Characterizing 3D Floating Gate NAND Flash: Observations, Analyses, and Implications

QIN XIONG and FEI WU, Huazhong University of Science and Technology
ZHONGHAI LU, KTH Royal Institute of Technology
YUE ZHU and YOU ZHOU, Huazhong University of Science and Technology
YIBING CHU, Renice Technology Co. Limited
CHANGSHENG XIE, Huazhong University of Science and Technology
PING HUANG, Temple University

As both NAND flash memory manufacturers and users are turning their attentions from planar architecture towards three-dimensional (3D) architecture, it becomes critical and urgent to understand the characteristics of 3D NAND flash memory. These characteristics, especially those different from planar NAND flash, can significantly affect design choices of flash management techniques. In this article, we present a characterization study on the state-of-the-art 3D floating gate (FG) NAND flash memory through comprehensive experiments on an FPGA-based 3D NAND flash evaluation platform. We make distinct observations on its performance and reliability, such as operation latencies and various error patterns, followed by careful analyses from physical and circuit-level perspectives. Although 3D FG NAND flash provides much higher storage densities than planar NAND flash, it faces new performance challenges of garbage collection overhead and program performance variations and more complicated reliability issues due to, e.g., distinct location dependence and value dependence of errors. We also summarize the differences between 3D FG NAND flash and planar NAND flash and discuss implications on the designs of NAND flash management techniques brought by the architecture innovation. We believe that our work will facilitate developing novel 3D FG NAND flash-oriented designs to achieve better performance and reliability.

CCS Concepts: • **General and reference → Reliability**; **Empirical studies**; *Performance*; • **Hardware → Non-volatile memory**; **Memory and dense storage**;

Additional Key Words and Phrases: 3D floating gate NAND flash, MLC, error pattern

This research is sponsored by the National Natural Science Foundation of China under Grant No. 61300047, No. 61472152, No. 61572209; Wuhan Science and Technology Project No. 2017010201010108; Shenzhen Basic Research Project No. ICYJ20170307160135308; 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.

Authors' addresses: Q. Xiong, Y. Zhu, and Y. Zhou, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan 430074, China; emails: qinxiong@kth.se, {yuezhu, zhouyou}@hust.edu.cn; F. Wu (Corresponding author) and C. Xie, Wuhan National Laboratory for Optoelectronics and the Key Laboratory of Data Storage System, Ministry of Education of China, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan 430074, China; emails: {wufei, cs_xie}@hust.edu.cn; Z. Lu, School of Information and Communication Technology, KTH Royal Institute of Technology, 16 Kistagangen, Kista 16440, Sweden; email: zhonghai@kth.se; Y. Chu, Renice Technology Co. Limited, 15 Keyuan Road, Shenzhen 518057, China; email: eric@renice-tech.com; P. Huang, Department of Computer Information Sciences, Temple University, 1925 N. 12th Street, Philadelphia, PA 19122, USA; email: templestorager@temple.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 ACM 1553-3077/2018/04-ART16 $15.00

https://doi.org/10.1145/3162616

ACM Reference format:
Qin Xiong, Fei Wu, Zhonghai Lu, Yue Zhu, You Zhou, Yibing Chu, Changsheng Xie, and Ping Huang. 2018. Characterizing 3D Floating Gate NAND Flash: Observations, Analyses, and Implications. *ACM Trans. Storage* 14, 2, Article 16 (April 2018), 31 pages.
https://doi.org/10.1145/3162616

1 INTRODUCTION

Over the past decades, NAND flash memory has been widely used as a popular storage medium in various systems ranging from embedded systems to high-performance computing servers due to its fast access, low energy consumption, small size, and high shock resistance compared with spinning disks. With the development of semiconductor process, the gap of the price-per-bit between NAND flash-based storage and magnetic storage is narrowing. However, when the feature size shrinks to sub-30nm, flash scaling becomes particularly challenging because of technical complexity and the electrons that can be retained by a cell decrease to single digits [20]. Moreover, as more bits are stored in a cell, the voltage margin between two adjacent states is exponentially reduced. As a result, the interferences among cells are poised to be worsening, since the distances among them become shorter [27]. However, on one hand, high reliability (including endurance) is one of the most important requirements for storage systems. On the other hand, the demand of storage capacity is much more than ever before due to data explosion. The digital universe reached 12ZB in 2016, and International Data Corporation (IDC) predicts that it will rise to 40ZB by 2020 [17]. Unfortunately, the traditional planar NAND flash memory technology cannot adequately satisfy the requirements for reliability and capacity, which results in the invention of innovative flash architecture.

Three-dimensional (3D) NAND flash represents a promising opportunity to overcome the limitations of planar devices, since it allows flash storage to continue aligning with Moore's Law, bringing a significant improvement in density and lowering the cost of NAND flash while expected to provide higher reliability. Three-dimensional NAND flash enables capacity scaling in the Z-axis direction rather than X-axis and Y-axis directions, thus vendors can roll back the semiconductor process technology node several generations to improve reliability. The current feature sizes of 3D NAND flash are larger than 40nm, and there are two types of NAND flash cells in industry, *floating gate* (FG) [25] and *charge trap* (CT) [21]. The main difference between the two structures is that an FG cell uses a floating gate made of doped polycrystalline silicon to store electrons versus a silicon nitride film in a CT structure. In the era of planar NAND flash, several vendors have tried to adopt the CT structure without much success and the FG structure has prevailed in the market, since FG cells are much more reliable than CT cells. For 3D NAND flash, with the significant improvement of materials in CT cells, both FG and CT structures are employed by different vendors (e.g., Intel-Micron produces the 3D FG NAND flash products and Samsung adopts the CT technology in its VNAND). Both techniques have strengths and weaknesses: 3D FG NAND flash has (1) lower charge spreading (fewer read errors), (2) more stable charges (better data retention), and (3) direct connection between p-well and channel poly (allowing bulk erase), while 3D CT NAND flash has (1) smaller cell size and 3D pillar (better scalability) and (2) weak coupling effect (less interference among cells) [24]. Even though FG technique in planar NAND flash has been studied for years, due to the structural differences, the characteristics of 3D FG NAND flash are different from those of planar NAND flash and should be further investigated. Therefore, *in this article, our research focuses on 3D FG NAND flash*, and we expect that studying 3D CT-type NAND flash is an important area of future work.

As 3D NAND flash memory is increasingly adopted in storage systems, it is critical and urgent to have a good understanding of its characteristics. Flash memory exhibits unique physical features and error patterns than spinning disks, thus several management techniques are needed to build a flash-based storage system. We give a few examples. *Page mapping* dynamically allocates free flash pages to serve write requests, since flash memory performs out-of-place updates [35]. *Garbage collection* (GC) reclaims invalidated pages in victim flash blocks by first migrating valid pages in the blocks and then erasing the blocks. *Wear leveling* tries to ensure that all flash blocks wear out at a similar time. *Error correction codes* (ECCs) use additional parity bits to detect and correct bit errors of flash memory. For example, low-density parity-check (LDPC) codes, which are widely used in modern flash memory, perform soft-decision decoding, so their error correction capability depends on the precision of input soft information [34]. *Bad block management* monitors and predicts the health status of flash blocks, and shields bad blocks from usage. The efficiency of these techniques is a critical factor that determines the performance and lifetime of flash-based storage systems. Since these techniques are closely related to the characteristics of underlying flash memory, their efficiency highly depends on the understanding of the characteristics.

Unfortunately, only a few prior works have researched the characteristics of 3D NAND flash. Moreover, most of them are simulation-based or integrated circuit-level experiment-based analyses. In this article, we conduct extensive experimental evaluations on the state-of-the-art 3D FG *multi-level cell* (MLC) NAND flash chips and perform a comprehensive data analysis. Our goal is to obtain a deep understanding of the properties of 3D FG NAND flash, and we hope to motivate both academia and industry to carry out more researches in this area and design more effective storage devices via leveraging our obtained insights.

The main contributions of this article are as follows:

- We *comprehensively characterize the performance and reliability of the state-of-the-art 3D FG NAND flash* through experimental measurements and observations. To the best of our knowledge, this is the first article to empirically analyze the characteristics of 3D FG NAND flash.
- We *analyze the physical and underlying circuit mechanisms of our observations.* Since those mechanisms are inherent to 3D FG NAND flash, our observations can be applied to future generations.
- We *discuss the implications of our observations on the designs of 3D FG NAND flash management techniques.*

The rest of this article is organized as follows. Related works are given in Section 2. Section 3 introduces the background and the preliminaries. Our experimental setup is described in Section 4. Section 5 shows evaluation methodologies and results and then describes those observations and discusses their implications. The last section concludes this article.

2 RELATED WORK

In the past few years, a mass of literature (e.g., [3–8, 12, 15, 25]) has focused on experimentally characterizing planar NAND flash. A representative series of studies have been conducted by Cai et al., including error patterns in 30- to 40nm MLC NAND [5], threshold voltage distribution [4], and main sources of errors in 20- to 24nm MLC NAND flash: program disturbs [8], read disturbs [6], and retention [7].

Nowadays, since planar NAND flash has reached its scaling limit, and 3D technology is emerging, both academia and industry are turning their attention to 3D NAND flash. However, prior research has been based on simulations or integrated circuit-level experiments, which study internal electrical characteristics by using self-designed structures and custom-fabricated devices, respectively. In Reference [28], Seo et al. proposed a novel separated-sidewall control gate (S-SCG) structure and their simulation results showed that the structure could be used to construct terabit 3D NAND flash array with highly-reliable multilevel cells. Based on an analytic model, Yoo et al. found that neighboring cells with high V_{th} increase the V_{th} of a selected cell and proposed a new read operation scheme with variable pass voltage (V_{pass}) for dual control gate with surrounding floating gate (DC-SF) NAND flash cells [33]. The simulation results showed that by adopting the scheme, more stable read operations could be achieved. In Reference [24], a comprehensive introduction of reliability in 3D NAND flash was given. Nevertheless, the introduction focused more on the mechanisms of factors threatening reliability and the majority of data used to illustrate the challenges of reliability was from simulations of 3D NAND flash and experiments with planar NAND flash. In References [30] and [1], the experimental results (only a small fraction was about reliability) were measured based on custom-fabricated DC-SF NAND cells before 2014. However, in 3D FG NAND flash that was first released in the second half of 2015, the conventional FG structure was employed [26], and what academia and industry devoted to system-level design really care about is the characteristics of performance and reliability of 3D NAND flash products. To our knowledge, this article is the first work that experimentally and comprehensively measures, observes, and analyzes the performance and reliability of the state-of-the-art 3D FG NAND flash products.

In our early work [32], we showed a few observations of 3D FG NAND flash, briefly analyzed them, and gave a simple introduction of characteristic-based NAND flash management policies. However, due to the incomplete observations and the oversimplified analyses and implications, the guidance that the early work can give to 3D FG NAND flash-oriented designs is limited. In this work, we widen and deepen the content in Reference [32] by making several new contributions: (1) This article describes 22 observations about performance and reliability, of which 8 observations are extended from the 5 observations in Reference [32] and the other 14 observations are newly added to comprehensively characterize the state-of-the-art 3D FG NAND flash memory; (2) analyzing those observations in detail from physical and circuit-level perspectives (the analyses of the original observations are extended); (3) discussing the implications on designs of several flash management techniques based on the corresponding types of observations (performance, endurance, program disturb error, read disturb error, and retention error); and (4) summarizing the differences between 3D FG NAND flash and planar NAND flash and their implications for different flash management algorithms.

3 BACKGROUND AND PRELIMINARIES

3.1 3D Floating Gate NAND Flash Structure

A 3D FG NAND flash chip consists of thousands of blocks and each block is a 3D organized array of FG transistors. In an FG cell, a *vertical channel* is completely surrounded by a *Floating Gate* and a *Control Gate*; *Inter-Ploy Dielectric* and *Tunnel Oxide* complete the cell structure [26]. Figure 1 illustrates a bird's-eye view and the circuit diagrams of the side views of a 3D FG NAND flash array. A *word line* (WL) is typically composed of hundreds of thousands of cells that are in a row, and dozens of WLs form a tier, in which all the Control Gates of cells are short-circuited together. Vertical stacked transistors, which form a *string*, are connected to a *Bitline* on the top and a *Source Line* on the bottom through a *Bitline Selector* (BLS) and a *Source Line Selector* (SLS), respectively. Since BLS and SLS are used to control the connection and do not store information, they are fabricated as standard transistors (i.e., without a floating gate). The gates of BLSs and SLSs that control the connection of a WL are connected together and form a BLS line and an SLS line, respectively. Since a BLS line needs to select a single WL out of a tier, BLS lines cannot be

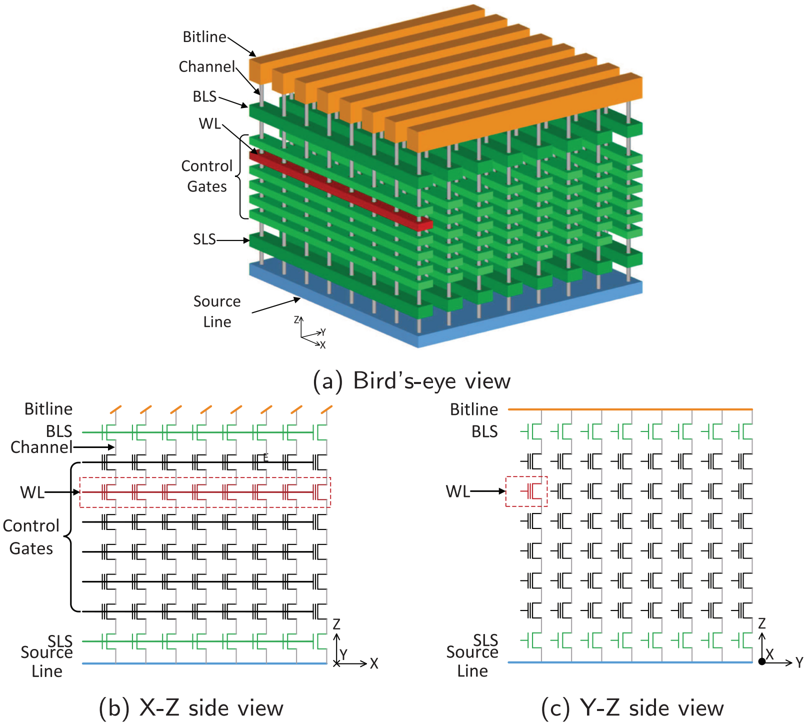


Fig. 1. Bird's-eye view and circuit diagrams of the side views of 3D FG NAND flash array. To clearly illustrate the structure, the Control Gates of WLs in the same tier are not short-circuited.

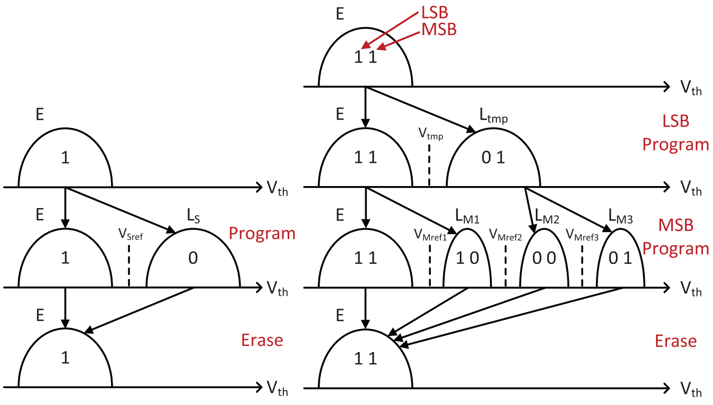
short-circuited. In contrast, all the SLS lines of a block are connected together as SLS lines need to be selective only when an erase operation is performed on that block.

3.2 Basic Operations

There are three basic operations for NAND flash memory: *erase*, *program*, and *read*.

Erase: An erase operation resets the data of all cells in a block simultaneously, which means that erase operations are performed at block-level. By applying a high positive voltage (e.g., 20V) between Channels and Control Gates of all cells in a selected block, the electrons held by all floating gates are tunneling out through Fowler-Nordheim (FN) tunneling mechanism [13]. All cells in the block are expected to return to the E state after an erase operation, as shown in Figure 2.

Program: Program operations change the data stored in the selected cells by injecting charges to those floating gates through the FN mechanism. When performing a program operation, only the BLS line corresponding to the programmed WL is turned on to select the WL, and the SLS line and the other BLS lines remain OFF state; the voltage of a bitline is either 0V or V_{cc} according to the value of the bit expected to be stored in the cell; the programming voltage V_{pgm} is applied to the target WL to charge the floating gates and the WLs on the other tiers are configured to the inhibited voltage V_{inh}. The V_{pgm} starts at an initial voltage and is increased by a small increment V_{step} for each program pulse, which increases the number of charges in the floating gate. Consequently, the threshold voltage V_{th} is shifted, and with more electrons in a floating gate, the V_{th} is higher. After each program pulse, the data are read out, and if a cell reaches its target V_{th} level, then the corresponding bitline is set to V_{cc} to inhibit the following program pulse. This process finishes



(a) SLC (b) MLC

Fig. 2. Program and erase operations of SLC and MLC.

when either all cells are programmed to the target levels or the number of program pulse reaches its limitation. This programming scheme is called *Incremental Step Pulse Programming* (ISPP) [29]. For *single-level cell* (SLC) flash, each WL stores only one logical page and for MLC flash, each cell can store two bits: a *least significant bit* (LSB) and a *most significant bit* (MSB), as shown in Figure 2. All the LSBs and the MSBs in a WL form a *lower page* and an *upper page*, respectively. Floating gate NAND flash is programmed at page-level: Program operations are carried out in two pages separately in terms of time and in lower page, upper page order. A cell starts to be programmed at the E state and moves into a temporary state (L_{tmp}) if the LSB is 0; otherwise, it remains in the E state. While programming the MSB, the LSB is firstly read out and the final state is then determined based on the combined value of these two bits. Specifically, the states E, L_{M1}, L_{M2} and L_{M3} correspond to the values 11, 10, 00, and 01, respectively. *In this article, we use the format AB to represent the two bits in an MLC cell, where A denotes the LSB and B indicates the MSB.*

Read: When reading a page, the SLS line and the BLS line corresponding to the target WL are in ON state, and the other BLS lines are turned off; all the tiers except the one that contains the target WL are biased at V_{pass}, so that all cells in these tiers act as pass transistors, regardless of their threshold voltages. A read reference voltage V_{ref} is applied to the target WL one or more times and by sensing the current, the cell states can be detected. If a cell is ON, then the V_{th} is lower than the V_{ref}; otherwise, it is higher than the V_{ref}. For SLC pages and lower pages in MLC mode, only one V_{ref} needs to be applied. To read an SLC page, V_{Sref} is used as the read reference voltage: If $V_{th} < V_{Sref}$, then the corresponding bit value is 1 and if not, the bit stores 0. To read a lower page, V_{Mref2} is applied: If $V_{th} < V_{Mref2}$, then the cell is in the E or the L_{M1} state and the LSB value is 1; otherwise, the cell is in the L_{M2} or the L_{M3} state meaning that it holds 0 in the LSB. To read an upper page, two read reference voltages, V_{Mref1} and V_{Mref3}, should be applied in turn. If the V_{th} is lower than V_{Mref1} or higher than V_{Mref3}, then the cell stays in either the E or the L_{M3} state, holding 1 in the MSB. Otherwise, if the V_{th} is between V_{Mref1} and V_{Mref3}, the cell is in the L_{M1} or the L_{M2} state, indicating that it stores an MSB value of 0.

3.3 Metrics

We now briefly introduce the main metrics used to describe the performance and reliability of NAND flash.

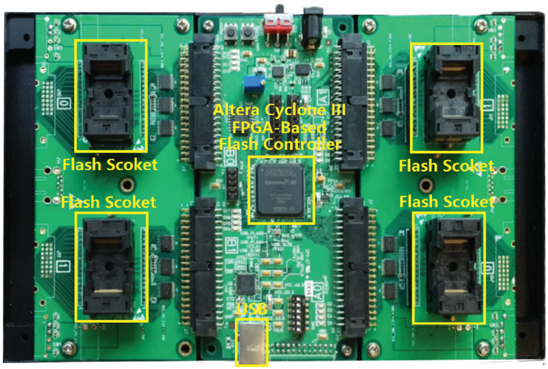


Fig. 3. 3D NAND flash evaluation platform.

Erase Latency: The time from the beginning of an erase operation until returning success or failure.

Program Latency: The time consumed for moving data from data register to the target page.

Read Latency: The time consumed for moving data from the target page to the data register.

Endurance: The number of *program/erase (P/E) cycles* that the flash can withstand before a failure occurs.

Program Disturb Error: When programming a WL, the other WLs in the same block are influenced by parasitic capacitance or nonzero voltages applied to control gates.

Read Disturb Error: When reading a WL, V_{pass} is applied to other WLs in the same block, resulting in a *weak programming* effect. Since a WL can be read unlimited times, read disturbs may introduce non-negligible errors.

Retention Error: Ideally, NAND flash can store data for a very long period of time (e.g., 5 years). However, V_{th} is not a constant over time due to charge loss. When V_{th} shifts across a read reference voltage, a retention error occurs.

In the rest of this article, we measure, observe, and analyze these features of 3D FG NAND flash in order.

4 EVALUATION SETUP

Evaluation Platform. We have built an FPGA-based NAND flash testing platform that gives us the ability to directly control the pins of NAND flash devices, as shown in Figure 3. The testing board has four sockets to hold tested flash chips and is connected to a host via a USB port. A custom flash controller is implemented in Altera Cyclone III FPGA EP3C55F484C7N. It receives commands and patterned data that are used to program flash memory from host and returns results (e.g., latencies) and data read from NAND flash chips. The controller supports timing measurements in a precision of $1\mu s$ through monitoring the ready/busy (R/$\bar{B}$) signal, which excludes the time consumed by transferring command and data to or from chips.

High-Low Temperature Test Chamber. To measure errors after a long retention age (e.g., 5 years), we use a high-low temperature test chamber to bake flash chips to accelerate retention error tests. According to Arrhenius Law [2, 7], we can calculate baking time t_{bake} (under temperature T_{bake}) corresponding to retention ages t_{room} (under room temperature, 25°C) used in our experiments, as shown in Table 1. Please note that although the temperature-accelerated test method of retention errors is widely accepted and applied in both academia and industry, due to the abnormal behaviors of retention loss of NAND flash [23], the test results of this method *cannot exactly represent the retention characteristics under room temperature.*

Table 1. Accelerated Actual Retention Ages

t_{room}	1 month	1 year	5 years
t_{bake} (T_{bake})	2.64h (70°C)	4.14h (90°C)	8.08h (100°C)

Table 2. Parameters of NAND Flash Chips[1]

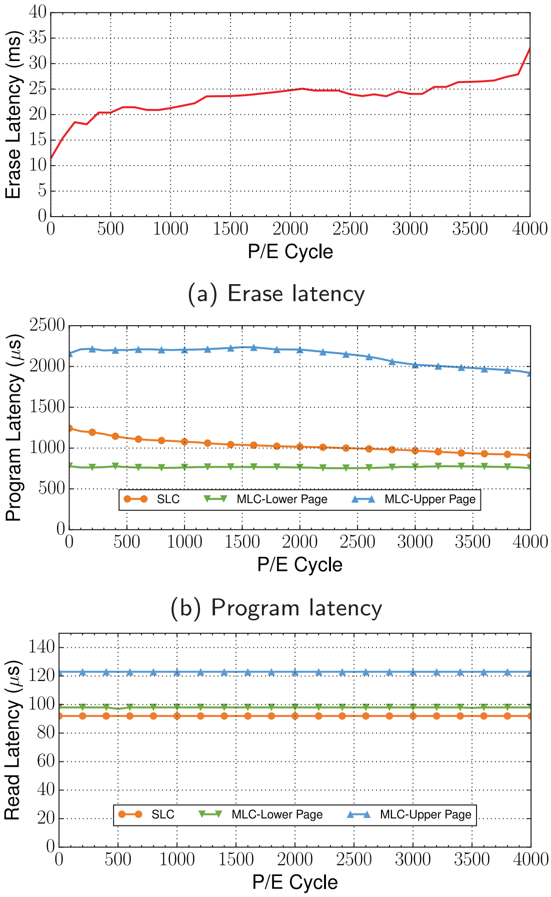
Parameters	Value
Capacity	256Gb (32GB)
Chip size	1 die/chip×4 planes/die
Plane size	548 blocks/plane
Block size	1024 pages/block
Page size	(16384 + 2208) bytes/page
Number of tiers	32 active tiers+6 dummy tiers
Read latency (typical)	75μs
Program latency (typical)	1050μs
Erase latency (typical)	10ms

1. All these parameters are from the datasheet.

3D FG NAND Flash. For our experiments, we choose the state-of-the-art Intel-Micron 32-tier 3D FG MLC NAND flash chips manufactured in 2016[1]. The main parameters are listed in Table 2. The NAND array has 38 tiers in total, consisting of 32 active tiers and 6 dummy tiers (3 at the top and 3 at the bottom). The dummy tiers are adopted to shield the Edge Disturbs [22] caused by hot carriers that are generated by the potential drop between the program inhibited string and the select transistor. Actually, 34 tiers, including 32 active tiers and 2 dummy tiers (1 in the top three dummy tiers and 1 in the bottom three), are used to store data. For these 34 tiers, the two dummy tiers and the first and the last active tiers work in SLC mode, and the middle 30 tiers are in MLC mode. In each tier, there are 16 WLs. From the bottom to the top, we label the 34 tiers as 0 to 33 and the 544 WLs as 0 to 543. The 544 WLs are firstly numbered sequentially within a tier and then within the next above tier. We also use Tier x WL n to locate the nth WL in Tier x, i.e., WL $(x * 16 + n)$. All the nth WLs in those tiers constitute a *vertical plane n* (VPn) and, consequently, there are 16 VPs in a block. In summary, each block contains 34 data tiers, 16 VPs, 544 data WLs, and 1024 pages.

To reduce disturbance during program operations, manufacturers recommend programming a block in *page order* and two pages of an MLC WL are not labeled contiguously. Both lower and upper pages are programmed from WL 32 to WL 511 (others are in SLC mode), but the upper page in WL n is programmed just before programming the lower page in WL $n + 31$. There are dozens of program operations between programming the lower page and the upper page in the same WL. In contrast, *WL order programming* means program operations are performed from WL 32 to WL 511 in turn and both the lower and the upper pages of a WL are programmed before operating the next WL. Note that whatever the order is employed, in a WL, the lower page MUST be programmed before the upper page.

[1] Until now, *only* Intel-Micron has released 3D FG NAND flash products. The chips we characterize in this article are of the first generation. Although Intel-Micron is currently sending the engineering samples of the second generation to very limited manufacturers of storage devices, it is still confidential and small-scale manufacturing is scheduled to begin in the second half of 2017. Hence, we believe the first generation is still the state-of-the-art 3D FG NAND flash product in the market now.



(a) Erase latency

(b) Program latency

(c) Read latency

Fig. 4. Performance over P/E cycles.

5 EVALUATION RESULTS

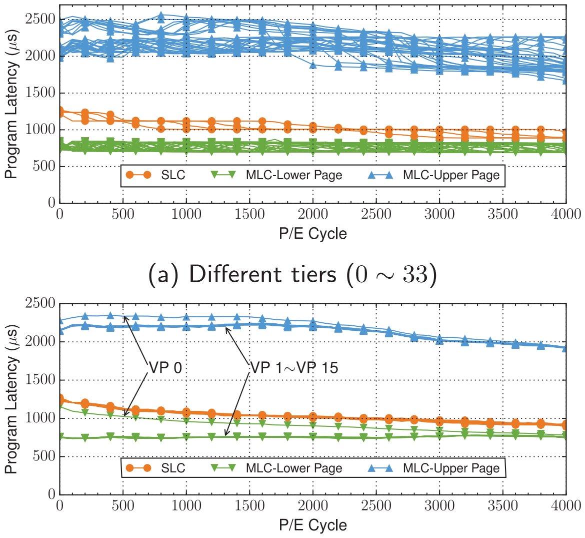
5.1 Performance

We measure and record a set of latencies of all basic operations every 100 P/E cycles from fresh blocks until they fail. Since in most cases, blocks fail when P/E cycles are between 4,000 and 6,000 (the details are described in Section 5.2.1), we report typical experimental data with P/E cycles from 0 to 4,000. Figure 4 shows the latencies of erase, program, and read operations as functions of P/E cycles. Only the performance of erase operations in the very early stage of flash's lifetime reaches the typical value (10ms); the lowest average values of program and read operations are still higher than those typical values (1311μs vs. 1050μs and 88μs vs. 75μs, respectively).

Observation P1. Erase and program latencies of SLC pages and upper pages in MLC mode vary predictably as P/E cycles increase. As shown in Figure 4, erase performance decreases dramatically as the devices wear out, especially during the first 300 and the last 100 P/E cycles, resulting in nearly 200% slower erase operations over the lifetime. Program operations show much fewer variations: the latencies of SLC pages and upper pages in MLC mode only decrease by about 27% and 11%, respectively, and the performance of lower pages fluctuated in a range of 3%. These phenomena are caused by the intrinsic property of FG cells that more defects on the tunnel oxide are accumulated as P/E cycles increase. During each P/E cycle, electrons are charged to and discharged from the floating gate through the tunnel oxide by applying a strong electric field between the floating gate and the substrate. Each time electrons tunnel through the tunnel oxide, defects, which trap electrons, may be generated, resulting in the rise of the threshold voltage in E state and the *trap-assisted tunneling* (TAT). When there are numerous trapped electrons, they can form paths to conduct electrons to and from the floating gate. The undesired TAT effect decreases the insulation of the tunnel oxide [11] and has a positive correlation with the number of trapped charges. For erase operations, the trapped electrons apply an opposite electric field against the erase electric field, which makes the electrons harder to tunnel from the floating gate, affecting the latency more than the TAT effect. For program operations of SLC pages and lower pages, the initial states are all the E state, hence both the increases of V_E and the TAT effect shorten the latencies. However, fewer electrons are needed to raise V_{th} from V_E to V_{tmp} and larger V_{step} is adopted when programming lower pages, the effects of V_E and TAT on the latency of lower pages are much weaker than those on the latency of SLC pages. These are also the reasons for the fact that a lower page is programmed faster than an SLC page. For upper pages, they are programmed from the E or the L_{tmp} state and programming it from L_{tmp} to L_{M3} is most time-consuming, thus the latency is influenced by the TAT effect and not by V_E. Therefore, over the lifetime of a NAND flash chip, (1) a fresh chip contains some defects caused in manufacture and those defects trap charges in the early stage, which leads to a rapid rise in V_E (during the first few hundreds of P/E cycles, erase latency jumps, while the program latency of SLC pages, lower pages and upper pages obviously declines, slightly decreases and stays stable, respectively); (2) in the middle stage, more defects are generated, resulting in slow climbs of the number of trapped charges and the TAT-based *Stress Induced Leakage Current* (SILC) (erase latency and the program latency of SLC pages grows and falls steadily, respectively, the program latency of lower pages fluctuates mildly, and that of upper pages gradually begins to slide); and (3) in the end, the tunnel oxide wears out very quickly, giving rise to spiralling trapped electrons and SILC (erase latency and the program latencies of SLC pages and upper pages vary at increasing rates while the program latency of lower pages slowly slips to its minimum value). In Figure 4(c), all read latencies keep steady over the lifetime, because the read processes are changeless regardless of the condition of the tunnel oxide. Reading an SLC page and a lower page are most time saving, since only one read reference voltage should be applied to detect the V_{th}, followed by reading an upper page, which gets data after adopting two read reference voltages.

Compared with planar NAND flash, 3D FG NAND flash brings more affecting factors on performance and reliability variations. Since read performance remains unchanged among pages of the same type (SLC pages, lower pages or upper pages), here we investigate the variations in program performance among tiers and vertical planes, as shown in Figure 5.

Observation P2. Program latency among tiers varies greatly while almost all VPs show similar program performance except VP 0. Figure 5(a) and (b) shows the program latency of each tier and each VP along the lifetime, respectively. The lines of the same type of pages in each figure form a set. The more compact the lines in a set are, the fewer performance variations the corresponding type of pages has. Intuitively, the latency distributions of SLC pages, lower pages, and especially upper pages among tiers are more irregular and dispersive over the whole lifetime, compared with those among VPs. This effect is due to the *cross-tier process variations.* Processes (e.g., etch and deposit, etc.) are unable to guarantee the consistency of geometries of cells over tiers and even vary significantly, which leads to the decentralized distributions of program performance among tiers. But for WLs in the same tier, they (except for the first WL) show similar performance. Since each VP contains one WL in each tier, no obvious variation among VPs (except for VP 0) exists.



(a) Different tiers ($0 \sim 33$)

(b) Different vertical planes ($0 \sim 15$)

Fig. 5. Program performance variations with different physical locations.

The MLC pages in the VP 0 is shown as a special case, in which both lower pages and upper pages exhibit higher latencies than those in other VPs. However, as P/E cycles increase, the gap is narrowed and finally becomes negligible. These phenomena are due to the different specifications required for the first VP during manufacture, which affects not only performance but also reliability, as observed in the following subsection.

Implications. Block erase latency of 3D FG NAND flash memory is larger than that of planar NAND flash [31] and rapidly increases, from 10ms to more than 30ms, as P/E cycles increase. Moreover, the number of flash pages in a 3D FG NAND flash block is larger than that of planar NAND flash. Thus a GC operation becomes more time-consuming even with the same workload properties and GC algorithms. Since GC operations can block user I/O requests, the performance of flash-based storage systems significantly degrades. A potential solution is to *amortise GC overheads*, i.e., valid page migrations and block erases, over multiple I/O requests [18]. However, the most costly block erase operation is atomic for planar NAND flash memory and blocked I/O requests have to wait several milliseconds before an erase completion. Since 3D FG NAND flash memory has a larger flash block size and erase latency, GC overheads would be a more serious concern. Fortunately, 3D FG NAND flash memory inherently supports *erase suspend and resume* commands, providing a chance to remove the conflicts between costly block erases and I/O requests. A block erase operation can be immediately suspended when I/O requests arrive and be resumed after completing the I/O requests.

Furthermore, the erase latency of a flash block has a steep increase before an erase failure. The beginning of the steep increase can be regarded as a signal that the block's lifetime will end in the near future. Thus, bad block management can leverage this feature to *predict bad blocks*.

In addition, the program latencies of flash pages vary dramatically among different physical locations, such as between SLC/MLC-lower pages and MLC-upper pages, and even among different tiers. This page program heterogeneity motivates us to design a *program heterogeneity-aware page allocation policy*, which maintains multiple program points to free flash pages of different program

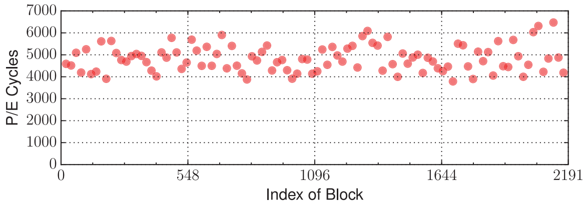


Fig. 6. The endurance distribution of 100 blocks.

latencies by scheduling the destination pages of valid page migrations in GC operations [16]. When serving a program request, a program point is chosen according to program requests' performance requirements. For example, fast-programing pages can be used to improve peak performance for latency-critical program requests.

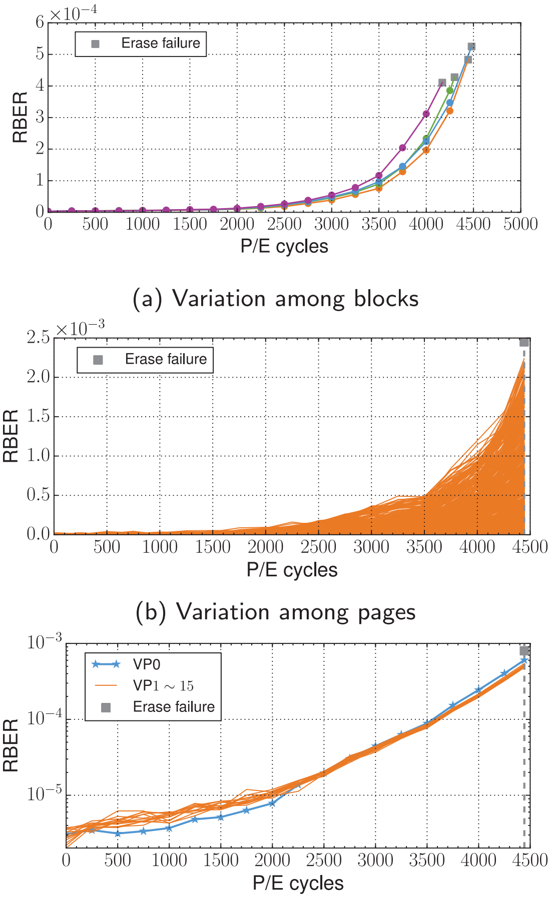
5.2 Reliability

Data integrity has always been a challenge in planar NAND flash and now in 3D NAND flash. The main cause is the wear of the tunnel oxide, resulting in irreversible physical damage. The other mechanisms (program/read disturbs, Gate Induced Drain Leakage (GIDL) [22], Random Telegraph Noise (RTN) [14], etc.) also make 3D FG NAND flash face the austere challenges of reliability. In this subsection, we focus on researching the patterns of endurance, program disturb errors, read disturb errors, and retention errors caused by these mechanisms.

5.2.1 Endurance. An erase operation is performed successfully when all cells in the block are reset to the E state, which is verified by checking if all the threshold voltages, V_{th}, are under an erase reference voltage (V_{Eref}). However, as the number of trapped electrons rises with P/E cycles, V_{th} shifts to a higher level continuously. Once there are enough trapped electrons in the tunnel oxide, the V_{th} becomes larger than V_{Eref}, which causes failures in the verifications, thus producing an erase failure. Program and read failures can also shorten the endurance of flash blocks. However, in our experiments on 3D FG NAND flash, all block failures are caused by erase failures.

We randomly choose 100 blocks and repeatedly erase and program them until a failure occurs in each block. For the sake of observing the trends of raw bit error rates (RBERs) over the lifetime, we sample data every 250 P/E cycles, and during each sample, we program all pages in a block sequentially and read each page immediately after it is programmed and compare the data with the original data. We use pseudo-random data for each program operation, since a data randomization scheme is applied in flash controllers or integrated into flash chips nowadays. Figure 6 shows the endurance distribution of tested blocks in a chip (containing 2,192 blocks).

Observation R1. In 3D FG NAND flash, the endurance of blocks is fluctuant in a chip and falls down significantly compared with planar NAND flash. As illustrated in Figure 6, most of the blocks can withstand the wear of 4,000 to 6,000 P/E cycles, and the actual endurance differs from block to block. In our evaluations, the earliest erase failure appears at 3,778 P/E cycles, and the latest one happens at 6,470 P/E cycles, which is 71% higher than the minimum. The average of endurance is 4,828 P/E cycles with the standard deviation of 597 P/E cycles. Although the nominal endurance of planar NAND flash devices in 24nm and 50nm is, respectively, 3,000 and 10,000, the actual numbers that they can be capable of are much larger in previous tests, often by a factor ranging from dozens to 100 [7, 12]. This 3D FG NAND flash is claimed to store roughly the same number of electrons per cell as the 50nm process did. Therefore, even for the best endurance in these tested blocks (6470 P/E cycles), it is still really dissatisfactory. This is because as an alternative solution in the 3D era, the first-generation products of 3D FG NAND flash has not been released until 2015, the



(a) Variation among blocks

(b) Variation among pages

(c) Variation among VPs

Fig. 7. Degradation speed variations over lifetime.

process is much more complex, and the process control and the quality are imperfect and unstable compared with planar NAND flash, which has been researched and produced for decades.

Observation R2. Blocks in a chip and pages in a block exhibit different degradation speeds and the speeds accelerate dramatically with P/E cycles. Figure 7(a) shows the RBER variations of 4 blocks in a chip. The RBERs of the four blocks start from about 6×10^{-6} and end with a range of 4×10^{-4} to 5.2×10^{-4}. With the rise of P/E cycles, the slopes become larger, indicating the faster growth of RBERs. Moreover, it is obvious that a block with a lower RBER has better endurance with high probability, because heavy wear of the tunnel oxide leads to a higher RBER. But this correlation is not always correct, since the endurance is determined by the first wear-out cell, not the average endurance of all cells. Thus, it is necessary to understand the variation inside a block. However, investigating each cell in a block is complex and unnecessary, because a page is the smallest unit for basic operations. Therefore, as shown in Figure 7(b), we illustrate the typical RBERs of all 1,024 pages in a block over the lifetime. We can observe that there are huge variations among pages, e.g., when the block wears 4,000 cycles, the RBER of Page 128 is 1.2×10^{-3}, while there are still dozens of pages with RBERs lower than 10^{-6}. In addition, the relative relationships among pages are not fixed, for example, Page 959 has a lower RBER than Page 1018 when P/E cycles are 2000 (1.7×10^{-5} vs. 9×10^{-5}), but when P/E cycles reach 4,000, the relationship is reversed (6.4×10^{-4} vs. 2.7×10^{-4}).

Observation R3. WLs in VP 0 degrade faster than those in other VPs. In the early stage of the lifetime, VP 0 shows lower RBERs than other VPs, whereas as P/E cycles increase, the RBERs of VP 0 climb faster and, finally, reach a higher level, as shown in Figure 7(c). This has been discussed in Section 5.1, due to the variation of specifications during fabrication. Since there are not many differences among VPs (only VP 0 shows a slightly higher degradation speed), we do not deliberately discuss variations among VPs in the rest of this article.

Implications. Since different blocks have different endurance, e.g., ranging from 3,778 to 6,470 P/E cycles in our experiment, wear leveling algorithm should not treat flash blocks equally. *A flash block with higher endurance should undertake more P/E cycles* to achieve a similar wearing degree to a flash block with lower endurance. Another implication is that P/E cycles are not an accurate metric to indicate the wearing degree of a flash block. Instead, *RBERs are a better wearing index*, although the statistics are more complicated.

Traditionally, a flash block is regarded as a bad block and will be abandoned once one of its pages becomes unreliable and unserviceable. However, different pages in a block also vary in degradation speeds, which means that some pages will become unreliable earlier than other pages. Therefore, competent flash pages in bad blocks are underused. To improve the space utilization, we suggest employing a *more fine-grained bad block management scheme*, which prolongs the lifetime of flash blocks with competent pages. When unreliable pages are detected in a flash block, the block remains to be in service but the unreliable pages are skipped. As the number of unreliable pages increases to a certain threshold, the block can be labeled as a bad block.

In addition, the endurance variations among flash blocks and among pages in a flash block also impose a great challenge on the ECC design due to a tradeoff between reliability and performance overheads. If a strong ECC is used to guarantee the reliability of the weakest flash page, then high decoding latency and large space overheads are also introduced. A *rate-adaptive ECC algorithm* may be used to provide different levels of error correction capabilities for different flash pages, but the hardware complexity and multi-rate ECC management overheads would significantly increase. Hence, we suggest employing a *redundant arrays of independent disks* (RAID) technique to enhance the storage reliability beyond ECC protection. For example, multiple flash pages either in a flash block or over flash blocks can form a RAID-5 stripe [19]. Since the pages in a stripe exhibit different RBERs, they are not likely to encounter a failure concurrently.

5.2.2 *Program Disturb Error.*

In 3D architectures, since all control gates of a tier in a block are short-circuited together, V_{pgm} is applied to all WLs in the tier containing the target WL during a program operation. Hence, in addition to parasitic capacitance coupling effects among WLs, the high voltages (V_{pgm}) can also lead to strong program disturbs through unintentional electron injection. Here, our first goal is to investigate which is the main factor of inducing program disturbs by conducting two experiments. The first one is to measure the sources of program disturbs caused by parasitic capacitance coupling effects. Since parasitic capacitance coupling effects greatly weaken with distance between two WLs, we count program disturb errors occurring in a target WL caused by programming the pages in a program disturbing WL set, including the two lower/upper adjacent WLs in the same VP and the two backward/forward adjacent WLs in the same tier to measure direct and indirect disturbs. For example, for a target WL (Tier x WL n), it can be affected by lower/upper pages in Tier $x-1$ WL n and Tier $x+1$ WL n (*direct disturbs in vertical direction*, DDVs), in Tier $x-2$ WL n and Tier $x+2$ WL n (*indirect disturbs in vertical direction*, IDVs), in Tier x WL $n-1$ and Tier x WL $n+1$ (*direct disturbs in horizontal direction*, DDHs),

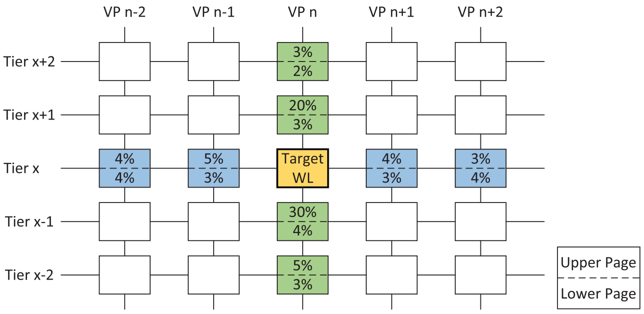


Fig. 8. Normalized contributions of two adjacent WLs in horizontal and vertical directions to program disturb errors.

and in Tier x WL $n-2$ and Tier x WL $n+2$ (*indirect disturbs in horizontal direction*, IDHs). In this experiment, we first erase a block, then program lower and upper pages in a target WL, and, finally, program one WL in the program disturbing WL set and read the target WL before and after performing each program operation. These three steps are repeated until disturbs from all WLs in the set are measured. By comparing the data before and after each program operation, we can get the number of errors caused by a program operation. Figure 8 shows the contribution of each page to the program disturb errors of the target WL.

Observation R4. Different pages in the vertical direction make different contributions to program disturb errors and the upper pages in adjacent tiers have the greatest impact. We can see that the program disturb errors caused by two upper pages in the vertical adjacent WLs account for a half of the total errors. The IDVs contribute only about 1/5 of the DDVs (13% vs. 57%), because the coupling capacitance decreases dramatically with the distance between two WLs. In the horizontal direction, the eight neighboring pages are responsible for 30% program disturbs, which are fewer than a half of those contributed by pages in the vertical direction. However, an interesting phenomenon can be observed in Figure 8, where the DDHs and the IDHs occupy the same proportions. This cannot be explained by coupling effects. Instead, it is due to high control gate voltages during programming.

Since all WLs in the tier containing the target WL withstand the same voltage (V_{pgm}) when performing a program operation, unselected WLs in that tier are supposed to suffer similar levels of program disturbs. To verify this and study program disturb–induced error accumulation within a tier, we run a second experiment. We first program WL 0 (both two pages) in Tier x and read it and then perform two sub-experiments: (1) each time after programming WL 0, we only program one WL (both two pages) in Tier x and then read WL 0 to measure program disturb errors caused by an individual WL and (2) program from WL 1 to WL 15 in Tier x (in WL order) and after programming each WL, read WL 0 to measure cumulative program disturb errors within a tier.

Observation R5. WLs introduce similar levels of program disturbs to the other WLs in the same tier and program disturb errors increase more slowly with program disturb counts. As shown in Figure 9, in Tier x, each WL can produce RBERs of about 9×10^{-6} to WL 0, which means program disturb errors within a tier are mostly caused by high applied control gate voltages, V_{pgm}. This is because, in 3D FG NAND flash, cells are circular, and within a tier, they are surrounded by interconnected control gates and do not share channels. These features significantly reduce the coupling capacitance between WLs within a tier. Because of the weak coupling effect between WL 0 and 1, we can also see that WL 1 results in slightly higher RBERs. Although a WL in Tier x induces more than a quarter of errors caused by Tier $x-1$ WL n, as shown in Figure 8, and a tier

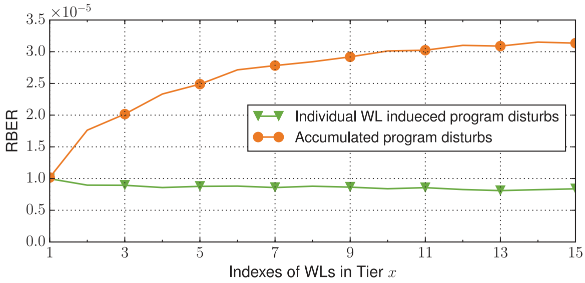


Fig. 9. Individual WL induced and accumulated program disturb errors of WL 0 in Tier x.

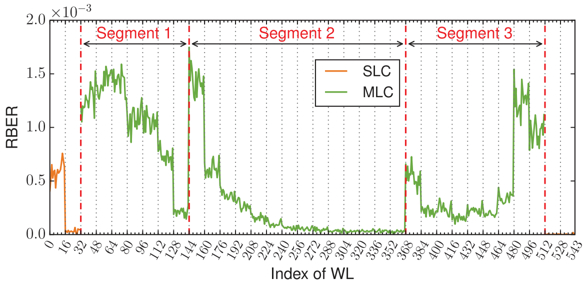


Fig. 10. Distribution of program disturb errors among WLs under the wear of 4, 000 P/E cycles. The WLs between two adjacent dot lines form a tier.

has 16 WLs, accumulated program disturb errors are still lower than those only caused by Tier $x - 1$ WL n, since as shown in Figure 9, the slope of RBERs becomes smaller as program disturb counts increase and the RBERs induced by 15 WLs are only triple those introduced by one WL in the horizontal direction.

Combining Figures 8 and 9, we can draw the following conclusion. **Observation R6.** Parasitic capacitance coupling effects and high control gate voltages are the two main factors of program disturb among tiers and among WLs within a tier, respectively, and the former can lead to more errors.

As observed above, large variations exist among pages/WLs in a block. Here, we investigate if location affects the distribution of program disturb errors or not.

Observation R7. Program disturb errors vary drastically with locations. As shown in Figure 10, we observe that in a tier, RBERs fluctuate slightly around the average RBERs of the tier, while among tiers, the maximum average RBERs can be up to 238 times and 54 times of the minimum average values in SLC mode (6×10^{-4} of Tier 0 vs. 2.5×10^{-6} of Tier 32) and MLC mode (1.4×10^{-3} of Tier 9 vs. 2.7×10^{-5} of Tier 22), respectively. To observe and analyze the varying patterns clearly, we roughly divide the MLC tiers into three segments: Segment 1 (Tier 2 to Tier 8), Segment 2 (Tier 9 to Tier 22), and Segment 3 (Tier 23 to Tier 31). In Segment 1 and Segment 2, the RBERs gradually decrease, while the RBERs show a U-shaped distribution in Segment 3. There are two reasons for the enormous variations among tiers: cross-tier process variations and location-related effects (e.g., the GIDL), which raise the RBERs locally. For the fabrication of 3D FG NAND flash, one of the most critical step is high aspect ratio etch [10]. Ion or neutral flux ratio changes with etch depth, which leads to the inconsistency of the geometries of cells at different depths (i.e., in different tiers), as shown in Figure 11. Cells at the top of the stack are always larger than cells at the bottom. These cross-tier process variations render the non-uniform performance and reliability of cells in

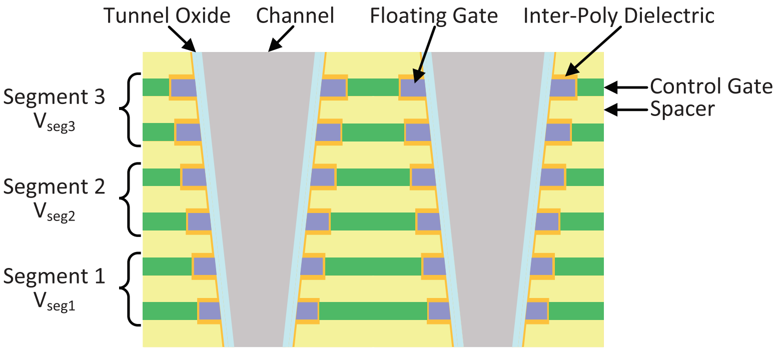


Fig. 11. An example of cross-tier process variations and segmented voltages. V_{seg1}, V_{seg2}, and V_{seg3} are the same types of voltages (e.g., program voltage) and have different values.

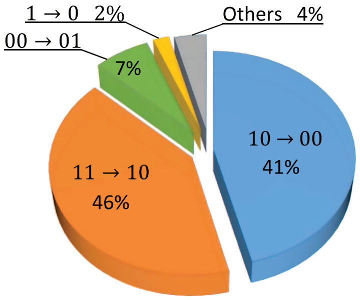


Fig. 12. Value dependence of program disturb errors.

different tiers, which is often contrary to vendors' expectations. An optional method to overcome the problem is that applying different voltages to different tiers in program and read operations according to their depths. However, this method greatly increases the complexity of control circuits in flash chips. As a tradeoff between accuracy and complexity, the stacked tiers are divided into multiple segments. Tiers in each segment are applied the same voltages, and different segments are applied different voltages. This policy causes cliffs of RBERs between adjacent segments. Moreover, for some specific tiers, the location-related effects can also affect the RBERs. For example, the GIDL[2] incurs numerous errors in Tier 0, which is the closest data tier to the SLS transistors, leading to much higher RBERs than those in other SLC tiers. The influence of cross-tier process variations also contributes to read disturb and retention errors.

We now explore correlations between the possibility of errors and the stored data values before and after an error occurs. We first program a block with random data in page order and after programming each page, read the page, and then when the programming of the whole block finishes, read the data stored in each page. Through comparing the two records, if an error occurs, then we can get the values before and after program disturbs. Figure 12 shows the possibility of different error patterns of program disturb errors. In the rest of this article, 1 → 0 and 0 → 0 only denote the errors in SLC WLs and two-bit errors occur in MLC WLs.

[2] The GIDL effects are activated when the SLS transistors are OFF and high voltages are applied to their drain and generate the electron-hole pairs that can be injected as hot carriers in the floating gates of nearby cells.

Observation R8. The frequencies of program disturb errors greatly depend on the values stored, among which the cells in the E and the L_{M1} states are the riskiest. As shown in Figure 12, we observe that (1) about 96% program disturb errors are caused by the shift of threshold voltages into the next higher V_{th} widows (i.e., 11 → 10, 10 → 00, 00 → 01, and 1 → 0) and (2) two error patterns, 11 → 10 and 10 → 00, are the two dominant types of program disturb errors, reaching up to 87% in total, where the former accounts for 46% and the latter is 41%. This is because, during a program operation, some electrons are unintentionally injected to floating gates in disturbed WLs by parasitic capacitance coupling effects and high control gate voltage. Electrons stored in a disturbed cell can generate an opposing electric field from the channel to the floating gate, which partially counteracts program disturbs by reducing the effective electric field from the floating gate to the channel. If a disturbed cell contains fewer electrons (e.g., in the E or the L_{M1} state), then more extra electrons can be injected into the floating gate due to the weak opposing electric field. By contrast, if a disturbed cell holds more electrons (e.g., in the L_{M2} or L_{M3} state), then the stronger electric field induced by electrons in the floating gate can effectively decrease program disturbs. As a result, program disturb errors exhibit strong value dependence, and the lower the V_{th} a cell has, the higher the likelihood that a program disturb error occurs.

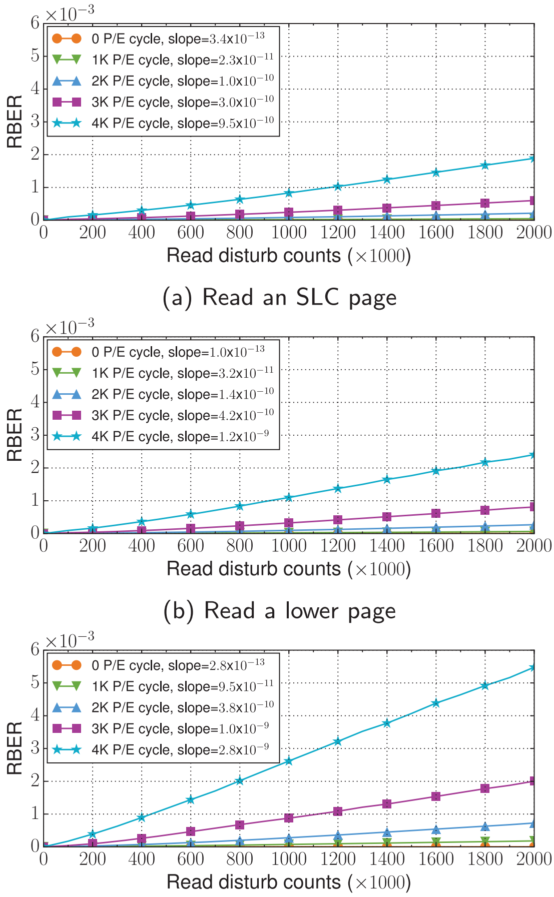
Observation R9. Upper pages are slightly more sensitive to program disturbs than lower pages. As shown in Figure 12, about 41% program disturb errors are contributed by lower pages ($\underline{1}0 \rightarrow \underline{0}0$), while upper pages bring about 53% errors ($1\underline{1} \rightarrow 1\underline{0}$ and $0\underline{0} \rightarrow 0\underline{1}$).

Implications. Observation R7 shows that different tiers show different inherent reliability and thus different resistances to program disturb errors. A possible design to reduce program disturb errors is to *rearrange the page programming order* in a flash block. Thus, pages in a more reliable tier can absorb more program disturb from programming of other pages.

Furthermore, the value-dependent error patterns in Observation R8 can *generate probability-based soft information*, which improves the precision of soft decoding of LDPC codes and thus the error correction capability.

5.2.3 Read Disturb Error. Read disturbs are the most common source of disturbs in NAND flash. Compared with program disturbs, which only occur a very limited number of times in a block before performing an erase operation, there is no limit on read disturbs. Due to the cumulative effect of successive *weak programs*, errors arise. To characterize read disturb error, we (1) randomly select several blocks, divide them into three groups, and choose an SLC page (Page 8), a lower page (Page 495), and an upper page (Page 558) as target pages for the three groups, respectively; (2) program each block by pseudo-random data and read the data as the baseline; (3) repeatedly read the target page of each block for 2 million times and read the whole blocks after every 100 thousand read operations to get the read disturbed data; and (4) after 2 million read operations, program and erase the blocks for 1,000 times. The last three steps are repeated until the blocks wear out.

Observation R10. The read disturb induced RBERs increase approximately linearly with the number of read operations, grow up dramatically with P/E cycles, and an upper page can introduce more errors than an SLC page and a lower page. Figure 13 shows trends of RBERs with read disturb counts ranging from 0 to 2 million under different levels of wear on flash blocks. The RBERs induced by a specific page under a certain amount of P/E cycles exhibit a roughly linear relation with read disturb counts. We also observe that the slopes of those lines, as listed in the legends in Figure 13, surge by hundreds of times from 0 to 1,000 P/E cycles and then roughly triple every 1,000 P/E cycles. Moreover, for an upper page, it can lead to more than twice as much read disturbs as an SLC or a lower page. The effects are because the number of unintentionally charged electrons caused by read disturbs have a positive correlation with the accumulated weak programming time



(a) Read an SLC page

(b) Read a lower page

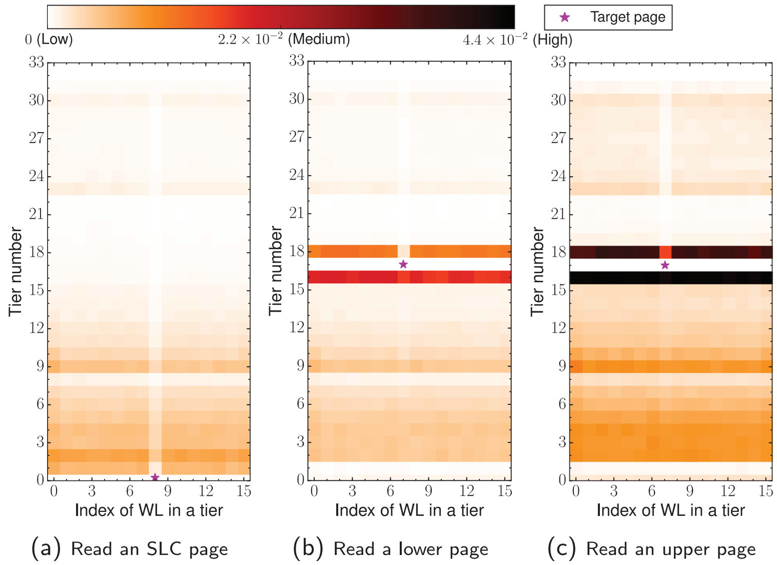
(c) Read an upper page

Fig. 13. RBERs vs. read disturb counts under different P/E cycles.

and the tunnel current. The time of V_{pass} applied to disturbed WLs is fixed during a read operation, and it is longer to read an upper page than an SLC or lower page (as described in Section 5.1), thus the effect of weak programs is accumulated over successive read operations and more significant when reading an upper page. For weak programming current, it has a negative correlation with the insulating abilities of the dielectric, which degrades rapidly at the beginning and then with a more stable speed. As a result, with more read disturb counts caused by an upper page and larger P/E cycles, errors occur with a higher probability.

Ideally, when reading a page, the tier containing the target page is applied V_{ref}, while the other tiers are applied V_{pass}; thus, the read disturb errors should distribute uniformly among the SLC WLs and the MLC WLs in other tiers separately. To observe whether there is location dependence of read disturb errors, we illustrate the distributions among all WLs in a block with different target pages, as shown in Figure 14.

Observation R11. The neighboring tiers of a target page suffer much more serious read disturbs than other tiers. In Figure 14(b) and (c), it is obvious that the RBERs of Tier 16 and Tier 18 are much higher than any other tiers. When the target page (Page 8) is in Tier 0, the RBERs of Tier 1 are



(a) Read an SLC page (b) Read a lower page (c) Read an upper page

Fig. 14. Distributions of read disturb errors among WLs in a block after 2 million read operations under the wear of 4,000 P/E cycles. (only Tiers 0, 1, 32, and 33 contain SLC pages).

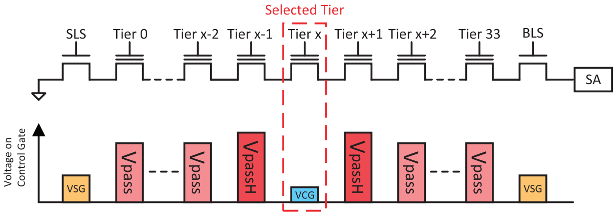


Fig. 15. Voltage distribution during a read operation with weakened coupling effect.

relatively higher but lower than Tier 2. However, this does not mean that the read operations on Page 8 disturb Tier 1 less than Tier 2, since pages in Tier 1 are in SLC mode and more robust. The reason of much higher RBERs in adjacent tiers is that the pass voltage applied to the adjacent tiers is higher than those applied to the other tiers. When applying the low-pass voltage to the adjacent tiers, for a cell in the WL storing the target page, if it is in a low V_{th} state and the adjacent cells in the same string are in high V_{th} states, then the V_{th} of the target will be coupled up when a read operation is performed. To reduce this coupling effect, V_{passH}, which is about 0.4V higher than the original V_{pass}, is applied to the adjacent tiers to increase pass voltage window, as shown in Figure 15[3]. As a result, a read operation has a stronger *weak programming effect* on the adjacent tiers, which leads to higher read disturb errors.

Observation R12. The VP that contains a target page is less disturbed by read operations, especially for those WLs in the upper tiers. During a read operation, there are two situations based on the applied read reference voltage after the capacitors are precharged (usually ~1.1V).

[3] As mentioned in Observation R7, in a read operation, different pass voltages are applied to different segments. In Figure 15, to highlight the higher pass voltage applied to the adjacent tiers, we use the same V_{pass} to denote the pass voltages applied to the other tiers.

For a cell in the target WL, if $V_{ref} < V_{th}$, then it is OFF, and the channel voltages of the cells above and below it in the same string are about 1.1V and 0V, respectively; if $V_{ref} > V_{th}$, then it is ON, and the voltages from the top cell to the bottom cell in the same string drop gradually from the voltage of the capacitor to 0V. Since a high channel voltage can diminish weak programming effect by reducing the voltage difference between gate and channel, the cells with higher channel voltages induce fewer read disturb errors. As a result, for WLs in the VP containing the target WL, especially for which in the upper tiers, the RBERs are lower.

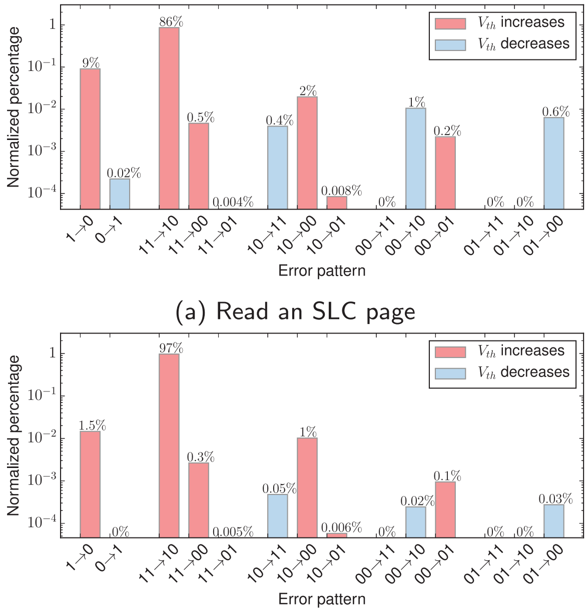
Observation R13. Except for the tier containing the target page and the two neighboring tiers, read disturb errors show the similar location dependence as program disturb errors. Except for the adjacent tiers, the other tiers can be divided into several segments according to the distributions of RBERs. In Figure 14, the RBERs from Tier 2 to Tier 8 (Segment 1) and from Tier 9 to Tier 22 (Segment 2) decrease separately, and in Segment 3, the RBERs are distributed uniformly from Tier 24 to Tier 29, which are lower than those of Tiers 23 and 30. This phenomenon is due to the cross-tier process variations and different voltages applied to different segments, as mentioned in Observation R7 and illustrated in Figure 11. The cross-tier process variations caused by high aspect ratio etch lead to gradual changes in a segment, while different voltages (to reduce the variations of performance and reliability among tiers caused by the cross-tier variations) result in jumps at the borders between segments.

During a program operation, the pulse voltage increases step by step, since with more electrons in a floating gate, V_{th} increases and charging becomes harder. Hence, the state of a cell also has an impact on the weak programming effect it suffers. Next, we explore how different stored values response to read disturbs.

Observation R14. Most of the read disturb errors are caused by V_{th} shifts to the adjacent higher V_{th} states, and the cells in the E state contribute the majority. We separately count read disturb errors of all possible error patterns and show the relative percentages in Figure 16. We can see that (1) the threshold voltages of more than 98% errors increase and more than 99% of them jump to the adjacent states and (2) the error rate of cells holding 1 or 11 accounts for more than 95%. These distributions are due to a lower V_{th} on a cell leading to a larger voltage difference through the tunnel, rendering more disturbs to the cell. Hence, we conclude that read disturb errors highly depend on stored values and the 11 state most likely gives rise to an error.

Observation R15. Upper pages are more sensitive to read disturbs than lower pages. In Figure 16, it is obvious that the jumps between the $\{E, L_{M1}\}$ states and the $\{L_{M2}, L_{M3}\}$ states, resulting in lower page errors, are much harder than the jump from the E state to the L_{M1} state ($1\underline{1} \rightarrow 1\underline{0}$), which is the major source of upper page errors. When reading a lower/upper page repeatedly, the errors in upper pages are about 75 times of those in lower pages (97.5% vs. 1.3%).

Implications. Intensive read operations gradually increase the RBERs of flash pages in two adjacent tiers, which are more vulnerable. A large read cache can largely alleviate this problem, but for flash-based storage system without large RAM, we need careful designs to overcome the reliability challenge brought by read disturbs. We suggest introducing a *refresh policy* that reprograms the data in flash pages when the pages have been read more than certain times. However, refresh operations can be costly, and thus the amount of target data should be reduced. Note that read disturbs occur in a certain pattern and cells in different tiers have different inherent reliability, as shown in Figure 14. The target data can be modestly chosen according to these two features to achieve a tradeoff between reliability and performance. We also suggest a *read disturb-aware page allocation policy* to distribute read-intensive data to flash pages in different tiers. Similarly to program disturb errors, the value dependence in Observation R14 can also help to improve the LDPC efficiency.



(a) Read an SLC page

(b) Read a lower/upper page

Fig. 16. Value dependence of read disturb errors.

5.2.4 Retention Error. Due to the imperfect electrical insulation and charge detrapping in the tunnel oxide, electrons stored in floating gates may escape towards the silicon substrate over time, leading to V_{th} shifts and even data corruptions. In this subsection, our goal is to understand the main features of retention-induced errors. We first choose nine groups of fresh blocks, one of which is kept fresh and the other groups are worn by random data to 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, and 4,000 P/E cycles, respectively. Then, we program all blocks with random data and read them immediately as *baseline data* and, finally, read out data from all blocks after a retention age t_{room}. Through comparing the baseline data and the data after retention, we can obtain the features of retention errors. In this experiment, retention ages are set as 1 day, 1 week, 1 month, 1 year, and 5 years, of which 1 day and 1 week are measured by putting chips under room temperature (25°C) for 1 day and 1 week, respectively, while 1 month, 1 year, and 5 years are measured by baking chips to accelerate this experiment (details are described in Section 4).

Observation R16. RBERs increase rapidly with retention ages and P/E cycles and reach very high levels after 1-week retention or longer. In Figure 17, error rates surge from 4.1×10^{-4} (1-day retention) to 1.6×10^{-1} (5-year retention) under 4,000 P/E cycles, with a 390× growth, and in the case of 1-year retention, the RBERs rise 1,500 times over 4,000 P/E cycles, reaching 1.2×10^{-1}. As discussed above, the insulation of the tunnel oxide degrades with P/E cycles, resulting in growing SILC, which is the main source of retention loss compared with charge detrapping. This explains why RBERs go up with retention ages and P/E cycles. We can also see that the growth rate first becomes larger and then smaller as the retention age increases. We use $AR = RBERs/Retention\ Age$ to denote the average RBERs generated in one day during the whole certain retention age. *AR*s of 1-day, 1-week, 1-month, 1-year, and 5-year retentions are 4.1×10^{-4}, 1.6×10^{-3}, 1.1×10^{-3}, 3.3×10^{-4}, and 8.8×10^{-5}, respectively. Moreover, even only after 1-week retention, the RBERs can reach more than 1%, which is usually unacceptable in practical applications.

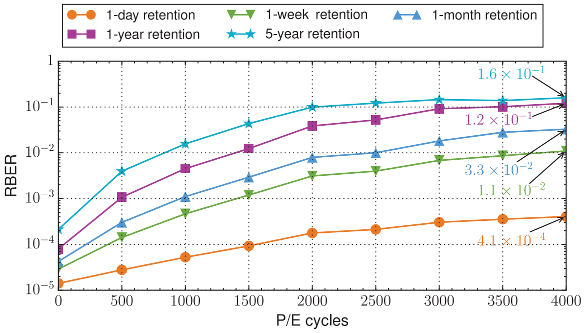
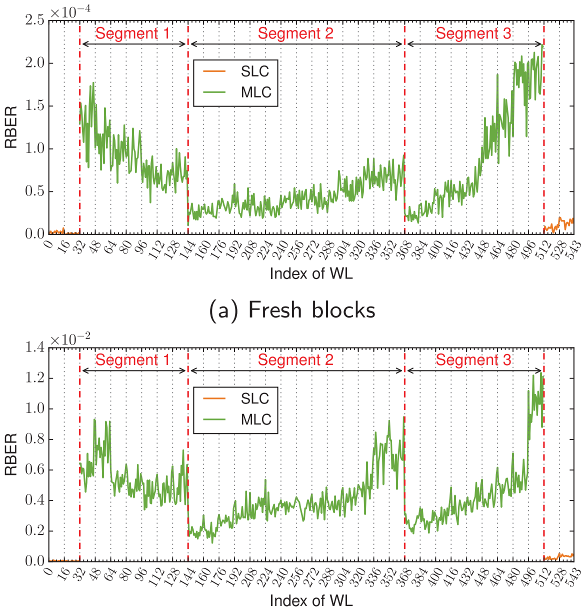


Fig. 17. RBERs vs. P/E cycles with various retention ages.



(a) Fresh blocks

(b) Under the wear of 4000 P/E cycles

Fig. 18. Distribution of retention errors among WLs after 1 week retention. The WLs between two adjacent dot lines form a tier.

Compared with program and read operations, which affect nearby WLs more than far WLs, retention can exert influence uniformly on all cells in a block, since all cells are subject to the same period of time during retention. Hence, the distribution of retention errors is more suitable to reflect process variations in chips. As shown in Figure 18(a), even in fresh blocks, retention errors also vary significantly among tiers. This proves the existence of the cross-tier process variations in 3D FG NAND flash chips. To compare the location dependence between program/read disturb errors and retention errors, we show the relations between retention errors and locations in Figure 18(b) under the same degree of wear (4,000 P/E cycles).

Observation R17. Retention errors show only partly similar location dependence to program/read disturb errors. As shown in Figure 18(b), similarly to program/read disturb errors, retention errors can also be divided into the same three segments with clear boundaries, and the

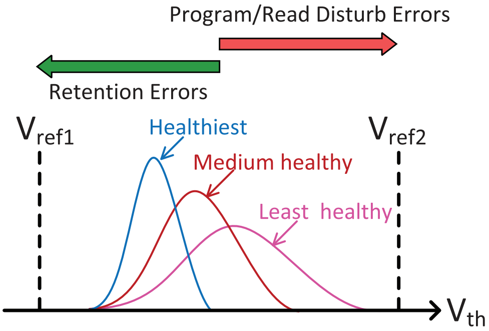


Fig. 19. Distributions of V_{th} under various health conditions.

RBERs in Segment 1 decrease from Tier 2 to Tier 8. This is also due to the cross-tier process variations and different voltages applied to different segments, as mentioned in Observation R7 and illustrated in Figure 11. However, in Segment 2 and 3, they exhibit distinct tendencies: program/read disturb errors fall down in Segment 2 and show a U-shaped distribution in Segment 3 (as shown in Figures 10 and 14), while retention errors keep upward tendencies in both segments. These differences can be explained by the various V_{th} distributions and the different ideal V_{th} shift directions between program/read disturbs and retention, as shown in Figure 19. If V_{th} shifts to a voltage lower than V_{ref1}, then a retention error occurs, and if V_{th} shifts to a voltage higher than V_{ref2}, then a program or read disturb error arises. For a segment, even if the same voltages are applied, due to the cross-tier process variations caused by multi-tier-stack structure, cells in different tiers are under various wear degrees (health conditions). Here, for simplicity, we qualitatively illustrate three V_{th} distributions to represent three health conditions. The less healthy a cell is, the higher the voltage V_{th} shifts to and the larger the leakage current the tunnel oxide conducts. A higher V_{th} means a smaller voltage margin (between V_{th} and V_{ref2}) for a program/read disturb error and a larger voltage margin (between V_{ref1} and V_{th}) for retention errors. Therefore, if WL n is less healthy than WL m, for program/read disturbs, then WL n has smaller margins and larger leakage currents, leading to more program/read disturb errors; for retention, although WL n has larger leakage currents, it is also with larger margins, the retention errors of WL n are not necessarily more than those of WL m. For example, as shown in Figures 10 and 18(b), Tier 3 has both more program disturb errors and retention errors than Tier 8, while Tier 9 has much higher program-induced RBERs than Tier 22, which means Tier 9 is under a worse health condition than Tier 22, but instead, for retention errors, the RBERs of Tier 22 are higher than Tier 9. This is why the distribution of retention errors is different from that of program/read disturb errors.

From above discussions, we believe that: **Observation R 18.** When suffering the same interference, program and read disturb errors can be used to indicate the health conditions among WLs, but retention errors cannot.

Unlike program disturbs and read disturbs, which shift the V_{th} of disturbed cells to higher levels, electron loss during retention leads to V_{th} reduction. As a consequence, retention errors show distinct value dependence from program and read disturb errors, as shown in Figure 20.

Observation R19. About 99.9% of retention errors are contributed by V_{th} drops to the neighboring lower states, in which more than a half manifest as 00 → 10. Charges escape from floating gates mainly through SILC flows, whose intensities positively correlate with the V_{th}. As a result, during retention, the higher the V_{th} of a cell is, the more electrons the cell loses. Intuitively, cells storing 01 should introduce the highest RBERs, followed by those holding 00 and 10. However,

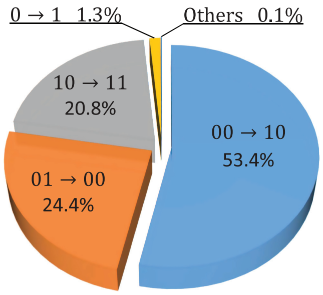


Fig. 20. Value dependence of retention errors.

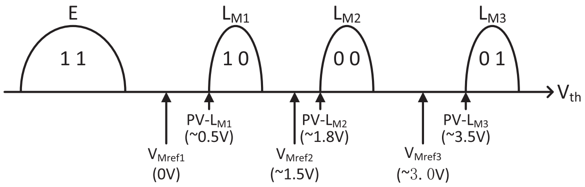


Fig. 21. Actual read reference voltages and program verify voltages in MLC.

as shown in Figure 20, the most common retention error is 00 → 10. This is due to the unequal voltage margins between read reference voltages and program verify (PV) voltages of the L_{M1}, the L_{M2}, and the L_{M3} states, as illustrated in Figure 21. For the L_{M1} and the L_{M3} states, the margins are both ∼0.5V, while the $PV\text{-}L_{M2}$ is only about 0.3V higher than V_{Mref2}. Therefore, compared with cells in the L_{M1} and the L_{M3} states, just a relatively fewer electron loss can make the V_{th} of a cell in the L_{M2} state lower than V_{Mref2}, leading to the occurrence of an error 00 → 10. For cells storing 01 and 00, although the margins are the same, the SILC of the former is larger due to higher V_{th}. Consequently, the error pattern 00 → 10 occurs most frequently, followed by 01 → 00 and 10 → 11.

Observation R20. Lower pages are slightly more sensitive to retention than upper pages. Unlike program and read disturbs, where upper pages are more interfered, retention can lead to more errors in lower pages ($\underline{0}0 \rightarrow \underline{1}0$ with 53.4%) than upper pages ($0\underline{1} \rightarrow 0\underline{0}$ with 24.4% and $1\underline{0} \rightarrow 1\underline{1}$ with 20.8%). The reason is explained in Observation R19.

Implication. 3D FG NAND flash memory shows more retention errors than planar NAND flash memory. Moreover, different tiers have dramatically different retention abilities. Thus, it is more challenging to ensure data persistence on 3D FG NAND flash memory. A *periodic refresh policy* has been proposed to solve the retention problem of planar NAND flash memory [9]. This refresh policy can also be applied to 3D FG NAND flash memory, but should be improved in two aspects. First, *the refresh frequency needs to be increased* to accommodate the weak retention ability of 3D FG NAND flash memory. Second, *different refresh frequencies* should be set for different tiers, i.e., adaptively to their retention abilities. Such a *retention heterogeneity-aware refresh policy* can achieve a fine tradeoff between reliability and refresh performance overheads. Moreover, a *retention-aware page allocation policy* can be adopted to store cold data in tiers with higher retention abilities. Furthermore, similarly to Observation R8, the value-dependent error patterns in Observation R19 can also be used to improve the efficiency of LDPC.

5.3 Summary

Based on the observations described and analyzed above, we summarize the main differences of 3D FG NAND flash from planar NAND flash and their implications on the designs of NAND flash

Table 3. Differences of 3D FG NAND Flash from Planar NAND Flash and Their Implications on the Designs of NAND Flash Management Techniques

Algorithm	Difference from planar NAND flash	Implication
Page mapping	• Larger block • Larger program latency variations among tiers (Observation P2) • More program disturb errors, read disturb errors and retention errors (Observations R5, R10, R16) • Distinct location dependence (Observations R3, R7, R11, R12, R17) • Distinct value dependence (Observations R8, R9, R14, R15, R19, R20)	• Program heterogeneity-aware page allocation policy • Rearrange the page programming order • Read disturb-aware and retention-aware page allocation policies
Wear leveling	• Shorter endurance (Observation R1) • Larger degradation speed variations among blocks (Observation R2)	• A flash block with higher endurance should undertake more P/E cycles • RBER-based wear leveling
GC	• Larger block • Larger and more rapidly increasing erase latency (Observation P1)	• Distributed GC to amortise overheads • Using erase suspend and resume commands to improve I/O performance
ECC	• Larger degradation speed variations among blocks and pages (Observation R2) • More program disturb errors, read disturb errors and retention errors (Observations R5, R10, R16) • Distinct location dependence (Observations R3, R7, R11, R12, R17) • Distinct value dependence (Observations R8, R9, R14, R15, R19, R20)	• Rate-adaptive ECC • Generating probability-based soft information to improve the precision of LDPC through the value dependence
Bad block management	• More rapidly increasing erase latency (Observation P1) • Larger degradation speed variations among pages (Observation R2)	• Using erase latencies to predict bad blocks • Skipping unreliable pages to delay the generations of bad blocks
Refresh	• Larger degradation speed variations among blocks and among pages (Observation R2) • More read disturb errors and retention errors (Observations R10, R16) • Distinct location dependence (Observations R3, R7, R11, R12, R17)	• Higher refresh frequency • Frequency-adaptive refresh policy
Internal RAID	• Larger degradation speed variations among blocks and among pages (Observation R2) • Distinct location dependence (Observations R3, R7, R11, R12, R17)	• Pages with different RBERs form a stripe to provide higher reliability

management techniques in Table 3. Compared with Planar NAND flash, the main differences are as follows:

- *Larger block.* Since 3D FG NAND flash enables increasing capacity in the vertical direction, the size of a block in 3D FG NAND flash is usually larger than that in planar NAND flash. A block in Intel-Micron 3D Gen1 FG MLC NAND flash contains 1,024 pages, while the number is usually no more than 512 in planar MLC NAND flash.
- *Larger program latency variations among tiers.* In 3D FG NAND flash, due to the cross-tier process variations, the geometries of cells cannot be guaranteed to be consistent over tiers,

resulting in program latency variations among tiers. Planar NAND flash does not have a concept of tier.

- *Larger and more rapidly increasing erase latency.* For an erase operation, it is completed successfully if all cells in the block return to the E state. In 3D FG NAND flash, a block contains much more cells than that in planar NAND flash, and those cells are with more process variations. To guarantee that all cells in a block are erased, an erase operation in 3D FG NAND flash takes a longer time than that in planar NAND flash. Since the 3D FG NAND flash technology is not mature yet and the process is much more complex, the quality of the tunnel oxide is much poorer compared with planar NAND flash. Thus, in 3D FG NAND flash, as P/E cycles increase, the defects in tunnel oxide accumulate faster, resulting in more rapidly increasing erase latency.
- *Shorter endurance.* Once an erase operation cannot be completed in a specified time, an erase failure occurs and the block fails. As mentioned above, erase latency in 3D FG NAND flash grows faster than that in planar NAND flash, resulting in shorter endurance.
- *Larger degradation speed variations among blocks and pages.* Compared with planar NAND flash, due to larger process variations, the quality of 3D FG NAND flash blocks fluctuates more greatly, leading to larger degradation speed variations among blocks. Because of the cross-tier process variations, 3D FG NAND flash also exhibits larger degradation speed variations among pages.
- *More program disturb errors, read disturb errors, and retention errors.* In planar NAND flash, program disturb errors are mainly caused by parasitic capacitance coupling effects, while in 3D FG NAND flash, except for that reason, high control gate voltages among WLs within a tier can also make a significant contribution, leading to more program disturb errors. For a read operation, in planar NAND flash, V_{passH} is applied to only 1 or 2 adjacent WLs, while it affects 1 or 2 adjacent tiers (16 or 32 WLs for the tested chips) in 3D FG NAND flash, resulting in more read disturb errors. As mentioned above, the quality of the tunnel oxide in 3D FG NAND flash is poorer, which means that cells have lower abilities to hold electrons during retention, causing more retention errors.
- *Distinct location and value dependence.* Unlike planar NAND flash, 3D FG NAND flash employs multi-tier-stack structure, which introduces serious cross-tier process variations caused by high aspect ratio etch. The cross-tier process variations and the corresponding approach adopted by the vendors (applying different voltages to different segments), and higher pass voltages applied to adjacent tiers in read operations, as well as location-related effects (e.g., the GIDL), lead to the distinct location dependence. The unequal voltage margins between read reference voltages and program verify voltages, different directions of threshold voltage shift and different degrees of interference caused by different types of operations result in distinct value dependence.

To achieve better performance and reliability through utilizing the characteristics of 3D FG NAND flash, some implication on the designs of NAND flash management techniques can be obtained from the observations.

- *Page mapping.* A flash controller can record multiple write points (optional free pages for the next programming). Due to the program latency variation, according to the criticality of application's write requests, the flash controller can provide multi-level write performance service by allocating fast-programming pages to critical requests and slow-programming pages to non-critical requests. Through this program heterogeneity-aware page allocation policy, the peak performance for burst critical workloads can be improved. Because of more program disturb errors, read disturb errors, and retention errors and distinct location dependence, the flash controller can select more reliable pages for the critical data by (1) rearranging the page programming order (e.g., allocating pages from the last 32 SLC pages to cold data before those pages should be programmed if programming the block in page order, and this policy does not violate the interleaved program operations among MLC pages) and (2) employing read disturb-aware and retention-aware page allocation policies (e.g., selecting pages suffering fewer retention errors in the multiple write points to cold data). After adopting such policies, the reliability of flash-based systems can be improved.

- *Wear leveling.* In general, wear leveling schemes intend to make all the blocks undergo the same P/E cycles to equalize the wearing degree. However, not all blocks with the same P/E cycles are in the same wear conditions. Hence, the lifetime of flash-based systems can be extended by imposing more P/E cycles on flash blocks with higher endurance and using RBER as the wearing index to perform more accurate wear leveling.
- *GC.* Due to the larger blocks and longer erase latency in 3D FG NAND flash, a complete GC operation consumes more time and causes more serious performance degradation. To provide sustained high performance, a flash controller can give application's requests higher priorities by (1) distributing a GC operation (executing GC operations in the idle time) and (2) using erase suspend and resume commands to avoid blocking read requests.
- *ECC.* Due to the variations of RBERs among blocks and pages and over the lifetime of NAND flash, employing a constant rate is not proper. A low rate means that the reliability of data cannot be guaranteed when and for where RBERs are high, and a high rate indicates encoding and decoding latencies are unnecessarily prolonged when and for where RBERs are low. By employing rate-adaptive ECCs (e.g., adopting a low rate in the early stage of flash's lifetime and a high rate in the late stage), both the performance and reliability are improved. By utilizing the characteristics of errors, probability-based soft information can be generated to enhance the LDPC decoding performance and thus reduce the decoding latency.
- *Bad block management.* Bad blocks can be predicted by using their erase latencies, therefore the data in those blocks can be migrated before they fail to ensure the reliability. When a page in a block becomes unreliable, the flash controller does not regard the block as a bad block but continues to use it until the number of unreliable pages exceeds a pre-set threshold, wherefore the lifetime is prolonged.
- *Refresh.* Since 3D FG NAND flash has more read disturb errors and retention errors to guarantee the reliability, a flash controller should use lower thresholds (e.g., read counts and retention ages) to increase refresh frequencies. Like the rate-adaptive ECC, a frequency-adaptive refresh policy (e.g., employing a low frequency for where retention errors are low and vice versa) can improve both the performance and reliability.
- *Internal RAID.* Due to wear leveling techniques, blocks in a flash-based system are in a similar degree of wear. As a technique borrowed from hard disk arrays, internal RAID in flash-based systems faces a higher risk of failure because of that feature of flash-based systems. For example, among RAID techniques, RAID-5 is the most popular approach and can recover data if a data/log block (a block in RAID does not represent a physical block in NAND flash) fails. If another data/log block fails during the process of recovery, then the data in the stripe cannot be recovered. For flash-based systems, the probability of this situation is much higher than that in hard disk arrays. This phenomenon motivates us to utilize the degradation speed variations and distinct location dependence to maximize the variations among pages in a stripe. Thus, the risk of simultaneous failures can be minimized, resulting in higher reliability.

In this article, our observations are based on the evaluations of 3D FG NAND flash. As a promising scheme, 3D CT NAND flash also brings new benefits and problem to NAND flash-based systems. Due to the enormous differences of the materials and structures between FG and CT, these observations cannot be generalized to 3D CT NAND flash. Based on our understanding of CT technique, we briefly compare the observed characteristics of 3D FG NAND flash with the projected characteristics of 3D CT NAND flash.

- *Performance.* Since storage layers in FG and CT cells employ conductor and insulator materials, respectively, the tunnel oxide in CT cells is much thinner than that in FG cells. Thus, in 3D CT NAND flash, injecting electrons from and to storage layers are easier, resulting in smaller erase and program latencies. Moreover, the coupling effect among cells in 3D CT NAND flash is dramatically decreased, which allows higher programming speed. For read operations, there is no obvious difference between the two types of 3D NAND flash.
- *Endurance.* As mentioned above, CT cells have thinner oxide barriers. Therefore, the voltage of an erase or a program operation in 3D CT NAND flash is lower than that in 3D FG NAND flash, resulting in fewer defects generated in a P/E cycle. The endurance of 3D CT NAND flash is usually better than that of 3D FG NAND flash.
- *Program disturb error.* Since 3D CT NAND flash employs insulating storage layer and different cell structure, the coupling effects among cells are very weak, resulting in much fewer program disturb errors than 3D FG NAND flash.
- *Read disturb error.* As discussed in Section 5.2.3, because of the coupling effect, the neighboring tiers are applied higher pass voltages, causing high read disturb errors in 3D FG NAND flash. On account of the weak coupling effect, 3D CT NAND flash does not need to employ that policy, which means fewer read disturb errors are generated.
- *Retention error.* Retention errors are the major issue of 3D CT technique, especially at a high temperature. In 3D CT NAND flash, the nitride layer of the cells in a string is continuously connected along the channel, forming a charge spreading path. This degrades the retention characteristic, leading to a much more severe problem of retention errors in 3D CT NAND flash than that in 3D FG NAND flash.

6 CONCLUSIONS

In this article, we conduct a comprehensive study on the state-of-the-art 3D FG NAND flash and provide the detailed characterizations of its performance and reliability. Based on the data measured on our FPGA-based evaluation platform, we make several observations and give detailed analyses from physical and circuit-level perspectives. 3D FG NAND flash shows higher RBERs, distinct location dependence of performance and errors caused by cross-tier process variations, and different value dependence of error patterns, compared with planar NAND flash. Moreover, we discuss several implications on flash management, such as read disturb-aware and retention-aware page allocation policies (page mapping), value dependence-based LDPC (ECC), bad block prediction (bad block management), and read disturb-aware and retention heterogeneity-aware refresh policies (refresh algorithm), and so on. We believe that our work presents an insight into the characteristics and a more efficient usage of 3D FG NAND flash.

In the future, we plan to characterize 3D CT NAND flash and develop characteristic-based methods to improve the performance and reliability of 3D FG/CT NAND flash-based storage systems.

REFERENCES

[1] Seiichi Aritome, Yoohyun Noh, Hyunseung Yoo, Eun-Seok Choi, Han-Soo Joo, Youngsoo Ahn, Byeongil Han, Sungjae Chung, Keonsoo Shim, Keunwoo Lee, Sanghyon Kwak, Sungchul Shin, Iksoo Choi, Sanghyuk Nam, Gyuseog Cho, Dongsun Sheen, Seungho Pyi, Jongmoo Choi, Sungkye Park, Jinwoong Kim, Seokkiu Lee, Sungjoo Hong, Sungwook Park, and Takamaro Kikkawa. 2013. Advanced DC-SF cell technology for 3D NAND flash. *IEEE Trans. Electr. Devices* 60, 4 (Mar. 2013), 1327–1333.

[2] David A. Baglee. 1984. Characteristics & reliability of 100A oxides. In *Proceedings of the IEEE International Reliability Physics Symposium (IRPS'84)*. 152–155.

[3] Simona Boboila and Peter Desnoyers. 2010. Write endurance in flash drives: Measurements and analysis. In *Proceedings of the USENIX Conference on File and Storage Technologies (FAST'10)*. 115–128.

[4] Yu Cai, Erich F. Haratsch, and Onur Mutlu amd Ken Mai. 2013. Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling. In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE'13)*. 1285–1290.

[5] Yu Cai, Erich F. Haratsch, Onur Mutlu, and Ken Mai. 2012. Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis. In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE'12)*. 521–526.

[6] Yu Cai, Yixin Luo, Saugata Ghose, and Onur Mutlu. 2015. Read disturb errors in MLC NAND flash memory: Characterization, mitigation, and recovery. In *Proceedings of the Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'15)*. 438–449.

[7] Yu Cai, Yixin Luo, Erich F. Haratsch, Ken Mai, and Onur Mutlu. 2015. Data retention in MLC NAND flash memory: Characterization, optimization, and recovery. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA'15)*. 551–563.

[8] Yu Cai, Onur Mutlu, Erich F. Haratsch, and Ken Mai. 2013. Program interference in MLC NAND flash memory: Characterization, modeling, and mitigation. In *Proceedings of the IEEE International Conference on Computer Design (ICCD)*. Asheville, NC, USA, 123–130.

[9] Yu Cai, Gulay Yalcin, Onur Mutlu, Erich F. Haratsch, Adrian Cristal, Osman S. Unsal, and Ken Mai. 2012. Flash correct-and-refresh: Retention-aware error management for increased flash memory lifetime. In *Proceedings of the IEEE International Conference on Computer Design (ICCD'12)*. 94–101.

[10] Jim Cooke. 2017. Overcoming challenges in 3D NAND volume manufacturing. In *Proceedings of the Flash Memory Summit*. 1–21.

[11] Robin Degraeve, F. Schuler, Ben Kaczer, Martino Lorenzini, Dirk Wellekens, Paul Hendrickx, Michiel van Duuren, G. J. M. Dormans, Jan Van Houdt, L. Haspeslagh, Guido Groeseneken, and Georg Tempel. 2004. Analytical percolation model for predicting anomalous charge loss in flash memories. *IEEE Trans. Electr. Devices* 51, 9 (Aug. 2004), 1392–1400.

[12] Peter Desnoyers. 2010. Empirical evaluation of NAND flash memory performance. *ACM SIGOPS Operat. Syst. Rev.* 44, 1 (Jan. 2010), 50–54.

[13] Ralph Howard Fowler and L. Nordheim. 1928. Electron emission in intense electric fields. In *Proceedings of Royal Society of London A* 119, 781 (May. 1928), 173–181.

[14] Andrea Ghetti, Christian Monzio Compagnoni, Alessandro S. Spinelli, and Angelo Visconti. 2009. Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer flash memories. *IEEE Trans. Electr. Devices* 56, 8 (Jul. 2009), 1746–1752.

[15] Laura M. Grupp, Adrian M. Caulfield, Joel Coburn, Steven Swanson, Eitan Yaakobi, Paul H. Siegel, and Jack K. Wolf. 2009. Characterizing flash memory: Anomalies, observations, and applications. In *Proceedings of the Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'09)*. 24–33.

[16] Laura M. Grupp, John D. Davis, and Steven Swanson. 2013. The Harey tortoise: Managing heterogeneous write performance in SSDs. In *Proceedings of the USENIX Annual Technical Conference*. 79–90.

[17] International Data Corporation. Where in the world is storage. Retrieved from https://www.idc.com/IDC_Storage-infographic.jsp.

[18] Myoungsoo Jung, Wonil Choi, Shekhar Srikantaiah, Joonhyuk Yoo, and Mahmut T. Kandemir. 2014. HIOS: A host interface I/O scheduler for solid state disks. In *Proceedings of the Annual International Symposium on Computer Architecture (ISCA'14)*. 289–300.

[19] Jaeho Kim, Jongmin Lee, Jongmoo Choi, Donghee Lee, and Sam H. Noh. 2013. Improving SSD reliability with RAID via elastic striping and anywhere parity. In *Proceedings of the Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'13)*. 1–12.

[20] Kinam Kim. 2008. Future memory technology: Challenges and opportunities. In *Proceedings of the International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA'08)*. 5–9.

[21] Chang-Hyun Lee, Jungdal Choi, Changseok Kang, Yoocheol Shin, Jang-Sik Lee, Jongsun Sel, Jaesung Sim, Sanghun Jeon, Byeong-In Choe, Dukwon Bae1, Kitae Park, and Kinam Kim. 2006. Multi-level NAND flash memory with 63 nm-node TANOS (Si-Oxide-SiN-Al2O3-TaN) cell structure. In *Digest of Technical Papers of the IEEE Symposium on VLSI Technology*. 21–22.

[22] Jae-Duk Lee, Chi-Kyung Lee, Myung-Won Lee, Han-Soo Kim, Kyu-Charn Park, and Won-Seong Lee. 2006. A new programming disturbance phenomenon in NAND flash memory by source/drain hot-electrons generated by GIDL current. In *Proceedings of the IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW'06)*. 31–33.

[23] Kyunghwan Lee, Myounggon Kang, Seongjun Seo, Dong-Hua Li, Jungki Kim, and Hyungcheol Shin. 2013. Analysis of failure mechanisms and extraction of activation energies (E_a) in 21-nm NAND flash cells. *IEEE Electr. Device Lett.* 34, 1 (Jan. 2013), 48–50.

[24] Rino Micheloni. 2016. *3D Flash Memories*. Springer.

[25] Rino Micheloni, Luca Crippa, and Alessia Marelli. 2010. *Inside NAND Flash Memories*. Springer Science & Business Media.

[26] Krishna Parat and Chuck Dennison. 2015. A floating gate based 3D NAND technology with CMOS under array. In *Technical Digest of the International Electron Devices Meeting (IEDM'15)*. 48–51.

[27] Ki-Tae Park, Myounggon Kang, Doogon Kim, Soon-Wook Hwang, Byung Yong Choi, Yeong-Taek Lee, Changhyun Kim, and Kinam Kim. 2008. A zeroing cell-to-cell interference page architecture with temporary LSB storing and parallel MSB program scheme for MLC NAND flash memories. *IEEE J. Solid-State Circ.* 43, 4 (Mar. 2008), 919–928.

[28] Moon-Sik Seo, Bong-Hoon Lee, Sung-Kye Park, and Tetsuo Endoh. 2012. Novel concept of the three-dimensional vertical FG NAND flash memory using the separated-sidewall control gate. *IEEE Trans. Electr. Devices* 59, 8 (Jun. 2012), 2078–2084.

[29] Kang-Deog Suh, Byung-Hoon Suh, Young-Ho Lim, Jin-Ki Kim, Young-Joon Choi, Yong-Nam Koh, Sung-Soo Lee, Suk-Chon Kwon, Byung-Soon Choi, Jin-Sun Yum, Jung-Hyuk Choi, Jang-Rae Kim, and Hyung-Kyu Lim. 1995. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme. *IEEE J. Solid-State Circ.* 30, 11 (Aug. 1995), 1149–1156.

[30] Sung-Jin Whang, Ki-Hong Lee, Dae-Gyu Shin, Beom-Yong Kim, Min-Soo Kim, Jin-Ho Bin, Ji-Hye Han, Sung-Jun Kim, Bo-Mi Lee, Young-Kyun Jung, Sung-Yoon Cho, Chang-Hee Shin, Hyun-Seung Yoo, Sang-Moo Choi, Kwon Hong, Seiichi Aritome, Sung-Ki Park, and Sung-Joo Hong. 2010. Novel 3-dimensional dual control-gate with surrounding floating-gate (DC-SF) NAND flash cell for 1Tb file storage application. In *Technical Digest of the International Electron Devices Meeting (IEDM'10)*. 668–671.

[31] Yeong-Jae Woo and Jin-Soo Kim. 2013. Diversifying wear index for MLC NAND flash memory to extend the lifetime of SSDs. In *Proceedings of the International Conference on Embedded Software (EMSOFT'13)*. 1–10.

[32] Qin Xiong, Fei Wu, Zhonghai Lu, Yue Zhu, You Zhou, Yibing Chu, Changsheng Xie, and Ping Huang. 2017. Characterizing 3D floating gate NAND flash. In *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. 32–33.

[33] Hyun-Seung Yoo, Eun-Seok Choi, Han-Soo Joo, Gyu-Seog Cho, Sung-Kye Park, Seiichi Aritome, Seok-Kiu Lee, and Sung-Joo Hong. 2011. New read scheme of variable Vpass-read for dual control gate with surrounding floating gate (DC-SF) NAND flash cell. In *Proceedings of the IEEE International Memory Workshop (IMW'11)*. 1–4.

[34] Kai Zhao, Wenzhe Zhao, Hongbin Sun, Tong Zhang, Xiaodong Zhang, and Nanning Zheng. 2013. LDPC-in-SSD: Making advanced error correction codes work effectively in solid state drives. In *Proceedings of the USENIX Conference on File and Storage Technologies (FAST'13)*. 243–256.

[35] You Zhou, Fei Wu, Ping Huang, Xubin He, Changsheng Xie, and Jian Zhou. 2015. An efficient page-level FTL to optimize address translation in flash memory. In *Proceedings of the European Conference on Computer Systems (EuroSys'15)*. 1–16.

Received June 2017; revised September 2017; accepted November 2017