvements in capacity and lower price-per-bit can be achieved, while vendors can employ larger feature sizes to guarantee the reliability. Nowadays, in comparison with 15 ∼ 20nm used in planar NAND flash, the processes of 3D NAND flash are usually larger than 40nm. In 3D era, there are two noteworthy technology trends:

- *Charge Trap is the mainstream.* Unlike planar NAND flash products, which are based on *Floating Gate (FG)* [17] technology, all vendors except Micron/Intel joint venture have chosen *Charge Trap (CT)* [15] in their 3D NAND products. The main difference between these two technologies is that an FG cell employs conductive polycrystalline silicon as the medium to store electrons while a CT cell adopts insulating charge trapping layer, which is typically made of silicon nitride. Compared with FG NAND flash, CT NAND flash is fundamentally with better scalability and less coupling effects among cells [16].
- *TLC dominates. Triple-level cell (TLC)* NAND flash, which stores 3 bits of data within a single cell, only accounts for a small fraction of the planar NAND flash market due to the poor reliability compared with *single-level cell (SLC)* (1 bit per cell) and *multi-level cell (MLC)* (2 bits per cell). Due to the employment of the rolled back process technology node, the emerging 3D TLC is able to achieve the technical specifications of the last planar MLC flash, and has accounted for over 50% of the industry bits by 2017.

In NAND flash, once a page is programmed, it cannot be over-programmed until being erased (*out-of-place update*). However, program and erase operations are performed at different granularities (program→page and erase→block). In order to simulate NAND flash as a block device in general-purpose file systems by hiding the out-of-place update and block-erase properties, flash translation layer (FTL) is employed between file systems and NAND flash. Due to the unique characteristics of NAND flash, FTL needs to be optimized by utilizing them to improve the performance and reliability of NAND flash-based storage devices. For planar NAND flash, the characteristics have been comprehensively and deeply investigated, and several NAND flash management methods have been proposed based on the characteristics and achieved high efficiency. Nevertheless, only a few prior studies have been conducted on the characteristics of 3D NAND flash. Xiong et al. have measured 3D FG NAND flash, extracted several observations from the experimental results, and given some implications on the designs of 3D FG NAND flash-based storage devices. Due to the differences in structures and materials between FG and CT NAND flash, those observations and implications of 3D FG NAND flash cannot be applied to 3D CT NAND flash. Hence, it is extremely critical and urgent to have a profound understanding of the characteristics of 3D CT NAND flash.

In this paper, we comprehensively characterize the reliability and performance features of 3D CT NAND flash and make meticulous analyses. We hope our work can give rise

CASES'18, September 2018, Torino, Italy
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.475/123_4

to a better understanding of the characteristics and help researchers design optimized flash management algorithms for 3D CT NAND flash.

The rest of this paper is organized as follows. Section 2 introduces background of the work. Section 3 shows experimental setups. Evaluation results, analyses and applications are presented in Section 4. Section 5 demonstrates related works and Section 6 concludes this paper.

2 BACKGROUND

2.1 3D Charge Trap NAND Flash

NAND Flash Cell. A NAND flash cell stores 1-bit or several bits of data by dividing the threshold voltage (V_{th}) into multiple regions. The region in which the V_{th} of a cell locates represents the current value of data. As illustrated by Fig. 1, seven *read reference voltages* (V_{ref}), V_{ref1} to V_{ref7}, divide a TLC cell into eight states, E and S_1 to S_7, which can be decoded into 111, 110, 100, 000, 010, 011, 001 and 101, respectively. In this paper, we use the format ABC to represent the 3-bit data stored in a TLC cell, where A, B and C denote the least significant bit (LSB), the center significant bit (CSB) and the most significant bit (MSB), respectively.

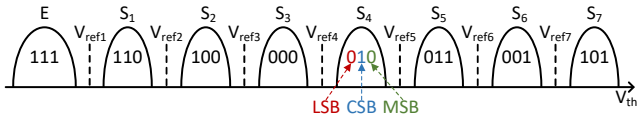


Figure 1: Threshold voltage distribution of TLC NAND flash.

3D CT NAND Flash Organization. When tens of millions of NAND flash cells are arranged together in a 3D structure, a NAND flash array can be formed. Fig. 2 shows the bird's eye view of a 3D CT NAND flash array. The medium responsible for data storage is made up of an array of word lines (WLs), which are continuously connected from top to bottom along the channel side, and are connected to the source line and bitlines through source line selectors (SLSs) and bitline selectors (BLSs), respectively. A WL is the basic unit of program operations and each WL is composed of 3 pages: lower page, middle page and upper page, which correspond to LSB, CSB and MSB, respectively. The 4 WLs in the same X-Y plane form a tier and their control gates are connected together (in order to clearly illustrate WLs, the WLs in the same tier are not connected in Fig. 2). Cells sharing a channel make up a string and the WLs in the same X-Z plane are in the same strings.

As with planar NAND flash, there are three basic operations for 3D CT NAND flash: *read*, *program* and *erase*.

Read. A read operation obtains data stored in the target page. When a target page is read, SLS and the corresponding BLS are turned on, and all the tiers exclusive of the target page are applied a read pass voltage (V_{pass}) so that data in the target page can be properly propagated to the output. Meanwhile, a sequence of read reference voltages (V_{ref}) are applied to the tier containing the target page in order to read out the data. According to the type of the target page, read reference voltages are divided into 3 sets: lower page→{V_{ref3}, V_{ref7}}, middle page→{V_{ref2}, V_{ref4}, V_{ref6}}, and upper page→{V_{ref1}, V_{ref5}}. During a read operation, the read reference voltages in the corresponding set are sequentially applied. For example, when reading a lower page, V_{ref3} is first applied, followed by V_{ref7}. If V_{ref3} cannot turn on the transistor while V_{ref7} can, then the threshold voltage of the cell is in between V_{ref3} and V_{ref7}, indicating that the LSB data in the cell is 0, otherwise it is 1.

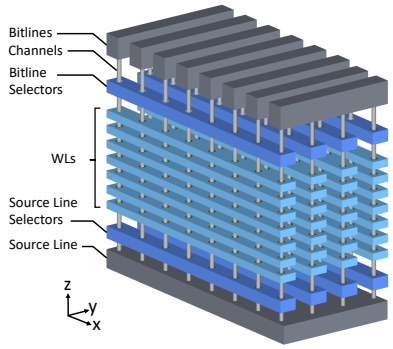


Figure 2: Bird's eye view of 3D CT NAND flash array.

Program. Program operations store specific data into the target pages. In 3D CT NAND flash, program operations are performed in a one-shot behavior, which means that the three pages in the selected WL are programmed simultaneously. Similar to the read operations, during a program operation, SLS and the BLS corresponding to the target WL are turned on, and all the tiers except for the tier containing the WL are applied a V_{pass}, so that the target WL can be selected. Meanwhile, a series of incremental staircase program pulses and program verify (PV) processes are performed. The program pulse is implemented by applying a high voltage, V_{pgm}, to the control gates of the tier containing the target WL so that electrons can be charged into the storage layers of the selected WL. After each program pulse, a PV process follows. The PV process is basically a read operation that verifies whether the threshold voltage of each cell exceeds the target value (i.e. whether enough electrons are charged into each cell). If there are still some cells whose target value has not been reached, then V_{pgm} is incremented by ΔV_P and another program pulse is needed. Otherwise, the program operation is completed.

梯度递增编程脉冲

Erase. An erase operation wipes all the data in the selected block. Similar to program operations, the erase operation involves a sequence of erase pulses and hard erase verify (HEV) processes. The erase pulse is achieved by applying a high voltage, V_{erase}, between channels and control gates, which move the thresholds of all cells in a block towards the E state. After each erase pulse, an HEV process follows. The HEV process checks if there are still some cells that have threshold voltages higher than V_{ref1}, in which case V_{erase} is incremented by ΔV_E and another erase pulse needs to be applied. Usually, a maximum number (E_{max}) of erase

X-Z平面上的所有WL都属于同一个通道，即如图2，X-Z平面上的8个柱子都是同一个通道的，而不是只有图示标注指向的那一个柱子

另：图7是一个块中各层的RBER示意图，由此可见，一层属于同一块，则不仅是X-Z平面上的8个柱子属于同一通道，图2中X-Y-Z三维上的8*4=32个柱子都属于同一通道

pulses is set. If E_{max} is reached, then the erase process exits with erase failure. Otherwise, when cells in the selected block are all set to the E state, the erase operation is successfully finished.

2.2 Metrics

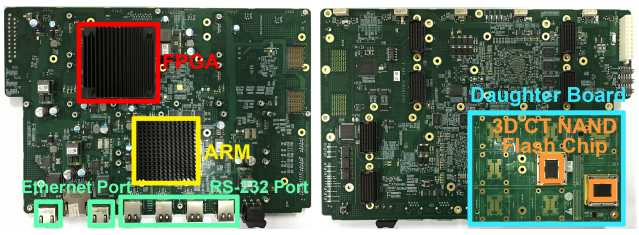
For data storage, reliability and performance are always the two most important technical indicators. The main metrics of the reliability and the performance of NAND flash are briefly introduced in this section and detailedly measured, observed, analyzed and discussed in the following sections.

2.2.1 Reliability. During the lifetime, NAND flash can fail due to a wide range of reasons, which jointly threat the data safety. After repeated erase and program operations, cells become unreliable due to trap creation in tunnel oxide and interfacial damages until erase or program failures arise. ***Endurance*** denotes the number of P/E cycles that NAND flash can withstand before a failure appears. Besides, data can be corrupted by four other main factors, ***retention***, ***fast detrapping***, ***read disturb*** and ***program disturb***, all of which do not damage cells permanently. Retention and fast detrapping errors result from charge leakage over time. Read disturb is a phenomenon that a read operation causes a weak programming effect on the other (unread) WLs in the same block. When programming a WL, electrons can be *unintentionally* injected into the cells of other WLs in the same block, caused by parasitic capacitance or a *weak programming effect*, called program disturb.

2.2.2 Performance. Nowadays, with the rapid development of NAND flash interface protocols, the latest interfaces could provide up to 800MT/s and 400MT/s when they run on Open NAND Flash Interface (ONFI) 4.0 and Toggle 2.0, respectively. Unfortunately, the performance of NAND flash chips is greatly limited by internal latencies. ***Erase latency***, ***program latency*** and ***read latency*** refer to the time consumed for wiping out all data stored in a block, for programming data stored in data register to the NAND array, and for copying data from NAND array to cache register and enabling data output from the cache register to host, respectively.

3 EXPERIMENTAL SETUP

3.1 Experimental Platform



(a) The front (b) The back

Figure 3: 3D NAND flash experimental platform.

In order to accurately measure the performance and reliability of 3D CT NAND flash, we build an ARM- and FPGA-based NAND flash experimental platform, *General Storage Tester (GST)*, that enables us to directly control raw NAND flash chips without ECC, as shown in Fig. 3. The *ARM* runs a stripped-down Linux operating system, which is responsible for 1) receiving commands/data from a personal computer (PC) via 1Gbps Ethernet ports or RS-232 ports; 2) parsing user commands into atomic commands supported by NAND flash chips; 3) transferring atomic commands/data to the FPGA; and 4) receiving returned data from the FPGA and sending them back to the PC. A custom flash controller is implemented in the *FPGA*, which transforms the commands/data from the ARM into the corresponding signals to control NAND flash chips and reads data from NAND flash chips to the ARM. By monitoring the ready/busy (R/$\overline{\text{B}}$) signal, the FPGA supports measuring the latency of each atomic command executed by NAND flash chips with the precision of $1\mu s$. An experimental platform can connect to up to 4 daughter boards, each of which supports at most 8 NAND flash chips.

3.2 Experimental Acceleration

We measure NAND flash chips over various P/E cycles and different retention ages at room temperature (25^{o}C) to imitate the real environment. In order to characterize reliability after long retention ages (e.g., 1 year), we accelerate retention error tests under high temperature according to the Arrhenius Law [2]. Table 1 shows the accelerated retention ages (t_{acce}) at various high temperatures (T_{acce}) to achieve the corresponding equivalent retention ages (t_{room}) at room temperature (T_{room}) used in our experiments. By putting the tested 3D CT NAND flash for 12.9 *hours at* 85^oC, it suffers a *1-year equivalent retention age at* 25^oC.

t_{room} (T_{room})	t_{acce} (T_{acce})
1 month ($25°C$)	$2.7h$ ($75°C$)
6 months ($25°C$)	$16.2h$ ($75°C$)
1 year ($25°C$)	$12.9h$ ($85°C$)
3 years ($25°C$)	$24.8h$ ($90°C$)
5 year ($25°C$)	$26.8h$ ($95°C$)

Table 1: Accelerated and equivalent retention ages.

3.3 Experimental Object

In our experiments, we choose a representative 3D CT NAND flash product, *BiCS2*[1] *TLC from Toshiba.* It uses Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) structure and its charge traps are similar to 3D CT NAND flash from SK Hynix and Samsung. The main parameters of our measured chips are listed in Table 2.

[1]Since the later generations, BiCS3 and BiCS4, only have *engineering samples* available and in order to provide guidance for the design of real flash-based storage systems through our characterization, we use the mature *customer samples* of BiCS2 as our experimental object.

Parameter	Value
Capacity	2.48Tb (310GB)
Page size	(16384 + 1952) bytes
Block size	576 pages
Plane size	1972 blocks
Die size	2 planes
Target size	2 dies
Chip size	4 targets

Table 2: Parameters of 3D CT NAND flash chips.

3.4 Experimental Methodology

Random Data. In actual products, data randomization mechanism is widely employed in flash controller or integrated into flash chips to reduce the raw bit error rates (RBERs) and prolong the lifetime. Therefore, we implement a pseudo-random binary sequence (PRBS) generator to produce random data and use a scrambler designed for 3D NAND flash to further randomize the data. The randomized and scrambled data[2] are adopted in each program operation to mimic real data.

P/E Cycling. Blocks are repeatedly erased and programmed with random data to simulate the real process of usage. For all experiments, a dwell time of $10s$, which is the idle time between an erase operation and a program operation to the same block, is employed at room temperature.

In this paper, the specified experimental methodologies of each metric are introduced in the corresponding part of Section 4.

4 EVALUATION RESULTS

4.1 Reliability

Reliability of NAND flash is crucial for the stability of storage systems. In this subsection, reliability issues in 3D CT NAND flash are detailedly researched.

4.1.1 Endurance. The endurance of a NAND flash block represents its lifetime. In order to quantitatively characterize the endurance of 3D CT NAND flash, we randomly choose 100 blocks from the flash chips as samples to represent general situations, repeatedly program and erase the blocks until they fail, and record the RBERs of each page and block for every 100 P/E cycles. Fig. 4 exhibits endurances of the selected blocks, and Fig. 5 plots the variations of RBERs with the increase of P/E cycles.

Observation 1. The endurance of 3D CT TLC NAND flash reaches similar levels of planar MLC NAND flash.

As shown in Fig. 4, the endurances of 3D TLC NAND flash blocks exhibit an average of 35,417 P/E cycles, showing a significant longer lifetime than planar TLC NAND flash (typical value: 3,000 P/E cycles), and is comparable with planar MLC NAND flash (typical value: 30,000 P/E cycles). The outstanding endurance benefits from the larger feature size, the new materials and structure. Among these selected blocks, the maximum endurance reaches 37,883 P/E cycles, and the minimum is 32,921 P/E cycles, resulting in a standard deviation of 1,064 P/E cycles.

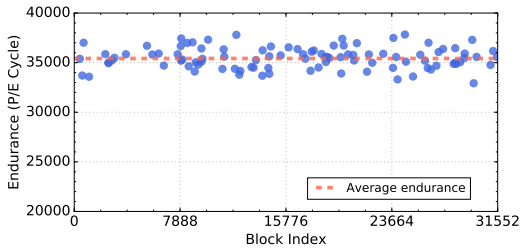


Figure 4: Endurance distributions among 100 blocks.

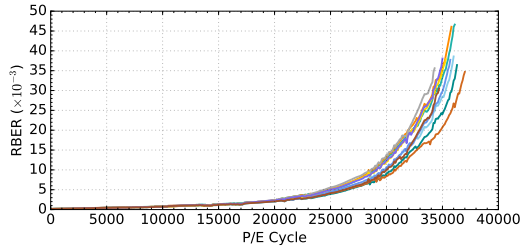
Observation 2. RBERs show a near exponential growing trend with the increase of P/E cycles.

Fig. 5a exhibits the growing trend of RBERs with P/E cycles for each block from the beginning of the test until they fail, with different colors referring to different blocks. We can observe from the figure that block RBERs remain low (under 2.5×10^{-3}) within the first 20,000 P/E cycles, and grow sharply subsequently, which can reach up to 4.6×10^{-2} before failures occur. Moreover, due to the process variations among different blocks, block RBERs exhibit various growing speeds, which usually have a positive correlation with endurance since those whose RBERs grow faster wears more seriously. However, we can also observe from the figure that although a few blocks have relatively slower growing speeds, they end up with shorter lifetimes. This phenomenon can be explained by the "short board effect" that the endurance of a block is dependent on the page with the worst reliability. Due to the process variations among different pages within a block, tunnel oxide of a page may wear seriously while all the other pages in the same block have just experienced a slight wear, in which case a block may fail earlier than expected.

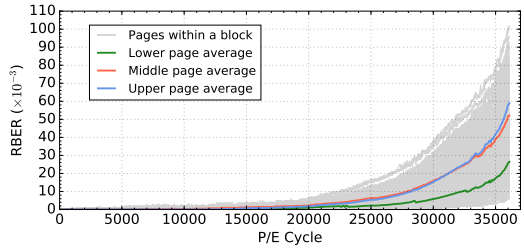
In order to further explore the page RBERs variation with P/E cycles, we analyze the RBERs of each page within a block, and the results are shown in Fig. 5b. We can observe from the figure that page RBERs also exhibit a near exponential growing trend with P/E cycles. However, page RBERs exhibit a dispersive distribution, with the highest RBERs end up with more than 100×10^{-3}, while the lowest less than 10^{-3}, indicating enormous process variations among pages. We can also observe that the average RBERs of upper and middle pages grow with a similar trend, while the average RBERs of lower pages grow more slowly and end up with approximately half of that of the other two types of pages. This phenomenon can be explained by the encoding rule of 3D CT NAND flash. According to Fig. 1, when threshold voltage shifts to adjacent states, the value of one bit will flip (change form "1" to "0" or from "0" to "1"), resulting in bit errors, and the probability that threshold voltages of CSBs shift to adjacent states are higher than that of LSBs and MSBs. Moreover, since P/E operations shift threshold voltages to higher values, states with lower threshold voltages tend to shift more. Therefore, the MSBs, which have states

[2]No data randomization mechanism is integrated into Toshiba BiCS2 chips. The generated data are directly programmed to flash cells.

with lower threshold voltages compared with LSBs, are more likely to flip. Hence, the upper and middle pages exhibit higher average RBERs than lower pages.



(a) Block unit



(b) Page unit

Figure 5: RBERs variations with P/E cycles.

Application. Traditionally, a block is identified as bad block if one of its pages fails. However, due to the huge page variations inside a block, when a page fails, the other pages in the block might be in quite healthy states, yet they can no longer be used, resulting in the underutilization of storage capacity. A *fault-tolerant block* approach that tolerates multiple failed pages can overcome this problem. Through the implementation of this approach, a block is identified as a bad block only if the number of failed pages exceeds a preset threshold, increasing the usage rate of NAND flash and prolonging system lifetime.

4.1.2 Retention error. Retention errors occur when a chip is placed for a long time without any operations, which usually happens to storage systems for cold data. In order to characterize retention errors, we randomly choose 100 fresh blocks and divide them into 4 groups, then employ random data to repeatedly program and erase each group to 0, 10,000, 20,000 and 30,000 P/E cycles, respectively. After that, we place the chips under room temperature for 1 day and 1 week, then under high temperature to accelerate the retention process, as shown in Table 1. At the end of each retention period, we read the selected blocks under room temperature to get the RBERs. Fig. 6 exhibits the change of block RBERs with retention time, and Fig. 7 plots the distributions of RBERs before and after the retention process.

Observation 3. RBERs show a near-logarithmic growth with the increase of retention time.

As shown in Fig. 6, RBERs grow sharply within the first 6 months of retention time, and increase mildly afterward. The gradually decreasing growing speed of RBERs results from the fact that when NAND flash chips are placed for a long time without operations, electrons gradually escape from the storage layer since they form an electric field. The field intensity is proportional to the number of electrons inside the storage layer. Therefore, as the number of electrons inside the storage layer gradually decreases with the increase of retention time, the electric field intensity is reduced, thus slowing down the increasing of retention errors. We can also observe from the figure that a large P/E cycle corresponds to higher RBERs. This phenomenon can be explained by the characteristic of NAND flash that the tunnel oxide wears more seriously with a larger P/E cycle, thus electrons are easier to escape from the storage layer, resulting in higher RBERs. Moreover, an interesting phenomenon can be observed from the figure that the RBERs experience a slight drop after 1 day's retention under the P/E cycle of 30,000. Since the P/E operations and the retention process shift threshold voltages to different directions, the retention process partly "repairs" the errors caused by P/E operations.

此现象的出现是由于 P/E 操作使存储单元的阈值电压向右偏移，而数据放置会导致存储单元的阈值电压向左偏移，因此，在某个特定的时间点，数据放置在一定程度上“修复”了由 P/E 操作造成的错误。在此之后，由于数据放置使阈值电压继续向左偏移，从而造成错误率的继续上升。

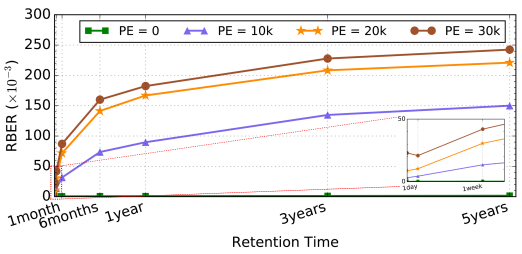
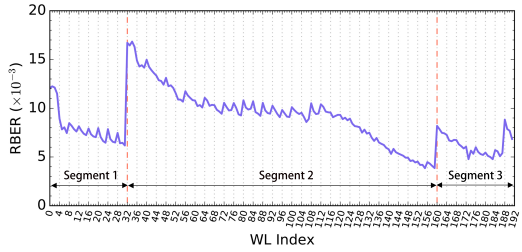


Figure 6: RBERs variations with retention time.

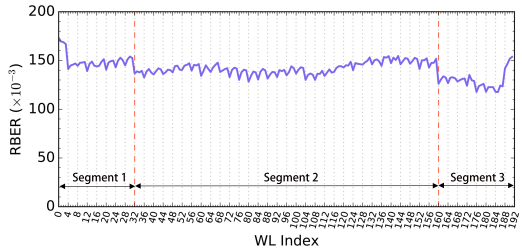
Observation 4. RBERs exhibit cross-tier variations after repeated P/E operations, while retention process reduces the variations.

Fig. 7a exhibits the distributions of RBERs within a block after performing 20,000 P/E operations. We can observe from the figure that the curve can be roughly divided into 3 segments according to the variation regularity: Segment 1 (Tier 0 to Tier 7), Segment 2 (Tier 8 to Tier 39) and Segment 3 (Tier 40 to Tier 47). The RBERs differences among tiers can be explained by the segmented voltage and the fringe effect in 3D CT NAND flash. The fabrication process of 3D CT NAND flash results in variations in different tiers. In order to reduce these cross-tier variations, the tiers are divided into several groups and applied with different voltages, resulting in the segmented RBERs distributions. Moreover, as a result of the fringe effect, RBERs of Tier 0 and Tier 47 are higher than adjacent tiers. However, these cross-tier variations are reduced by the retention process. As shown in Fig. 7b, the RBERs within a block exhibit a near uniform distribution after 6 months of retention time. This can be explained by the intrinsic characteristics of retention process. Different from P/E operations which apply different voltages to different segments, the retention process exerts uniform influences on all WLs. Therefore, after a long period of retention time, differences among tiers are reduced. Moreover, we can observe from Fig. 7b that the average RBERs of Segment 1 is slightly

higher than that of Segment 2, which is also slightly higher than that of Segment 3 after retention. The average RBERs of the three segments shows the process variations. Since cells in Segment 1 have the smallest size, their tunnel layers wear more easily, thus electrons are easier to escape from the storage layer, resulting in higher RBERs. On the contrary, cells in Segment 3 have the largest size, thus exhibiting the lowest RBERs among the three segments.



(a) Before test



(b) 6 months of retention

Figure 7: RBERs distributions under 20000 P/E cycles before test and after 6 months of retention time.

硕士论文中对应段更好：p36应用5

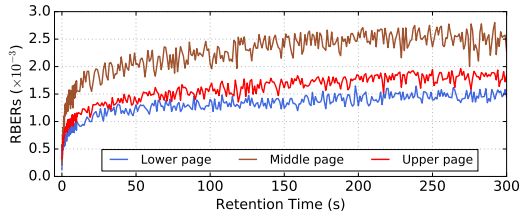
Application. Placing data for a long period of time (cold data) causes high retention errors. However, the effect of retention only exists in a single P/E cycle. Therefore, a *refresh* strategy, which moves the cold data (with error corrections) periodically, can be adopted to eliminate this problem.

4.1.3 Fast detrapping phenomenon. As mentioned in Sec. 2, we program a cell by charging a certain amount of electrons into the storage layer. However, electrons charged into the storage layer may escape immediately after programming, resulting in undesirable errors, this phenomenon is called "fast detrapping". In order to characterize the fast detrapping phenomenon, we choose blocks that are programmed and erased to different P/E cycles, and for each block, perform the following operations: 1) program a WL (instead of a whole block in order to observe the RBERs variations immediately after programming) with random data and record the programmed data; 2) read the three pages in the programmed WL every 100 milliseconds for 300 seconds. Finally, we compare the programmed data and the data read out to obtain the variations of page RBERs within the first 300 seconds after programming, and a representative fast detrapping phenomenon is shown in Fig. 8.

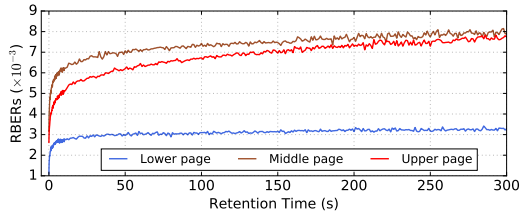
Observation 5. RBERs rise sharply within a small period of time after programming, and remain nearly constant after saturation point is reached.

Fig. 8 exhibits the growing trends of page RBERs shortly after a WL has been programmed under the P/E cycle of 0 and 30,000. We can observe from the figure that RBERs of the three pages increase rapidly within the first 10 seconds after programming, with lower, middle and upper page reaching 8.2×, 7.2× and 3.4× of the initial RBERs in fresh chips, respectively. Under the situation of 30,000 P/E cycles, the three values are 2.6, 2 and 1.9, respectively. This phenomenon is due to the mechanism that a portion of electrons are trapped in shallow traps during the programming procedure, thus are less stable compared with electrons trapped in deep traps. Therefore, they can easily escape from the storage layer through defects in oxide immediately after programming, shifting threshold distributions towards lower values, thus resulting in the sharp rise of RBERs in the initial stage. After the majority of the unstable electrons escape from the storage layer, charge loss becomes slow and steady, similar to the retention phenomenon in traditional NAND flash. With the increase in P/E cycles, threshold distributions shift to higher values after program operations, which is in the opposite shift direction of fast detrapping phenomenon and weakens it. Thus, the rising rates under 30,000 P/E cycles are lower than those under 0 P/E cycles.

P/E高则氧化层磨损更多，编程时采用步长脉冲实现编程及其应用，氧化层磨损越厉害，则相同电压脉冲下注入的电子更多，使阈值电压右偏（右偏只是相较于某一具体数据的峰来说有所偏移）



(a) Fresh block



(b) Under the PE cycle of 30,000

Figure 8: Fast detrapping phenomenon.

Application. Fast detrapping phenomenon results from the electrons trapped in shallow traps during program operations, hence a *re-program* approach that enhances electrons into deep traps by implementing an additional program operation can alleviate this problem. However, the additional program operation increases the program latency thus degrading the system performance. Another approach is to adjust the read reference voltages. Since threshold voltages are shifted to lower values after fast detrapping, moving read reference voltages to lower values can obtain a better result.

4.1.4 Read disturb error. Read disturb errors occur when pages in a block are repeatedly read without any erase operations. In order to characterize the read disturb error in 3D CT NAND flash, we choose blocks at different P/E cycles, equally divide them into two groups (group A and group B), and for each group we perform the following operations: 1) erase and program an entire block with random data; 2) read a specific page (for group A we choose Page 3, Page 4 and Page 5, for group B we choose Page 279, Page 280 and Page 281 to research on the influence of location and types of pages on read disturb errors) for $2,000,000$ times and record data of the entire block every $100,000$ times to get the distributions of RBERs. Fig. 9 shows the variations of block RBERs with the increase of read disturb counts under different P/E cycles, and Fig. 10 plots the distributions of RBERs within the blocks after being read disturbed.

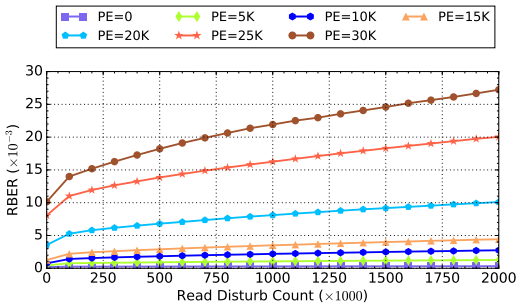


Figure 9: Variation of read disturb induced RBERs with different read disturb counts and P/E cycles.

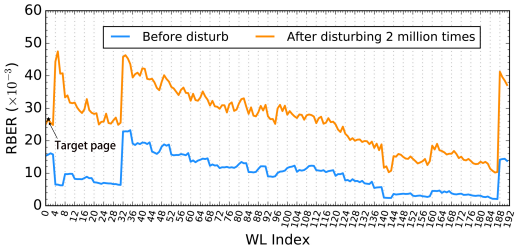
Observation 6. ***Read disturb induced RBERs grow with increased read disturb counts, with the earlier stage growing much faster.***

According to our tested results, reading a lower page, middle page and an upper page produce similar levels of read disturb errors, hence in Fig. 9 we only present REBRs variations when reading a middle page as a representative case. we can observe from the figure that with the increase of read disturb counts, the RBERs rise sharply within the first $10,000$ read disturb counts, and grows relatively slowly subsequently. Moreover, blocks under higher P/E cycles are more vulnerable to read disturb and exhibits higher increase of RBERs. These phenomena can be explained by the intrinsic cause of read disturb that a pass through voltage, V_{pass}, is applied to tiers not being read in the block. The continuous exertions of V_{pass} gradually shift the threshold voltages to higher values, resulting in the rise of RBERs. As more electrons are unintentionally charged into the storage layers and raise the threshold voltages, the opposite electric field against charging electrons becomes larger, resulting in the decrease of RBERs growing speed, as shown in Fig. 9. In addition, blocks under higher P/E cycles have more defects in the tunneling oxide, thus electrons are much easier to be injected into the storage layer through read disturb.

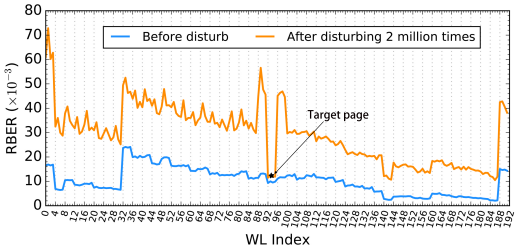
Observation 7. ***Read disturb exerts non-uniform influences on different tiers inside a block, with neighboring tiers suffering more seriously.***

Fig. 10a and Fig. 10b exhibit cases in which a page at the edge (Page 4) and a page in the middle (Page 280) of the block are selected as the source of read disturb, respectively. Since all types of pages generate similar results, Fig. 10 gives the general situations. We can observe from the figure that when a target page is repeatedly read, RBERs of the tier containing the target page experience only a slight increase, while RBERs of all the other tiers in the block rise dramatically. Moreover, tiers that are adjacent to the tier containing the target page suffer relatively more read disturb than the other tiers. This is due to the read mechanism that the pass voltage, V_{pass}, applied to tiers exclusive of the target page generates a *weak programming effect*, which can be accumulated and results in read disturb errors through repeated read operations. Nevertheless, the tier that contains the target page only needs to be applied read reference voltages, V_{ref}, which are lower than V_{pass}, thus RBERs of the tier increase slightly. In addition, the tier containing the target page and the adjacent tiers are weakly coupled during a read operation, resulting in higher RBERs in the adjacent tiers after 2 millions of read operations.

目标页的相邻层受到了相比其它层更为严重的读干扰，这是由于为了减少耦合效应， 目标页的相邻层被施加了稍高于Vpass电压的通过电压（VpassH），从而受到了更严重的弱编程效应造成的



(a) Reading a page in WL 1



(b) Reading a page in WL 93

Figure 10: Distribution of read disturb induced RBERs under the P/E cycle of 30,000.

Application. In some applications, read disturb errors generated by the intensive read of certain pages could be disastrous. A straightforward way to alleviate this problem is to implement a large read cache. However, for some embedded systems without a large RAM, it remains a serious problem. For such systems, a feasible approach is to detect read-hot pages and for each read-hot page, make several copies to different blocks. Through implementation of this approach, read-hot data are dispersed so that system reliability is improved, and meanwhile read access to these pages can also be accelerated.

4.1.5 Program disturb error. Due to the three-dimensional structure adopted in 3D NAND flash, program disturb exhibits more complicated spatial features, which are far different from those in planar NAND flash. As investigated in 3D FG NAND flash, program disturb can cause a serious reliability issue. Nevertheless, the distinct materials and string structure in 3D CT NAND flash lead to disparate characteristics of program disturb, compared with 3D FG NAND flash.

A criminal WL is the WL being programmed and the source of program disturb, and a victimized WL suffers program disturb from a criminal WL. In 3D NAND flash, there are three spatial relationships between a criminal and a victimized WL: 1) Y mode (the two WLs are in the same tier); 2) Z mode (the two WLs are in the same strings); and 3) YZ mode (the two WLs are in neither the same tier nor the same strings). In order to explore the characteristics of program disturb in 3D CT NAND flash, we 1) randomly choose several blocks; 2) repeatedly program and erase those blocks to different preset P/E cycles; 3) randomly program a WL (not the last WL in a tier) as the victimized WL in each block and place those blocks for 30 minutes to eliminate the influence caused by fast detrapping phenomenon; and 4) for each block, read the victimized WL, program the next 8 WLs of the victimized WL[3] (e.g., victimized WL index: n, criminal WL indexes: $n+1$, $n+2$, $\cdots$, $n+8$), and read the victimized WL after each program operation. Since in the tested 3D CT NAND, each tier contains 4 WLs, Y-, Z- and YZ-modes program disturb can be covered by the above scheme.

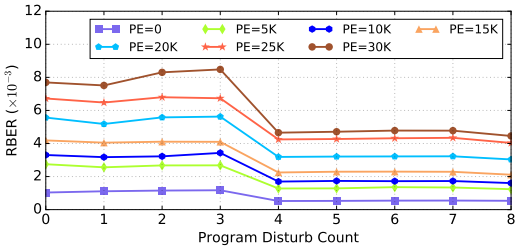


Figure 11: Variations of RBERs caused by program disturb with various P/E cycles and disturb counts.

Observation 8. Z-mode program disturb has the most significant impact on the reliability.

As shown in Fig. 11, the RBERs vary gradually when the program disturb is in Y mode or YZ mode (the program disturb counts are 1, 2, 3, 5, 6 and 7). But for Z-mode program disturb, the RBERs are strongly and weakly affected by the 4th and the 8th program disturb, respectively. When performing a program operation on a criminal WL, a victimized WL can be disturbed by weak programming effect and coupling effect. The coupling effect mainly exists between WLs in the same strings, sharply weakens with the increase of distance, and is much stronger than the programming effect. Thus, programming the WL $n+4$, which is the next WL in the same strings with the victimized WL, affects the reliability most, and the other program operations have weaker effects.

Just like read disturb, program disturb also unintentionally injects electrons into the storage layer. However, as shown in Fig. 11, the RBERs decrease after being disturbed by program operations. In order to investigate this phenomenon, we analyze and count the state shifts before and after program disturb, as shown in Fig. 12. Since more than 99.9% of the shifts occur among adjacent states, Fig. 12 only shows the adjacent state shifts. If a cell shifts from a lower state to a higher one, it is a positive state shift (e.g., $S_6 \rightarrow S_7$), otherwise, a negative state shift (e.g., $S_7 \rightarrow S_6$).

Observation 9. Program disturb can eliminate fast detrapping errors to some extent.

As mentioned in Section 4.1.3, electrons sitting in shallow traps can escape from the storage layer easily, leading to the result that threshold voltages shift to lower values, even negative state shifts. As a result, the injected electrons caused by program disturb may make the cells, which have shifted to lower states, back to the original states. As shown in Fig .12, the 4th program disturb dramatically reduces the negative state shift counts, and slightly drives up the positive state shift counts, resulting in the lower RBERs.

Application. WLs in 3D CT NAND flash are programmed in a fixed order and the effect of program disturb from a criminal WL to a victimized WL can be quantified. Moreover, program disturb and fast detrapping shift threshold voltages to opposite directions. Therefore, both program disturb and fast detrapping errors can be reduced by carefully controlling the program pulses so that a majority of the threshold voltage shifts can be counteracted, which has already been implemented in the design of flash chips.

4.2 Performance

The performance of NAND flash is critical for real-time applications, and may contain valuable information to evaluate the health conditions of NAND flash. In order to characterize the performance of 3D CT NAND flash, we employ random data to repeatedly program and erase fresh blocks and record the latencies of each erase, program and read operation for every 100 P/E cycles until they fail. Fig. 13 shows the latencies of erase (a), program (b) and read (c) operations within the lifetime. We further explore the performance variations inside each block by analyzing the program latency distribution among WLs, and the results are shown in Fig. 14.

Observation 10. The performances of erase and program operations vary predictably with the increase of P/E cycle.

As shown in Fig. 13a and Fig. 13b, with the continuous wear of NAND flash, the erase latency exhibits a ladder-shaped growth, forming several steps and fluctuates near the joint of each two steps, while the program latency declines gradually. These phenomena are due to the intrinsic mechanisms of program and erase operations. In 3D CT NAND flash, program and erase operations are achieved by Fowler-Nordheim (FN) tunneling effect [12], which charges and discharges electrons into and from the storage layer,

[3]The WL-order programming scheme is suggested by the manufacturers and employed in practice.

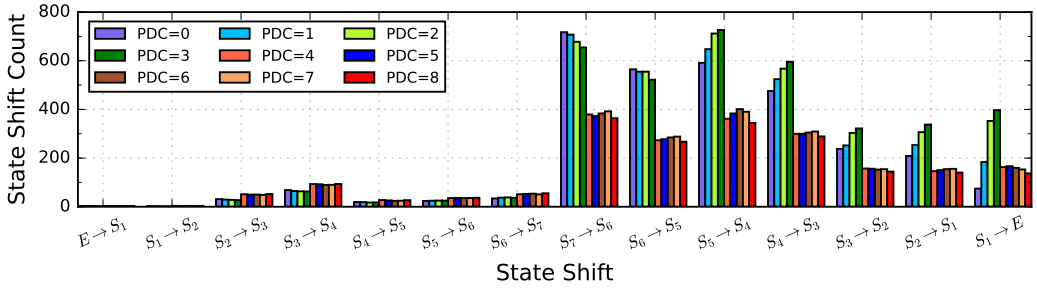
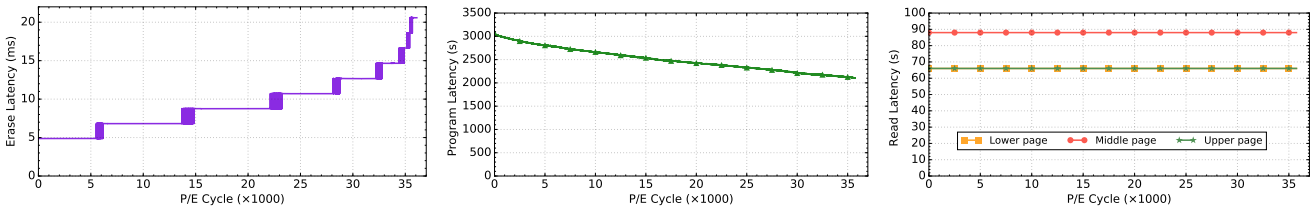


Figure 12: State shifts with different program disturb counts (PDC) under the P/E cycles of $30,000$.



(a) Erase latency **(b) Program latency** **(c) Read latency**

Figure 13: NAND flash operation performances over P/E cycles.

respectively. However, as electrons are repeatedly tunneled through the tunnel layer, defects, which unintentionally trap electrons and cause the Trap-Assisted Tunneling (TAT) effect, are formed. The TAT effect reduces the insulativity of the tunnel layer [10], making electrons easier to be tunneled through. For erase operations, although the TAT effect accelerates the detrapping process, electrons trapped in the defects form an opposite electric field against the erase pulse and make electrons in the storage layer harder to be tunneled out. Moreover, due to the unintentionally trapped electrons, the threshold voltages of the cells are shifted to higher values, further increasing the erase time. Therefore, with more defects accumulating in the tunnel layer, a larger opposite electric field is formed and the higher values the threshold voltages are shifted, thus more time is needed to erase the cells to a voltage beneath V_{ref1}. Considering the mechanism of erase operations, which is composed of continuous erase-verify processes, the harder electrons can be tunneled out, the more erase-verify processes are performed, thus the erase latency is increased in a ladder shape. For program operations, the TAT effect accelerates the programming process. In addition, as the threshold voltage of a cell is shifted to a higher value, fewer electrons are needed to be programmed into the storage layer, hence reducing the program time. Since the step length of the pulse of program operations are much shorter than erase operations, the decreasing of the program latency exhibits much shorter steps, thus showing a more smooth curve.

Observation 11. ***Read performance remains constant during the entire lifetime of NAND flash.***

As shown in Fig. 13c, the read latencies of all types of pages remain unchanged regardless of the P/E cycle, with reading lower and upper pages taking 66μs, and middle pages costing relatively longer (88μs). Unlike erase or program operations which involve moving electrons through the tunnel layer, the read operation is not affected by the wear degree of NAND flash, thus read latency stays constant as P/E cycle increases. The differences of read latencies among different types of pages are due to the encoding rule of 3D CT NAND flash, as shown in Fig. 1. When reading a lower/upper page, only two read reference voltages (lower page→$\{V_{ref3}, V_{ref7}\}$, upper page→$\{V_{ref1}, V_{ref5}\}$) need to be applied in sequence, while reading a middle page involves applying three read reference voltages ($\{V_{ref2}, V_{ref4}, V_{ref6}\}$), resulting in a longer read time.

Observation 12. ***Program performance exhibits segmented distribution among WLs.***

As shown in Fig. 14, the program latencies are not evenly distributed among all WLs in a block. In fact, the first 32 WLs have much higher program latencies, and the program latencies of the last 32 WLs are slightly lower. As mentioned in Sec. 4.1.2, different voltages are applied to different groups of WLs. Since cells in Segment 1 have the smallest size due to the unsatisfactory high aspect ratio etch procedure, they are applied with the lowest program voltage, thus more time is needed for tunneling the same amount of electrons into the storage layer. Therefore, the average programming speed of Segment 1 is the highest among the three segments. On the contrary, cells in Segment 3 have the largest size, hence exhibiting the highest programming speed.

Application. In storage systems, the loss of data has always been a serious problem. In order to overcome this problem, traditional techniques employ mechanisms such as *redundant arrays of independent disks* (RAID) to store redundant data,

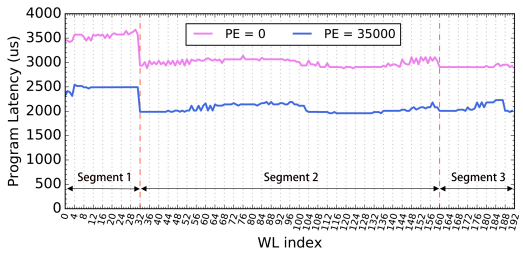


Figure 14: Program latency distributions among WLs under different P/E cycles

trading storage capacity for data reliability. However, with the maturing of machine learning algorithms and the development of hardware resources, this problem can be mitigated from another perspective: *lifetime prediction*. As we can observe from Fig. 13, performances such as erase and program latencies exhibit regular variations throughout the entire life cycle of 3D CT NAND flash, thus can be used as important factors to predict its lifetime.

5 RELATED WORK

Since NAND flash-based storage devices gradually take the place of HDDs, the focus of both academia and industry is on NAND flash, especially the characteristics of NAND flash. During the past few years, plenty of research works have experimentally characterized planar NAND flash [3–9, 11, 13, 17]. Cai et al. conducted the major series of investigations on various reliability issues, including error patterns [5], threshold voltage distribution [6], three main sources of errors (retention [7], read disturb [8] and program disturb [9]), and Vulnerabilities in programming [4]. These studies increase designers knowledge about the inherent features of NAND flash, and motivate them to design high efficient NAND flash management algorithms based on the features.

Due to the scaling limitation of planar NAND flash, the 3D structure has been proposed and is attracting more and more attention. Most research works about 3D NAND flash are at simulation levels or integrated circuit levels, which usually use self-designed structures and custom-fabricated chips [1, 16, 19]. Xiong et al comprehensively evaluated the commercial 3D FG NAND flash product for the first time [20]. Based on their observations, Zhu et al built a read disturb error model and proposed a location-aware redistribution method (ALARM) [21]. ALARM utilizes the intrinsic features of read disturb errors in 3D FG NAND flash and redistributes read-hot data to locations inducing less read disturb errors to improve the reliability. However, in 3D era, the CT technology has replaced the FG technology as the mainstream, and the characteristics of commercial 3D CT NAND flash products have not been comprehensively investigated. The lack of research in this area makes NAND flash management algorithms inefficient since those algorithms cannot utilize the characteristics of 3D CT NAND flash. To our knowledge, this is the first study comprehensively characterizing advanced commercial 3D CT NAND flash products.

6 CONCLUSION

In this paper, we implement comprehensive characterizations of advanced 3D CT NAND flash from reliability and performance aspects. According to the experimental data measured on the ARM- and FPGA-based platform, we make multiple distinct observations and give detailed analyses. 3D CT NAND flash exhibits outstanding endurance, diverse degradation speed among pages and blocks, fast detrapping phenomenon, cross-tier variations of errors, and slight program disturb, etc. Based on those observations, approaches to optimize flash management algorithms of 3D CT NAND flash in real applications are briefly discussed. We believe this work can give researchers and designers deeper understandings of the characteristics of 3D CT NAND flash and improve the efficiency of NAND flash management algorithms.

REFERENCES

[1] S. Aritome, Y. Noh, and H. Yoo et al. 2013. Advanced DC-SF cell technology for 3D NAND flash. *IEEE TED* 60, 4 (2013), 1327–1333.
[2] D. A. Baglee. 1984. Characteristics & reliability of 100A oxides. In *Proc. of IRPS*. 152–155.
[3] S. Boboila and P. Desnoyers. 2010. Write Endurance in Flash Drives: Measurements and Analysis. In *Proc. of FAST*. 115–128.
[4] Y. Cai, S. Ghose, and Y. Luo et al. 2017. Vulnerabilities in MLC NAND Flash Memory Programming: Experimental Analysis, Exploits, and Mitigation Techniques. In *Proc. of HPCA*. 49–60.
[5] Y. Cai, E. F. Haratsch, and O. Mutlu et al. 2012. Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis. In *Proc. of DATE*. 521–526.
[6] Y. Cai, E. F. Haratsch, and O. Mutlu et al. 2013. Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling. In *Proc. of DATE*. 1285–1290.
[7] Y. Cai, Y. Luo, and E. F. Haratsch et al. 2015. Data retention in MLC NAND flash memory: Characterization, optimization, and recovery. In *Proceedings of HPCA*. 551–563.
[8] Y. Cai, Y. Luo, and S. Ghose et al. 2015. Read disturb errors in MLC NAND flash memory: Characterization, mitigation, and recovery. In *Proc. of DSN*. 438–449.
[9] Y. Cai, O. Mutlu, and E. F. Haratsch et al. 2013. Program interference in MLC NAND flash memory: Characterization, modeling, and mitigation. In *Proc. of ICCD*. 123–130.
[10] R. Degraeve, F. Schuler, and B. Kaczer et al. 2004. Analytical percolation model for predicting anomalous charge loss in flash memories. *IEEE TED* 51, 9 (2004), 1392–1400.
[11] P. Desnoyers. 2010. Empirical evaluation of NAND flash memory performance. *ACM SIGOPS Operating Syst. Rev.* 44, 1 (2010), 50–54.
[12] R. H. Fowler and L. Nordheim. 1928. Electron emission in intense electric fields. 119, 781 (1928), 173–181.
[13] L. M. Grupp, A. M. Caulfield, and J. Coburn et al. 2009. Characterizing flash memory: Anomalies, observations, and applications. In *Proc. of MICRO*. 24–33.
[14] K. Kim. 2008. Future memory technology: Challenges and opportunities. In *Proc. of VLSI-TSA*. 5–9.
[15] C.-H. Lee, J. Choi, and C. Kang et al. 2006. Multi-level NAND flash memory with 63 nm-node TANOS (Si-Oxide-SiN-Al2O3-TaN) cell structure. In *Tech. Dig. of VLSI Technol.* 21–22.
[16] R. Micheloni. 2016. *3D Flash Memories*. Springer.
[17] R. Micheloni, L. Crippa, and A. Marelli. 2010. *Inside NAND Flash Memories*. Springer Science & Business Media.
[18] K.-T. Park, M. K., and D. Kim et al. 2008. A zeroing cell-to-cell interference page architecture with temporary LSB storing and parallel MSB program scheme for MLC NAND flash memories. *IEEE JSSC* 43, 4 (2008), 919–928.
[19] J. Wu, D. Han, and W. Yang et al. 2017. Comprehensive investigations on charge diffusion physics in SiN-based 3D NAND flash memory through systematical Ab initio calculations. In *Proc. of IEDM*. 1–4.
[20] Q. Xiong, F. Wu, and Z. Lu et al. 2017. Characterizing 3D floating gate NAND flash. In *Proc. of SigMetrics*. 32–33.
[21] Y. Zhu, F. Wu, and Q. Xiong et al. 2017. ALARM: A Location-Aware Redistribution Method to Improve 3D FG NAND Flash Reliability. In *Proc. of NAS*. 1–10.