Shaving Retries with Sentinels for Fast Read over High-Density 3D Flash

Qiao Li
City University of Hong Kong

Min Ye
City University of Hong Kong

Yufei Cui
City University of Hong Kong

Liang Shi
East China Normal University

Xiaoqiang Li
YEESTOR Microelectronics Co., Ltd

Tei-Wei Kuo
City University of Hong Kong

Chun Jason Xue
City University of Hong Kong

***Abstract*—High-density flash-memory chips are under tremendous demands with the exponential growth of data. At the same time, the slow read performance of these high-density flash-memory chips becomes a new challenge. In this work, we analyze the high raw bit error rates (RBER) issue by characterizing the error behaviours of 3D QLC flash-memory chips. A preferred read voltage to a QLC cell could vary among layers and might even change in a short period of time due to the temperature. A sentinel-cell approach is thus proposed to utilize the error characteristics among cells. We propose to infer the optimal read voltages of a wordline based on errors introduced on sentinel cells. An on-line calibration procedure is further presented to resolve the problem of possible non-uniform error distribution on some wordlines. With optimal voltages being inferred, the number of read retries will be significantly reduced. Experiments show that optimal read voltages can be instantly obtained in 94% cases on average over the evaluated QLC flash memory with at most 2 read retries, and with merely 0.2% space overheads for adopting sentinel cells. The number of read retries could be reduced by 82% on average, and the read performance can be improved by 74% on average through a series of extensive experiments over 3D TLC and QLC flash-memory chips.**

***Index Terms*—3D NAND flash, read retry, reliability**

I. Introduction

NAND flash memory is now the storage media of mobile devices and also heavily adopted in data servers. As the amount of data in applications continues to increase, the demand for greater-capacity NAND flash memory continues to grow. To increase the density and capacity of flash memory, high-density 3D flash-memory chips are replacing planar flash chips in storage systems [1][2][3][4][5]. Currently, 3D triple-level cell (TLC) flash memory with three bits per cell is the mainstream [6][7][8], and manufacturers will supply quad-level cell (QLC) flash memory soon by storing four bits per cell [9][10][4]. The pursuit of high-density and low-cost flash memory makes sacrifice on the reliability of flash memory. The raw bit error rates (RBER) of today's flash-memory chips are around 10^{-3} to 10^{-2}, which introduces new challenges to the performance of flash memory [11][12][13][14][15][16]. To guarantee data correctness in flash memory, the cost of error correction codes (ECC) grows quickly along with the increasing density and the shrinking size of flash-memory chips [17][18][19]. A read failure happens when a read operation with the default read voltages returns an RBER beyond the ECC capability [20][3][21][22]. After a read failure, the read voltages are tuned, followed by a read retry. Due to charge leakage or various disturbances, flash cells' voltage states might gradually shift over time. It is challenging, if not impossible, to estimate the amount of voltage shifts. Hence, multiple read retries are often applied, and the read performance could deteriorate significantly. MSB pages of high-density flash-memory chips are particularly vulnerable, as multiple read voltages are required for a single page read. A successful read needs to tune all the read voltages to proper positions.

A series of work studies the error characteristics of multi-level-cell (MLC) and TLC flash, where each cell stores two bits and three bits respectively, and developed error models. Based on the models, the shift of voltage states is estimated [23][24][25][26][27]. Accurate estimation of voltage state shifting is challenging, which requires large overheads to record additional information. It's more challenging on 3D NAND where errors are more diverse. Another method is to keep track of the optimal read voltages that generate the lowest RBER [28][29][30]. The overhead to periodically update the optimal read voltages at required granularity are prohibitively high. These approaches could get desired results on planar flash memory, but not on 3D flash due to different error characteristics [31][17][18]. For 3D flash-memory chips, Shim et al. [32] studied the process similarity among wordlines inside a layer and presented that all the wordlines inside the same layer have the same optimal read voltages over TLC flash-memory chips in their experiments. As presented in their work, significant variations exist among layers, and the information for each layer is recorded in the flash translation layer (FTL). Besides, it takes a high-cost approach similar to previous work [28][29][30] to find the optimal read voltages in the first place for each layer.

To avoid the high timing overheads in finding the optimal read voltages as previous approaches and the high space overheads in FTL for maintaining necessary information, this paper proposes an efficient and effective approach to infer the optimal read voltages on high-density 3D flash memory. In particular, a small set of cells is reserved as sentinel cells on each wordline, and one read voltage among all read voltages is selected as the sentinel voltage. The underlying rationale is

978-1-7281-7383-2/20/$31.00 ©2020 IEEE
DOI 10.1109/MICRO50266.2020.00048

that the voltage shift of the whole wordline is often consistent with that of sentinel cells even with only a small set of cells reserved. In addition, the optimal values of all the read voltages have strong correlations among each other. During write operations, we program sentinel cells to the two voltage states on the two sides of the sentinel voltage. Thus after a read operation, the exact errors of sentinel cells are known by comparing the original data and the readout data. The optimal value of the sentinel voltage can be estimated based on the error information. In this way, *cells' read-voltage information could stay with the cells of the same wordline, and no extra read or write overheads occur.* By comparison, previous work requires huge overhead to read and update extra information maintained in the FTL. After estimating the optimal value of the sentinel read voltage, the optimal value of the other read voltages can also be estimated by leveraging their correlation. Experiments on real QLC flash chips verify that the proposed sentinel-assisted scheme can find the optimal read voltages for 83% cases on average. This work further proposes a calibration method based on the difference in the first two reads (one read with the default read voltages and the other with the first inferred read voltages), which can identify the optimal read voltages for 11% more cases on average. Based on the evaluations on real flash chips, the proposed approach can reduce the number of read retry operations by 82% on average, and the read performance can be improved by 74% on average.

The major contributions of this work are as follows:

- The errors and the optimal read voltages of 3D TLC and QLC flash-memory chips are explored. We show that it is challenging to estimate the optimal read voltages of 3D flash under the impact of P/E cycles, retention time, and the high variations among layers. Optimal read voltages could also change dramatically in a short period of time under the influence of temperature.
- A sentinel-cell approach is proposed based on our observation that errors are nearly uniformly distributed along a wordline, and the optimal values for read voltages were found with a strong linear correlation among each other. Under the proposed approach, the number of read retry operations is be minimized by inferring the optimal read voltages based on the errors from sentinel cells without run-time overhead, followed by a calibration procedure.
- Experiments were conducted over real 3D TLC and QLC flash-memory chips to show the effectiveness of the proposed approach. As shown by the results, by reserving merely 0.2% cells on each wordline, the number of read retries can be reduced from 6.6 to 1.2 on average, which also improves the read performance by 74% on average.

II. Background and Motivation

A. Basics of NAND Flash Memory

A NAND flash memory chip contains thousands of flash blocks, which are two-dimensional (planar) or three-dimensional (3D) arrays of flash cells. Inside a flash block, flash cells are organized into many rows. Each row constitutes a wordline. A flash cell is a floating gate or a charge trap transistor to represent data by storing a certain amount of charges, which determines the threshold voltage (V_{th}) of the cell. Current flash memory stores multiple bits in each cell. Figure 1 shows an example of the V_{th} distribution of triple-level cell (TLC) flash. Each TLC cell stores three bits, most significant bit (MSB), center significant bit (CSB) and least significant bit (LSB). The LSB (CSB/MSB) of all cells in one wordline form the LSB (CSB/MSB) page of that wordline.

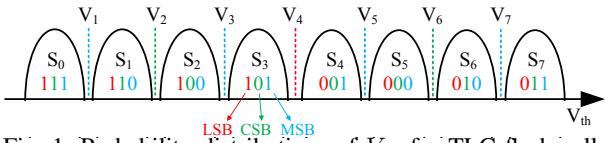

Fig. 1: Probability distributions of V_{th} for TLC flash cells.

To read a particular page from a flash block, the controller applies one or several read voltages to the control gates of all the cells on the wordline that contains the page. As shown in Figure 1, to read an LSB page, V_4 will be applied. For all the cells, if $V_{th} < V_4$, the reading bit is 1, otherwise, it is 0. To read a CSB page, two voltages, V_2 and V_6 will be applied. If $V_{th} < V_2$ or $V_{th} > V_6$, the reading bit is 1, otherwise, it is 0. To read an MSB page, the remaining four read voltages are applied to differentiate 1 and 0. In this work, we target both TLC and quad-level cell (QLC) flash. In QLC flash, up to eight voltages are used to read the MSB page. The details of the voltage distributions on QLC are not presented here as QLC is similar to TLC, except for more bits and more voltage states per cell.

The V_{th} of flash cells could be biased owing to different error sources [33][25][34]. For example, program and erase operations wear flash cells and damage the strength of charge trapping over time. Charge leakage over retention time and read disturb will become more severe with increased program/erase (P/E) cycles. Thus, the V_{th} in one state may shift to other states, which introduces errors on the stored data. To guarantee reliability, error correction codes (ECC) are adopted in NAND flash memory. The user data will be stored together with the parity generated from ECC encoding, where the combination of user data and parity is called one ECC frame. ECC parity is stored in the out-of-band (OOB) area, which contains additional spare bytes within a page. ECC can correct a limited number of errors. When the RBER of data exceeds the error correction capability, read retry operations will be applied to reduce RBER.

Due to charge leakage over retention time or disturbance during flash operations, the threshold voltage distribution could shift and may cross the default read voltages. In this case, a cell could be misread to a different value. If ECC fails to decode the data, the data read will be retried with tuned read voltages. Read retry operations will be conducted until a successful read, or a read failure when reaching the maximum number of read retries. The left figure in Figure 2 shows the number of bit errors given a certain voltage offset, illustrated in the right figure. The x-axis shows the offset to the default

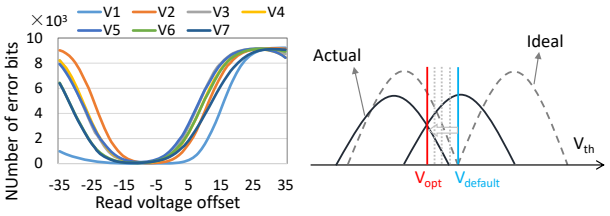

Fig. 2: The illustration of read retry with tuned read voltages.

read voltage, and positive/negative values represent tuning to higher/lower voltages. As presented in Figure 2, the number of bit errors heavily depends on the read voltage. There exists one optimal voltage which will introduce the lowest RBER for read voltage. The number of bit errors is smaller when the read voltage is closer to the optimal read voltage.

B. Observation and Motivation

In this section, experiments are first conducted on real high-density 3D TLC and QLC flash-memory chips to study the characteristics of RBER and read voltages. The detailed information and parameters of the evaluated flash-memory chips can be found in the experiment section. We present several observations on the optimal read voltages and RBER, which motivate us to propose a novel scheme to effectively minimize the number of read retry operations on high-density 3D NAND flash.

1) Impact of Optimal Read Voltage on RBER: We first present how optimal read voltages affect RBER. Instead of absolute voltage values, results are presented with normalized voltages. The offset to the default read voltage is used to represent the value of the optimal read voltage. The value is positive if the optimal read voltage is higher than the default read voltage, and vice versa. Figure 3 shows the RBERs of MSB page on the evaluated TLC and QLC flash chips, after one-year retention time with different numbers of P/E cycles. MSB page has the highest RBER among all the pages sharing the same wordline, and thus is used as an example to show the results. One layer consists of multiple wordlines, and data presented in the figures are the maximum RBER of each layer. From the results, we made the following two observations.

First, reading data with the optimal read voltages significantly reduces the RBER, especially for QLC flash memory due to its high error rate. The RBERs can be reduced by an order of magnitude for some layers on both TLC and QLC flash chips. For example, when P/E cycle is 5000, the RBER at default read voltages is near 10^{-1}, which is much higher than the error correction capability of the current ECC. However, the RBER at the optimal read voltages is less than 10^{-2}, which is within the correction capability.

Second, the optimal read voltages greatly reduce the variations of RBER among layers. When using the default read voltages, the RBER presents huge variations among layers, and the variations on the QLC flash chip are even more significant than TLC flash. The phenomenon has been well studied in existing work [35][18][36]. At the optimal read voltages, the variations are reduced. Even the highest RBER at optimal read voltages is low, compared with all RBER values at default read voltages.

2) Impact of Temperature: Existing work has presented that the reliability of flash memory chips would be affected by the temperature [25]. To study the temperature effects on high-density 3D flash memory, we evaluated the differences after retention time of one hour under the room temperature and high temperature. For room temperature, the flash chips are placed on the desk of an office with air conditioning, where the room temperature is 25°C. For high temperature, the flash chips are placed within a computer case, and the computer is running a program with a heavy computation workload, where the temperature is about 80°C. These two cases are possible working environments of flash memory chips. Figures 4 and 5 show the comparison of RBERs and optimal read voltages, respectively. The results of the QLC flash chip are presented, and the results on TLC are omitted due to the page limit. Only four read voltages are presented for the same reason and the results on the remaining read voltages show a similar trend.

Compared with the room temperature, the RBERs of the chips in the computer case are higher, and the optimal values of read voltages are smaller. The reason is that high temperatures can accelerate the retention effects of flash memory. Under high temperature, the charge leakage due to retention effects is accelerated and the voltage states shift to the left at a higher speed. Thus, there are more errors and the optimal read voltages shift to the left correspondingly. As observed in Figure 4, with retention time of only one hour, the RBERs under high temperature are greatly different from that under room temperature, so are optimal read voltages as shown in Figure 5. This indicates that the optimal read voltages will change sharply in a short time period if the working temperature is high. Note that the current method [28] by tracking the optimal read voltages requires updating the recorded optimal read voltages every 24 hours. This tracking method is not applicable for high working temperature. This characteristic introduces further challenges on accurate estimation of the optimal read voltage of flash memory in case of high temperature. The overheads would become higher and the accuracy would drop significantly.

3) Challenges: From the above results, we observe that applying optimal read voltages brings considerable benefits through reducing both RBER and the variations of RBER among layers. However, under the influence of temperature, the optimal read voltages may change dramatically within a short period of time, which makes it impossible to track the optimal read voltages. Moreover, the problem of finding optimal read voltages is challenging because multiple voltages are required to read a page in high-density flash memory. For example, four read voltages are used to read the MSB page of TLC in Figure 1. Any misplaced voltage will result in high RBER and a large number of read retries. The problem becomes even more severe as the number of bits per cell increases in flash memory, which increases the number of read voltages for read operations. In this paper, we propose a method that accurately estimates the optimal read voltages

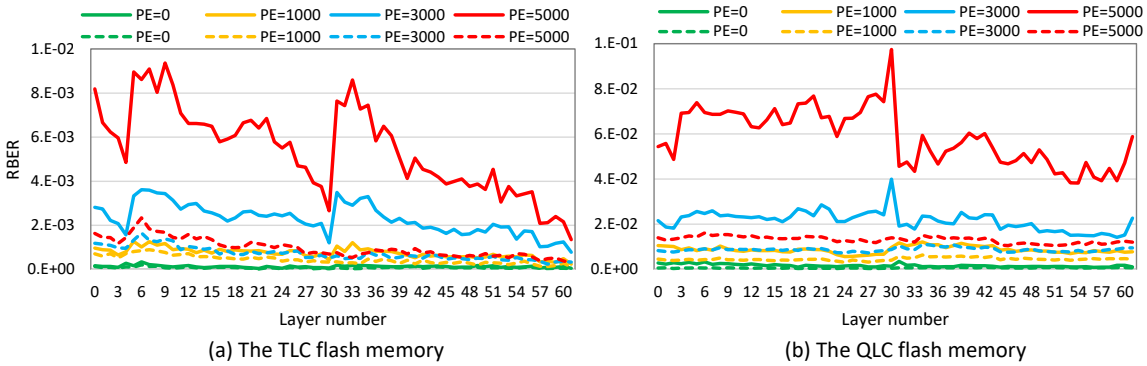

Fig. 3: The MSB RBERs of the evaluated flash chips at default read voltages (solid lines) and at optimal read voltages (dashed lines) for all layers inside one flash block.

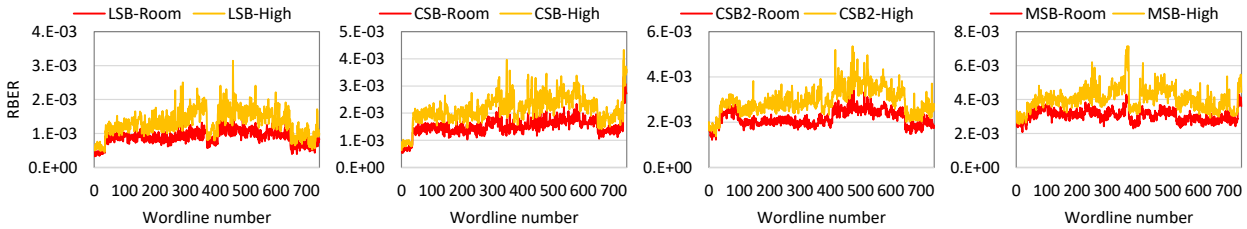

Fig. 4: The RBERs of four pages after retention time of one hour at room temperature (-Room) and at high temperature in a computer case (-High).

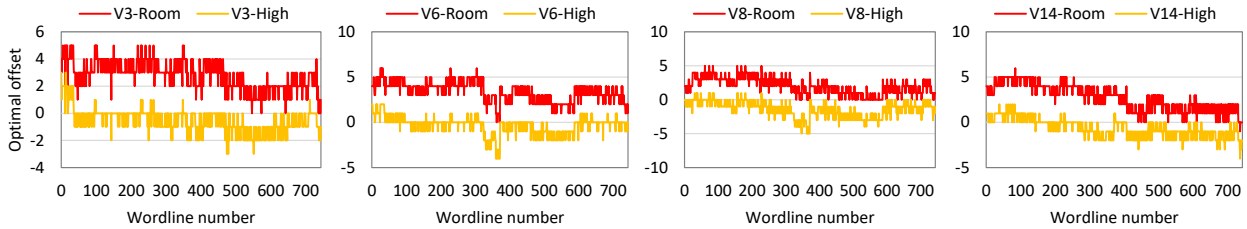

Fig. 5: The optimal value of randomly-selected four read voltages after retention time of one hour at room temperature (-Room) and at high temperature inside a computer case (-High).

with a minimized number of read retries.

III. Shaving Read Retries with Sentinels

In this section, we propose a method using sentinels to minimize the number of read retries. First, we present an approach that preserves sentinel cells on each wordline and selects a sentinel voltage among all read voltages. We then infer the optimal read voltages from the errors shown on sentinel cells as errors are nearly uniformly distributed on the cells along each wordline. Finally, a calibration approach is proposed to calibrate the optimal read voltages.

A. Sentinel Cell and Sentinel Voltage

1) Sentinel Cell: First, we present the definition of sentinel cells and the reason why it works. The error information and the optimal read voltages collected from real 3D flash chips are analyzed. The optimal read voltages of the evaluated QLC flash chip are shown in Figure 6. The number of P/E cycles is 3000, and the retention time is 1 year. The y-axis shows the offset of the optimal read voltages to the default read voltages. Since the erase state (S_0) has a wide distribution and thus the read voltage V_1 to differentiate S_0 and S_1 presents a much larger variation among layers within a block, we only present the rest read voltages in the figures. Similar to the RBER, the optimal read voltages of different layers also present great variations. Regarding the same read voltage, for example, V_{15} to differentiate the last two voltage states, S_{15} and S_{16}, the optimal values range from -9 to 0 among different layers. This characteristic also results from the variations among layers in 3D flash blocks. Therefore, the optimal read voltages of one wordline may introduce high RBER on other wordlines within the same block.

We further present the positions of the error cells inside one block of the evaluated QLC flash memory, as shown in Figure 7. Each blue dot represents a cell that has at least one bit error. We can make two observations via the density of blue dots from the figure. First, the dark and light horizontal stripes show the RBER varies greatly among different wordlines, and thus the optimal read voltages of different wordlines are different, as presented above. Therefore, applying the optimal

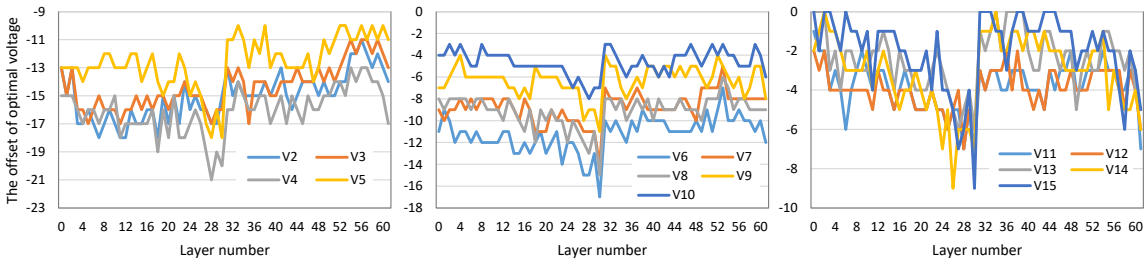

Fig. 6: The optimal read voltages of different layers within a block on the QLC flash chip.

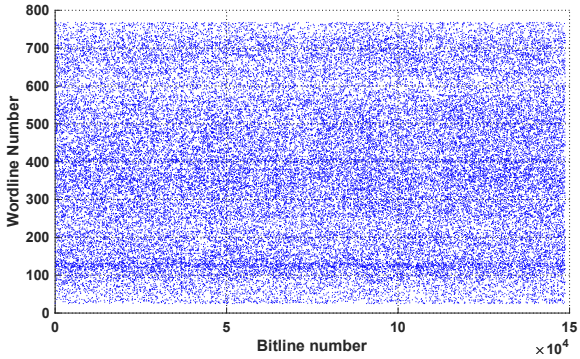

Fig. 7: The position of bit errors inside one flash block, which has endured 3000 P/E cycles and one-year retention time.

read voltages of one wordline to other wordlines inside one block, as proposed in previous work [28][25], is not applicable on 3D NAND flash memory. In other words, **the error characteristics of one wordline cannot represent the error characteristics of other wordlines inside a block.** Second, the errors are almost uniformly distributed along a wordline. This indicates that the errors on wordlines present locality. The rationale behind the locality is because there are two types of voltage shift. First, the main shift due to error sources, including P/E and retention time, is global for all the cells on the wordline. The cells on the same wordline have endured the same P/E cycles, retention time, and temperature. Second, there is a local drift for each cell due to random factors. Since global drift is much more significant, the errors of a small set of cells can predict the errors on the wordline. Therefore, we propose to reserve a small subset of cells on each wordline as sentinel cells to infer the optimal read voltages for the whole wordline. As evaluated in the experiment section, reserving 0.2% cells can provide an accurate inference of the errors and optimal read voltage.

2) Sentinel Voltage: We further propose to select one read voltage as sentinel voltage by exploiting the correlations among optimal read voltages. The rationale is that all the voltage states on the same wordline have endured the same wearing and retention time. It is expected that the optimal read voltages have some correlations with each other.

The optimal read voltages of all wordlines from multiple blocks under different P/E cycles and retention time are collected. V_8 of the QLC flash is used as an example to show the relationship. Figure 8 shows the correlation between the optimal value of V_8 and the optimal values of other read voltages on QLC. The optimal offset of every read voltage shows a nearly linear relationship with the optimal offset of V_8, as most points in Figure 8 spread along a straight line. Besides the presented V_8 voltage, the optimal offset of other pairs of read voltages also presents a similar strong linear correlation. This characteristic indicates that once the optimal offset of one read voltage, for example, V_8 is achieved, we can infer the optimal value of other read voltages. The same evaluations were performed on several flash chips with the same type and the same production batch. The linear correlation on one flash chip can be well applied to other chips. This property leads to the option of picking one read voltage as sentinel voltage and relieves the need to track all read voltages.

According to the above results, we can conclude that: 1) the optimal read voltages of a small set of cells can predict the optimal read voltages of all the cells on the same wordline; 2) the optimal value of one read voltage between a pair of adjacent voltage states can predict the optimal value of other read voltages on the wordline.

B. Inference of Optimal Read Voltages

To minimize the number of read retries, we propose to directly infer the optimal voltages with sentinel cells and a sentinel voltage. This work uses V_4 as the sentinel voltage for TLC flash memory and V_8 as the sentinel voltage for QLC flash memory. As all the optimal read voltages have a strong correlation between each other, other read voltages could also be selected as the sentinel voltage.

During write operations, sentinel cells will be evenly programmed to the two voltage states between which the sentinel voltage lies, i.e., S_3 and S_4 for TLC, S_7 and S_8 for QLC. During read operations, we can count the number of bit errors on sentinel cells by comparing the original data and readout data. Thus the optimal offset of the sentinel voltage, V_4 for TLC and V_8 for QLC, can be estimated for sentinel cells, which is also the inferred optimal V_4 (V_8 for QLC) for the wordline. Then, based on the correlation between optimal read voltages, we can calculate the optimal value of other read voltages from the optimal value of V_4 (V_8 for QLC). By finding the optimal read voltages quickly and accurately, the number of read retries is minimized. In the following, TLC is used as an example to present the proposed mechanism for a simple description. The method is widely applicable to different types of NAND flash memories.

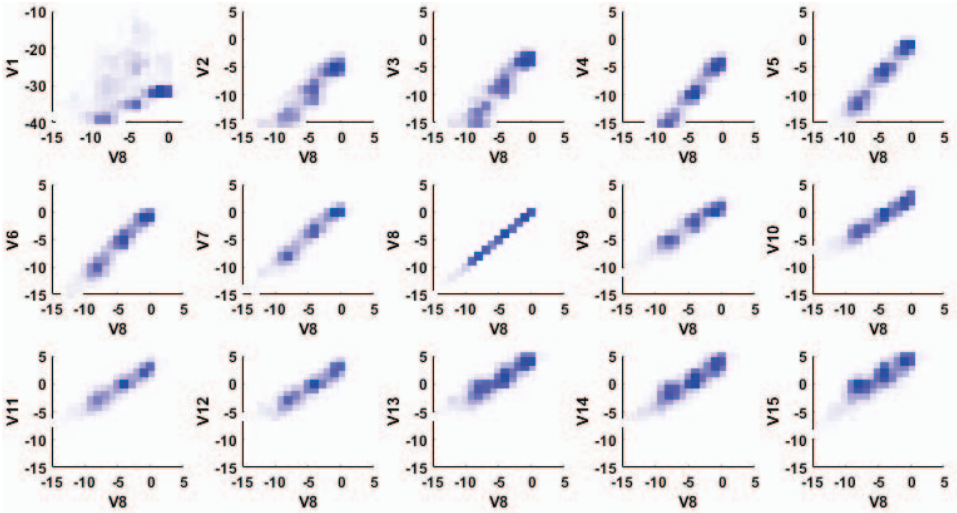

Fig. 8: The correlation between the optimal value of each read voltage and the optimal value of V_8 on the QLC flash chip. Darker square means more data at that point.

1) Error Difference : Suppose V_i is selected as sentinel voltage, we first define two types of errors for V_i. As shown in Figure 9, V_i is applied between two neighboring states, S_{i-1} and S_i, to differentiate these two states. The errors resulting from cells in S_{i-1} being misread as S_i are called *up errors* while those resulting from cells in S_i being misread as S_{i-1} are called *down errors*. The error difference d is calculated as the number of up errors minus the number of down errors for the sentinel voltage. The value of d indicates the V_{th} transformation of the two states at the two sides of the sentinel voltage. For example, in Figure 9, after the left shift of S_{i-1} and S_i, there are more down errors (instances of S_i below the threshold V_i) than up errors (instances of S_{i-1} above V_i), and d is smaller than 0. The optimal read voltage is lower than the default read voltage.

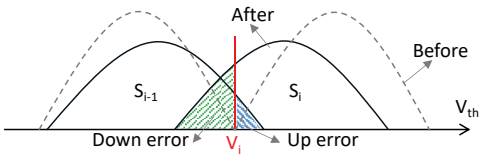

Fig. 9: The illustration of up errors and down errors.

2) Optimal Read Voltage Inference: To serve a read request, a set of default read voltages is applied on the wordline. On TLC flash chips, the voltage sets are $\{V_4\}$, $\{V_2, V_6\}$ and $\{V_1, V_3, V_5, V_7\}$ for LSB, CSB and MSB page read respectively. If a read fails due to too many errors, a read retry is required with tuned values of read voltages. At this point, we calculate the error difference of sentinel cells, d, based on the original programmed data and readout data on sentinel cells. If the read page is an LSB page, the default value of read voltage V_4, which is also the sentinel voltage, is applied in the failed read operation. In this case, up errors and down errors of sentinel cells can be directly counted, where the up errors are the cells originally programmed to S_3 but being misread as '0', similarly for down errors. If the read page is a CSB page or an MSB page, the sentinel voltage has not been applied and thus we need an extra read operation by applying the sentinel voltage to count the errors on sentinel cells. In the presented coding scheme of the evaluated flash chips in this work, this is also an LSB page read. The latency of this extra read operation is much less than a read retry operation, because read latency is proportional to the number of read voltages. Only one read voltage is applied for this extra read, while one read retry operation for CSB page or MSB page needs multiple read voltages.

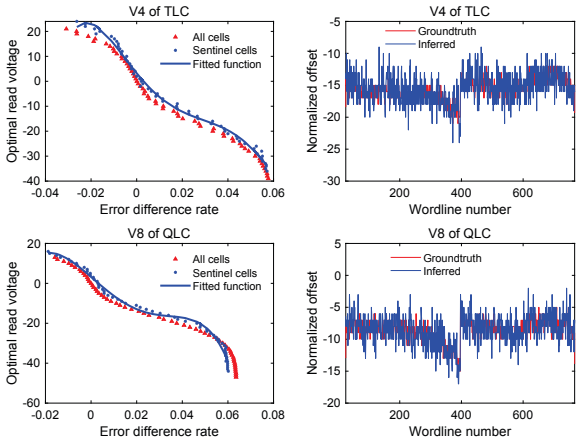

Fig. 10: The result of curve fitting (left) and the inferred optimal read voltage (right) for V_4 of TLC and V_8 of QLC.

Based on the error difference d, the optimal offset of the sentinel voltage on sentinel cells can be inferred. By collecting hundreds of pairs of d and $V_{optimal}$, we use polynomial function f to fit the relationship between the two, which is $V_{optimal} = f(d)$. Figure 10 shows an example of the fit result (left) and the predicted optimal voltage (right). The

relationship f in Figure 10 is well fitted as a polynomial function of degree 5. The right figure shows that the inferred optimal read voltage is equal to or very close to the real optimal voltage.

After computing the optimal offset of the sentinel voltage, the optimal offset for other read voltages can be inferred based on their linear correlations, as presented in Figure 8. Note that the correlations in Figure 8 are obtained from data collected from two flash chips of the same product. The correlation is common for all the flash chips of the same batch. Therefore, during the manufacturing process, we can conduct evaluations on one or several flash chips to collect data for the correlation. Then the correlation can be written into all the chips of the same batch without the need of evaluations on every chip.

C. Calibration of Inferred Read Voltage

The inference of optimal read voltages can achieve a high success rate as the errors are uniformly distributed along a wordline for most cases, as shown in Figure 7. Nevertheless, the inference may fail, which results from the difference between sentinel cells and all cells. Some sentinel cells cannot exactly represent the whole wordline. To deal with the read failures after the inference, we propose to calibrate the optimal offset of the sentinel voltage by finding the difference between sentinel cells and all cells. As Figure 10 shows, the direction of the inferred voltage is always correct, and the inferred read voltage is very close to the real optimal read voltage. Therefore, the difference between the sentinel cells and other cells is not large. We consider two inference failure cases as shown in Figure 11: 1) the inferred tuning offset is smaller than the real optimal offset and the read voltage should be tuned more in the same direction; 2) the inferred tuning offset is greater than the real optimal offset and the read voltage should be tuned back a little bit.

The challenge lies in how to differentiate these two cases with error information only provided by sentinel cells, while the error information on other cells within the wordline is *unknown*. We study the two failure cases and propose a unified calibration method.

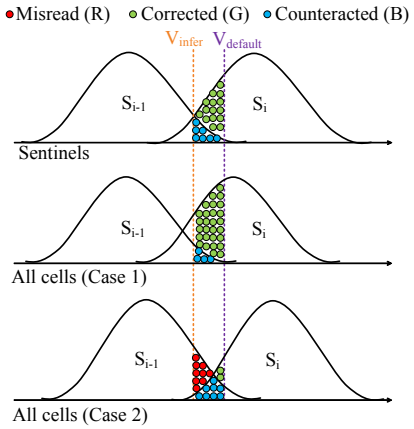

Fig. 11: Two potential cases of inference failures.

1) Analysis of Cell State Changing: As shown in Figure 11, when the read voltage is tuned from $V_{default}$ to V_{infer}, the read states of the cells between these two voltages will be changed from S_i to S_{i-1}. These cells consist of three types of cells: misread cells (R, red dots in Figure 11), corrected cells (G, green dots), and counteracted cells (B, blue dots). Misread cells are in the correct state S_{i-1} with $V_{default}$ but being misread as S_i with V_{infer}. Corrected cells are in the wrong state S_i with $V_{default}$ and are corrected by V_{infer}. Counteracted cells are in the overlapped area, and moving read voltage from $V_{default}$ to V_{infer} generates the equal number of cells being corrected and being misread in this area.

Based on Figure 11, the number of sentinel cells with state changed (denoted as NC_s) equals to the number of corrected cells (G) plus double number of counteracted cells (B), i.e., $NC_s = G + 2B$. Similarly, for Case 1 of all the cells, the number of cells with state changed (denoted as NC_a^1) can be calculated as $NC_a^1 = G + 2B$. While for Case 2 of all the cells, $NC_a^2 = G + 2B + R$. In the following, we discuss the comparison of the three distributions in Figure 11. To make a fair comparison on the numbers, NC_s will be divided by the reserving ratio.

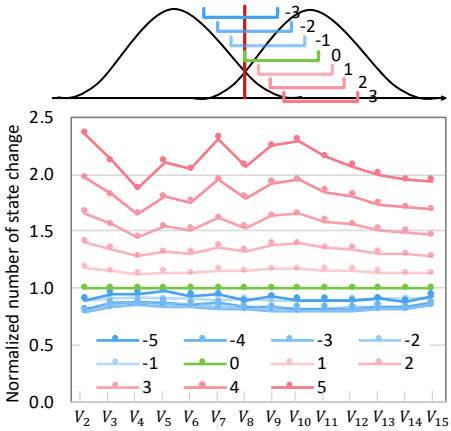

Fig. 12: The comparison of number of state-changing cells.

Compared to the distribution of sentinel cells, the cells with state changed in Case 1 are closer to the mean of the distribution. Therefore, with the same distance, the integral of the area between V_{infer} and $V_{default}$ is supposed to be larger, thus $NC_a^1 > NC_s/r$, where r is the reserving ratio of sentinel cells. In contrast, the area of Case 2 is further away from the mean than that of sentinel distribution, and thus $NC_a^2 < NC_s/r$. To verify this relationship, we conduct experiments on real flash chips. Figure 12 shows the comparison of the number of state-change cells by changing the position of the area. The numbers in the legend represent the position offset to the real optimal read voltage. The positive items in the legend represent Case 1 in Figure 11 and the negative items represent Case 2. The numbers in the figure are divided by the number of 0 position offset (green curve), which represents a successful prediction. Then, these numbers are the normalized average state-change numbers of all the

wordlines inside one block. The result in Figure 12 verifies the relationship of the number of state-changing cells, which is, the average number of Case 2 is greater than that of the successful prediction, while the number of Case 1 is less than it.

2) Calibration based on State Changing: Based on the above observation, we propose to calibrate the offset of the sentinel voltage when the read retry with the inferred voltages fails. The basic idea is to compare the number of cells whose states have changed after tuning the read voltage from $V_{default}$ to V_{infer}. If $NC_a > NC_s/r$ which fits Case 1 in Figure 11, V_{infer} of the sentinel voltage should be tuned further in the same direction. Otherwise, it fits Case 2 and the V_{infer} should be tuned in the opposite direction. In both cases, the tuning offset of read voltage is set as a small value Δ. As observed from the result in Figure 10, the inferred voltage is very close to the real optimal voltage. It is expected that an accurate calibration with a small offset Δ can tune the read voltages to the optimal read voltage.

D. Overhead Analysis and Discussion

The main overhead of the proposed approaches is the space taken up by sentinel cells on each wordline. Table I shows the statistics of the offset difference between the real and the predicted optimal sentinel voltage by adjusting the ratio of sentinel cells. The numbers in the table illustrate the difference of normalized voltages. Compared to the width of a voltage state, which is 256 for the TLC flash chip and 128 for the QLC flash chip, the average of offset difference in the table is very small. In other words, optimal read voltages of sentinel cells and those of all cells within the same wordline have a small difference. With more cells reserved as sentinel cells, the offset difference gets smaller, but the overheads become greater. In the following, the percentage of sentinel cells is set as 0.2%, considering the trade-off between accuracy and overheads.

Sentinel cells are stored in the OOB space, which is used to store ECC parity. Usually, ECC parity does not take up all the OOB space. If there are enough free cells to be reserved as sentinel cells, no extra overhead is introduced, and the error correction capability will not be impacted. All of the evaluated flash chips in this work have enough spare space for sentinel cells. Empirically, the OOB area takes up more than 10% of total wordline on average, while sentinel cells only need 0.2% of total wordline space.

We illustrate this using one of the tested chips in this work as an example. The page size is 18592 bytes, and the space to store user data is 16384 bytes, which leaves the OOB space 2208 bytes (11.9% of the total space). ECC parity takes 2016 bytes (10.9%). 192 bytes (1.0%) are available for sentinel cells which is much greater than the empirical value 0.2%. In the experiments, we will also evaluate the impacts if the sentinel cells take up the space of ECC parity.

The presented overheads stay with the data on the same wordline. Unlike previous methods where the optimal read voltages are maintained in the FTL [32][28][30], the error information shown on sentinel cells is accessed together with the normal read and write operations. No extra read or write operations will be introduced in the proposed approach.

TABLE I: The statistics of the offset difference between the real and the predicted optimal sentinel voltage.

Ratio of sentinel cells		0.02%	0.1%	0.2%	0.4%	0.6%
TLC	Average	2.35	1.77	1.60	1.50	1.44
	Standard deviation	1.81	1.47	1.29	1.19	1.16
QLC	Average	3.15	2.09	1.79	1.48	1.27
	Standard deviation	2.46	1.67	1.39	1.21	1.12

The relationship between the optimal read voltage and the error difference, as well as the correlations among optimal read voltages, will be stored inside flash chips. For each flash type, one or several flash chips are randomly selected for evaluation and analysis, where the error difference and optimal read voltage are collected to fit the relationships and the correlations. The temperature will not impact the relationship between the error difference rate and the optimal read voltage, while it does impact the correlation among optimal read voltages. For implementation, we maintain one table for the relationship between error difference and the optimal read voltage, and multiple tables to store the correlations among optimal read voltages, where each table corresponds to a temperature range. Afterwards, the relationships are programmed into all the flash chips of the same type. Our evaluation indicates that all the flash chips of the same type have similar reliability characteristics, with only marginal deviations due to process variation.

IV. Experiments

The proposed approach is evaluated on the YEESTOR 9083 SSD evaluation platform [37], which has a flash memory controller that enables users to implement custom flash memory solutions. Several TLC and QLC flash-memory chips from three different vendors are tested. In the following, the results of one TLC and one QLC flash chips from Micron [38] are presented. They both have 64 layers and the capacities are 64GB and 128GB for TLC and QLC, respectively.

The number of P/E cycles is set as 5000 for the TLC flash and 1000 for the QLC flash. The experiments are conducted on both flash chips with one-year retention time (under room temperature), where the retention time is accelerated by baking with high temperature. This case can also be considered as working under high temperature for dozens of hours, for example, in a computer case with running applications. We also conducted experiments to study read disturbance. On all the measured flash chips, read disturbance does not introduce reliability degradation until one million read operations. This means read disturbance is a less significant issue compared to retention and P/E cycles. Thus this work focuses on the evaluations of retention and P/E cycles.

A. Read Performance Improvement

To study the impact of the proposed approach on read performance, one block from the TLC flash chip is evaluated. The

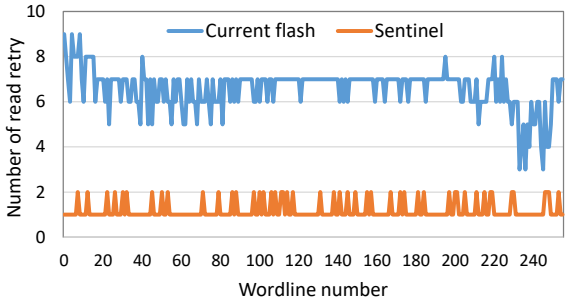

Fig. 13: The comparison of read retry number on the evaluated TLC flash memory chip.

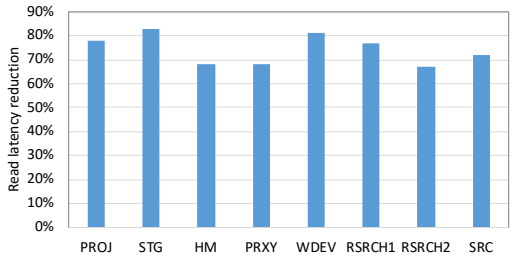

Fig. 14: The read latency reduction percentage of the proposed method on TLC flash.

evaluated QLC flash chip is not equipped with optimization approaches to tune the read voltages or predict the optimal read voltages. Thus we cannot evaluate the performance improvement on QLC flash chips at this point, while the evaluations on other aspects will be presented. The block of the evaluated TLC has endured for 5000 P/E cycles and the retention time is one year, which is accelerated by high temperature. As shown in Figure 13, we compare the number of read retry operations between the default methods in current flash chips ("current flash") and the proposed method ("sentinel"). In the current flash, many wordlines need more than 5 read retry operations for a successful read due to the high RBER. While the proposed sentinel method can accurately derive the optimal read voltages based on the information obtained from the default read for most cases. Compared with the current flash, sentinel based approach reduces the number of read retry by 82%, from 6.6 to 1.2 on average. For the evaluated TLC flash chip, the proposed sentinel-based approach totally removes read retries except the read operations used to assist the calibration.

We further conduct experiments on a trace-driven simulator SSDSim [39] to study the read performance improvement from the system perspective. The simulated flash system has the same settings as the real 3D NAND flash chips. Eight popular real workloads from the Microsoft Research Cambridge [40] are used for evaluation. Figure 14 shows the percentage of read latency reduction of the proposed sentinel approach compared with the method in the current flash. For the simulated situation, the proposed approach can effectively reduce read latency by 74% on average.

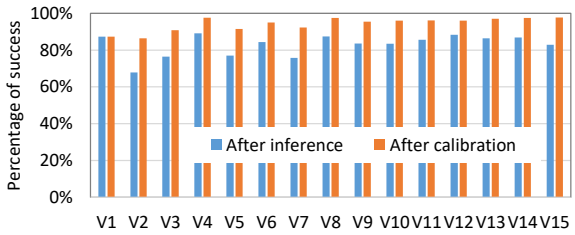

Fig. 15: The percentage of wordlines of which the optimal read voltage is successfully achieved.

B. Accuracy of Inference

We then present the performance of the optimal read voltage inference and calibration. Figure 15 shows the percentages of wordlines of which the optimal read voltages are successfully inferred and calibrated on the evaluated QLC flash memory. A successful inference is defined as the case that the difference of RBERs between inferred read voltages and the optimal is less than 5%. After inference, more than 83% wordlines can obtain the optimal read voltages and more than 94% wordlines obtain the optimal voltages after calibration. Note that the calibrated optimal read voltages of some cases can still successfully read the data even though the calibrated voltages are different from the optimal read voltages. The calibrated read voltages are much closer to the optimal read voltages compared to the default read voltages. Thus the RBER introduced by the calibrated read voltages is much lower than that of default voltages. Once it is lower than the error correction capability, the data can be correctly decoded and no more read retry is required.

The numbers of bit errors introduced by each read voltage are compared between four cases: default, inferred, calibrated, and optimal read voltages. Figures 16 and 17 show the results of the TLC and QLC flash chip, respectively. Each read voltage will introduce a certain number of errors, and the data is presented separately. Reading data with the default read voltages introduces lots of bit errors on both TLC and QLC flash chips. The inferred read voltage can greatly reduce the number of errors of all the wordlines. From V9 to V15 of the QLC flash chip, the number of errors introduced by optimal read voltages is similar to that of default, thus the reduction is less significant. Compared to the inferred read voltages, the calibrated read voltages can further reduce bit errors. Overall, the number of errors with the identified read voltages is close to that of optimal read voltages.

Note that it is impossible to obtain the same number of errors as the optimal read voltages. The reason is that there exist noises during a read operation as the cells with threshold voltages near the read voltages can easily be read as one of the two neighboring voltage states. Even two reads with the same read voltage might generate different RBERs. Therefore, the inference and calibration could not reach the same result as the optimal read voltages for some wordlines. The difference between the calibrated and real optimal for most wordlines in Figures 16 and 17 is acceptable.

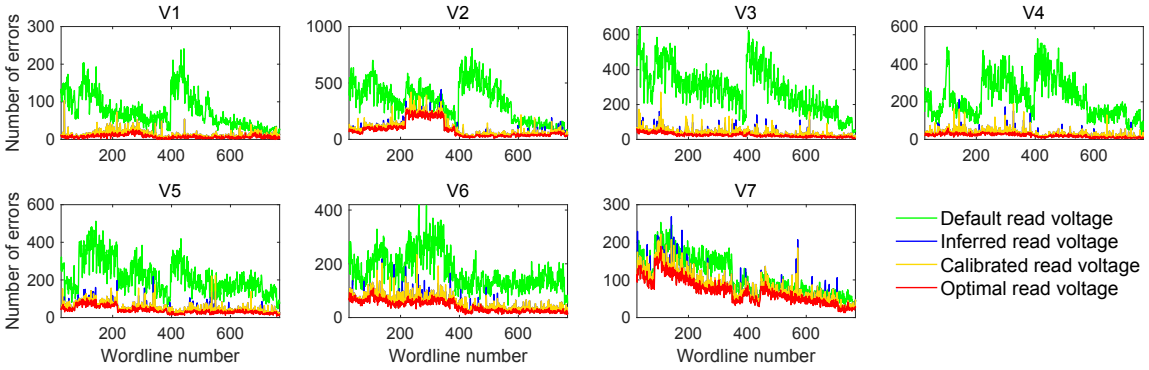

Fig. 16: The error number comparison of the TLC flash.

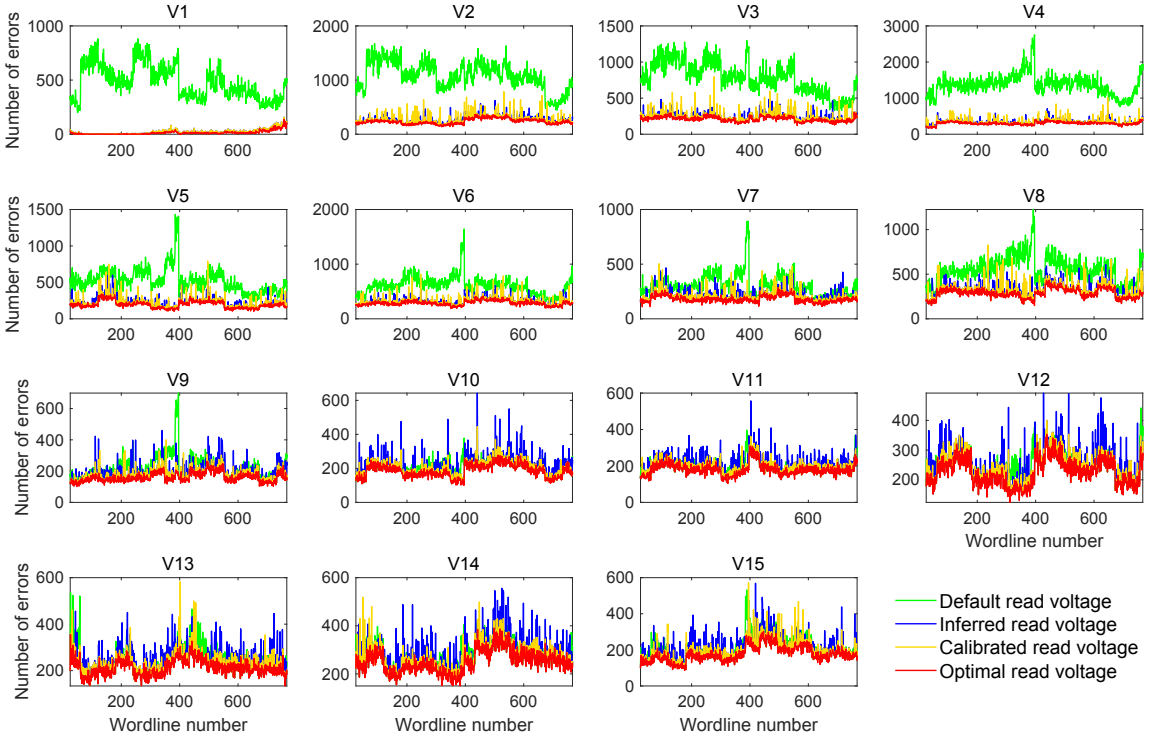

Fig. 17: The error number comparison of the QLC flash.

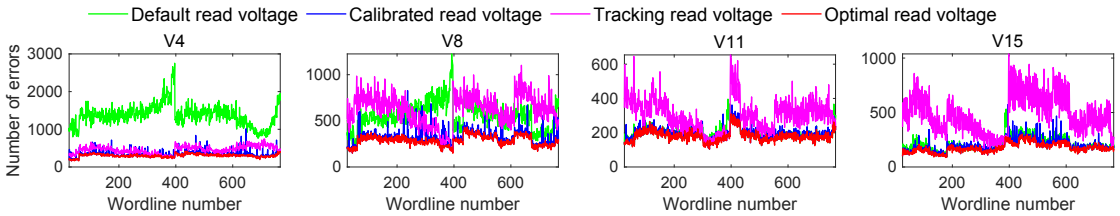

Fig. 18: The error number comparison of the QLC flash by considering existing tracking method.

We also compare the proposed approach with the existing method, which keeps track of the optimal read voltages of one wordline and uses it as the optimal read voltages of all other wordlines inside the same block. Figure 18 shows the results of comparisons, where four read voltages of the QLC flash are presented due to the page limit. As observed in the results, tracking does effectively reduce the bit errors for some wordlines. However, some wordlines cannot get any benefit from tracking, and the RBER may even exceed that of default read voltages. The reason is that the variations among wordlines are more severe in high-density 3D flash memory as presented in Section II. Our proposed method consistently

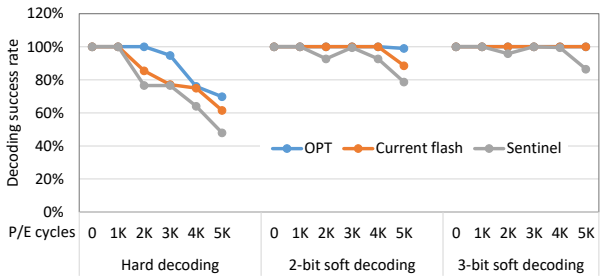

Fig. 19: The comparison of LDPC error correcting capability.

outperforms the existing tracking-based method.

C. Overhead

One potential overhead of the proposed approach is that sentinel cells take up some space of ECC parity. As discussed in Section III-D, if there are enough free cells to be reserved as sentinel cells, the error correction capability will not be impacted. All our evaluated flash chips have enough spare space for sentinel cells, where the OOB area takes up more than 10% on each wordline while sentinel cells only need 0.2%. In case of insufficient space for sentinel cells, we tested the decoding success rate among three methods: decoding with optimal read voltages ("OPT"), the decoding method in the current flash chip ("Current flash") and the proposed method by occupying some space of ECC parity with sentinel cells ("Sentinel"). To evaluate the worst case of the proposed approach, we suppose the space of all sentinel cells is taken from the space of ECC parity. Figure 19 presents the comparisons of the decoding success rate when LDPC is used as the default ECC, where the retention time is set as 1 year with different numbers of P/E cycles. As the figure shows, the success rates are all 100% when the number of P/E cycles is within 1000. After P/E cycles of 1000, the rate of "Sentinel" is slightly lower than the other two when using hard decoding and 2-bit soft decoding. The marginal loss in ECC capability can be well compensated by the improved optimal read voltage.

V. Related Work

To authors' best knowledge, this paper is the first to empirically study error characteristics considering read voltages on high-density 3D flash memory. This paper proposes a novel approach to reduce the number of read retry operations to improve the performance of 3D NAND flash memory. In the following, we will present the closely-related works.

3D NAND Flash Memory Characterization. Recent works study the error characteristics of 3D NAND flash memory, and identify differences between 3D and planar NAND flash memory due to memory cell design and architectural changes [41][17][18][35][32][42]. None of these works provide a detailed characterization of the impact of the read voltages. In addition, there is no public report on the error characterizations of QLC flash memory.

Read Retry Optimization. Currently, there are multiple approaches to minimize the number of read retries. Lots of works have studied the error characteristics of NAND flash memory, and estimated errors by considering various impacting factors, based on which the number of read retries is reduced [43][23][24][35][44]. Accurate estimation of voltage state shifting is challenging, which requires large overheads to record additional information in the FTL, especially on 3D NAND where there are more types of errors. Our evaluation on QLC flash chips shows even greater challenges. Another method is to track the optimal read voltages that generate the lowest RBER [28][29][25]. The overheads to periodically update the optimal voltages are significant, and the accuracy is limited when the temperature changes. Shim et al. [32] studied the process similarity among wordlines inside a layer of 3D TLC flash and presented that all the wordlines inside the same layer have the same optimal read voltages. This method reduces the overheads from each wordline to each layer, while the optimal read voltages are still needed to be maintained in the FTL. We should also note that the proposed method can be well combined with previous work. Read operations can start with the tracked optimal read voltages to reduce the failure rate of the first read operation, and our sentinel based prediction is applied once there is a read failure. Li et al. [21] proposed to predict the optimal read voltages based on the error difference by reserving a subset of cells on each wordline. However, the overhead is relatively large and their method works poorly for QLC flash. Our work differs from this method from three aspects: considering the characteristics of QLC, exploiting the correlation among optimal read voltages and further calibrating the read voltages in case of inference failure.

VI. Conclusion

Due to the increasing RBER on NAND flash memory, read retry operations are often required to re-read data with tuned read voltages, which prolongs read latency. Without accurate estimation of the errors, a successful read may need multiple trails to tune the read voltages. This problem becomes worse on high-density 3D flash as it is more challenging to model the errors and more read voltages are required for a read operation. In this work, we first characterize the errors and optimal read voltages on real high-density 3D NAND flash. Based on the observations, we reserve a subset of cells as sentinel cells on each wordline and select one read voltage as the sentinel voltage. The data corresponding to the sentinel voltage will be programmed on sentinel cells to infer the optimal read voltages for each wordline instantly. Finally, the inferred read voltages will be calibrated to address the bias between sentinel cells and data cells. Experiments on real flash chips verify that the proposed approach can effectively and efficiently infer the optimal read voltages and thus reduce the number of read retries with negligible overheads. The proposed method has a strong performance impact on real-world flash chips and is of great potential to be adopted in the industry.

Acknowledgment

The work was partial supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11204718) and NSFC 61772092.

REFERENCES

[1] J. H. Yoon, R. Godse, G. Tressler, and H. Hunter, "3d-nand scaling and 3d-scm—implications to enterprise storage," *Flash Memory Summit*, vol. 3, no. 4.2, pp. 3–4, 2017.

[2] C. Gao, M. Ye, Q. Li, C. J. Xue, Y. Zhang, L. Shi, and J. Yang, "Constructing large, durable and fast ssd system via reprogramming 3d tlc flash memory," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 493–505.

[3] B. S. Kim, J. Choi, and S. L. Min, "Design tradeoffs for SSD reliability," in *17th USENIX Conference on File and Storage Technologies (FAST)*. Boston, MA: USENIX Association, 2019, pp. 281–294. [Online]. Available: https://www.usenix.org/conference/fast19/presentation/kim-bryan

[4] A. Goda, "3-d nand technology achievements and future scaling perspectives," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1373–1381, 2020.

[5] C. M. Compagnoni and A. S. Spinelli, "Reliability of nand flash arrays: A review of what the 2-d–to–3-d transition meant," *IEEE Transactions on Electron Devices*, vol. 66, no. 11, pp. 4504–4516, 2019.

[6] D. Kang, W. Jeong, C. Kim, D.-H. Kim, Y. S. Cho, K.-T. Kang, J. Ryu, K.-M. Kang, S. Lee, W. Kim *et al.*, "256 gb 3 b/cell v-nand flash memory with 48 stacked wl layers," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 210–217, 2016.

[7] C. Kim, D.-H. Kim, W. Jeong, H.-J. Kim, I. H. Park, H.-W. Park, J. Lee, J. Park, Y.-L. Ahn, J. Y. Lee *et al.*, "A 512-gb 3-b/cell 64-stacked wl 3-d-nand flash memory," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 124–133, 2017.

[8] C. Siau, K.-H. Kim, S. Lee, K. Isobe, N. Shibata, K. Verma, T. Ariki, J. Li, J. Yuh, A. Amarnath *et al.*, "A 512gb 3-bit/cell 3d flash memory on 128-wordline-layer with 132mb/s write performance featuring circuit-under-array technology," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2019, pp. 218–220.

[9] N. Shibata, K. Kanda, T. Shimizu, J. Nakai, O. Nagao, N. Kobayashi, M. Miakashi, Y. Nagadomi, T. Nakano, T. Kawabe *et al.*, "A 1.33 tb 4-bit/cell 3d-flash memory on a 96-word-line-layer technology," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2019, pp. 210–212.

[10] H. Huh, W. Cho, J. Lee, Y. Noh, Y. Park, S. Ok, J. Kim, K. Cho, H. Lee, G. Kim *et al.*, "13.2 a 1tb 4b/cell 96-stacked-wl 3d nand flash memory with 30mb/s program throughput using peripheral circuit under memory cell array technique," in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2020, pp. 220–221.

[11] Q. Li, L. Shi, C. J. Xue, K. Wu, C. Ji, Q. Zhuge, and E. H.-M. Sha, "Access characteristic guided read and write cost regulation for performance improvement on flash memory," in *14th USENIX Conference on File and Storage Technologies (FAST 16)*, 2016, pp. 125–132.

[12] Y. Cai, S. Ghose, E. F. Haratsch, Y. Luo, and O. Mutlu, "Errors in flash-memory-based solid-state drives: Analysis, mitigation, and recovery," *arXiv preprint arXiv:1711.11427*, 2017.

[13] K. Zhao, W. Zhao, H. Sun, X. Zhang, N. Zheng, and T. Zhang, "Ldpc-in-ssd: Making advanced error correction codes work effectively in solid state drives," in *Presented as part of the 11th USENIX Conference on File and Storage Technologies (FAST 13)*, 2013, pp. 243–256.

[14] I. Narayanan, D. Wang, M. Jeon, B. Sharma, L. Caulfield, A. Sivasubramaniam, B. Cutler, J. Liu, B. Khessib, and K. Vaid, "Ssd failures in datacenters: What? when? and why?" in *Proceedings of the 9th ACM International on Systems and Storage Conference*. ACM, 2016, p. 7.

[15] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "A large-scale study of flash memory failures in the field," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 1. ACM, 2015, pp. 177–190.

[16] E. Xu, M. Zheng, F. Qin, Y. Xu, and J. Wu, "Lessons and actions: What we learned from 10k ssd-related storage system failures," in *2019 {USENIX} Annual Technical Conference ({USENIX}{ATC} 19)*, 2019, pp. 961–976.

[17] Q. Xiong, F. Wu, Z. Lu, Y. Zhu, Y. Zhou, Y. Chu, C. Xie, and P. Huang, "Characterizing 3D floating gate NAND flash," in *2017 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*. ACM, 2017, pp. 31–32.

[18] F. Wu, Y. Zhu, Q. Xiong, Z. Lu, Y. Zhou, W. Kong, and C. Xie, "Characterizing 3D charge trap NAND flash: Observations, analyses and applications," in *2018 IEEE 36th International Conference on Computer Design (ICCD)*. IEEE, 2018, pp. 381–388.

[19] D. Kang, W. Jeong, C. Kim, D.-H. Kim, Y. S. Cho, K.-T. Kang, J. Ryu, K.-M. Kang, S. Lee, W. Kim *et al.*, "256 Gb 3 b/cell V-NAND flash memory with 48 stacked wl layers," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 210–217, 2017.

[20] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "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)*. EDA Consortium, 2013, pp. 1285–1290.

[21] Q. Li, M. Ye, Y. Cui, L. Shi, X. Li, and C. J. Xue, "Sentinel cells enabled fast read for NAND flash," in *11th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 19)*, 2019.

[22] A. Fukami, S. Ghose, Y. Luo, Y. Cai, and O. Mutlu, "Improving the reliability of chip-off forensic analysis of nand flash memory devices," *Digital Investigation*, vol. 20, pp. S1–S11, 2017.

[23] K.-C. Ho, P.-C. Fang, H.-P. Li, C.-Y. M. Wang, and H.-C. Chang, "A 45nm 6b/cell charge-trapping flash memory using LDPC-based ECC and drift-immune soft-sensing engine," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE, 2013, pp. 222–223.

[24] F. Sala, R. Gabrys, and L. Dolecek, "Dynamic threshold schemes for multi-level non-volatile memories," *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2624–2634, 2013.

[25] Y. Luo, S. Ghose, Y. Cai, E. F. Haratsch, and O. Mutlu, "Heatwatch: Improving 3D NAND flash memory device reliability by exploiting self-recovery and temperature awareness," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2018, pp. 504–517.

[26] Z. Fan, G. Cai, G. Han, W. Liu, and Y. Fang, "Cell-state-distribution-assisted threshold voltage detector for nand flash memory," *IEEE Communications Letters*, vol. 23, no. 4, pp. 576–579, 2019.

[27] N. Papandreou, N. Ioannou, T. Parnell, R. Pletka, M. Stanisavljevic, R. Stoica, S. Tomic, and H. Pozidis, "Reliability of 3d nand flash memory with a focus on read voltage calibration from a system aspect," in *2019 19th Non-Volatile Memory Technology Symposium (NVMTS)*. IEEE, 2019, pp. 1–4.

[28] Y. Cai, Y. Luo, E. F. Haratsch, K. Mai, and O. Mutlu, "Data retention in MLC NAND flash memory: Characterization, optimization, and recovery," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2015, pp. 551–563.

[29] B. Peleato, R. Agarwal, J. Cioffi, M. Qin, and P. H. Siegel, "Towards minimizing read time for NAND flash," in *2012 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2012, pp. 3219–3224.

[30] N. Papandreou, T. Parnell, H. Pozidis, T. Mittelholzer, E. Eleftheriou, C. Camp, T. Griffin, G. Tressler, and A. Walls, "Using adaptive read voltage thresholds to enhance the reliability of mlc nand flash memory systems," in *Proceedings of the 24th edition of the great lakes symposium on VLSI*. ACM, 2014, pp. 151–156.

[31] B. Choi, S. H. Jang, J. Yoon, J. Lee, M. Jeon, Y. Lee, J. Han, J. Lee, D. M. Kim, D. H. Kim *et al.*, "Comprehensive evaluation of early retention (fast charge loss within a few seconds) characteristics in tube-type 3-d nand flash memory," in *2016 IEEE Symposium on VLSI Technology*. IEEE, 2016, pp. 1–2.

[32] Y. Shim, M. Kim, M. Chun, J. Park, Y. Kim, and J. Kim, "Exploiting process similarity of 3d flash memory for high performance ssds," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 211–223.

[33] R. L. Schroeder, Bianca and A. Merchant., "Flash reliability in production: The expected and the unexpected," in *14th USENIX Conference on File and Storage Technologies (FAST 16)*, 2016, pp. 67–80.

[34] E. H. Wilson, M. Jung, and M. T. Kandemir, "ZombieNAND: Resurrecting dead NAND flash for improved SSD longevity," in *2014 IEEE 22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems*. IEEE, 2014, pp. 229–238.

[35] Y. Luo, S. Ghose, Y. Cai, E. F. Haratsch, and O. Mutlu, "Improving 3d nand flash memory lifetime by tolerating early retention loss and process variation," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 2, no. 3, p. 37, 2018.

[36] Q. Li, L. Shi, Y. Cui, and C. J. Xue, "Exploiting asymmetric errors for ldpc decoding optimization on 3d nand flash memory," *IEEE Transactions on Computers*, 2019.

[37] YEESTOR, "Ys9083xt/ys9081xt ssd platform," http://www.yeestor.com/index-product-info-id-946-cid-33-pid-3-infopid-33.html, 2018.

[38] Micron, "Mt29f128g08cbcebj4-37itr 3d tlc nand flash," https://www.micron.com/products/nand-flash/3d-nand/part-catalog/mt29f1t08eehafj4-3r, 2018.
[39] Y. Hu, H. Jiang, D. Feng, L. Tian, H. Luo, and C. Ren, "Exploring and exploiting the multilevel parallelism inside SSDs for improved performance and endurance," *IEEE Transactions on Computers (TC)*, vol. 62, no. 6, pp. 1141–1155, 2013.
[40] D. Narayanan, E. Thereska, A. Donnelly, S. Elnikety, and A. Rowstron, "Migrating server storage to SSDs: analysis of tradeoffs," in *Proceedings of ACM European conference on Computer systems*, 2009, pp. 145–158.
[41] K. Mizoguchi, T. Takahashi, S. Aritome, and K. Takeuchi, "Data-retention characteristics comparison of 2d and 3d tlc nand flash memories," in *2017 IEEE International Memory Workshop (IMW)*. IEEE, 2017, pp. 1–4.
[42] W. Lin, J. Chen, X. Zhang, and Z. Cheng, "Improving 3d nand flash memory read performance by modeling the read offset," in *2019 IEEE 19th International Conference on Communication Technology (ICCT)*. IEEE, 2019, pp. 1472–1476.
[43] T. Parnell, N. Papandreou, T. Mittelholzer, and H. Pozidis, "Modelling of the threshold voltage distributions of sub-20nm NAND flash memory," in *2014 IEEE Global Communications Conference*. IEEE, 2014, pp. 2351–2356.
[44] Z. Peng, R. He, G. Han, G. Cai, and Y. Fang, "Neighbor-a-posteriori information assisted cell-state adaptive detector for nand flash memory," *IEEE Communications Letters*, vol. 23, no. 11, pp. 1967–1971, 2019.