Open Block Characterization and Read Voltage Calibration of 3D QLC NAND Flash

Nikolaos Papandreou*, Haralampos Pozidis*, Nikolas Ioannou*, Thomas Parnell*, Roman Pletka*,
Milos Stanisavljevic*, Radu Stoica*, Sasa Tomic*, Patrick Breen†, Gary Tressler†,
Aaron Fry‡, Timothy Fisher‡, Andrew Walls‡

*IBM Research - Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland
†IBM Systems, 2455 South Road, Poughkeepsie, NY, USA
‡IBM Systems, 10777 Westheimer Road, Houston, TX, USA

***Abstract*—3D QLC NAND has recently entered the SSD market offering capacity increase and cost reduction compared to 3D TLC NAND. However, the endurance of QLC NAND is limited. Moreover, due to reduction of the available margin between the programmed threshold voltage distributions, QLC NAND is more susceptible to bit errors. Read voltage calibration is a key element of modern NAND flash memory controllers to improve the overall bit-error rate and maintain enterprise level reliability. To reduce the calibration overhead associated with the increased number of pages and read voltages in QLC NAND, page grouping is an effective approach. This paper presents open block characterization and read voltage calibration results of state-of-the-art 3D QLC NAND. We present experimental measurements of the bit-error characteristics and threshold voltage distributions based on closed and open block test patterns. We discuss the reliability issues with open blocks in preserving uniform characteristics within a page group at the boundary programmed layer and analyze the performance of different calibration algorithms.**

***Index Terms*—3D NAND, quad-level cell (QLC), reliability, read voltage calibration, open block, threshold voltage.**

I. Introduction

NAND flash memory has been making steady advances in terms of die capacity since the transition of the manufacturers to 3D technology. Recently, quad-level cell (QLC) 3D NAND with 4 bits/cell storage has been introduced, offering more capacity and lower cost-per-bit than its predecessor triple-level cell (TLC) 3D NAND [1]–[3]. However, this capacity increase comes at the price of lower endurance, measured in program/erase (P/E) cycles that the device can withstand. This is because of the larger number of threshold voltage (V_{TH}) levels that QLC cells store compared to TLC (16 vs. 8 levels), and the reduced margin between the adjacent levels, which causes level crossing and thus bit errors. This is further exacerbated by progressive P/E cycling of flash cells, and by data retention and read-disturb. These operations cause shifts and broadening of the V_{TH} distributions, leading to rapid deterioration of the raw bit-error rate (RBER) in the NAND flash blocks [4]–[7].

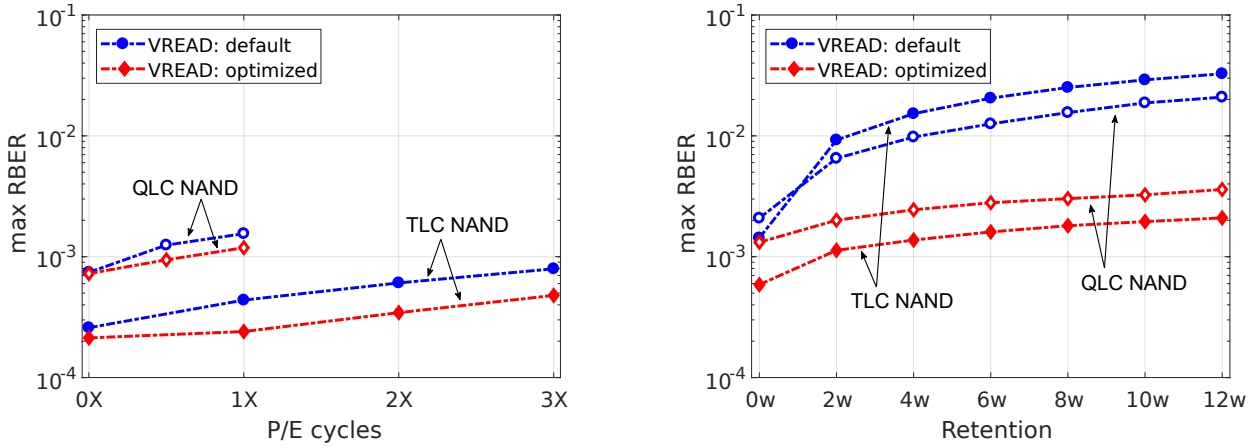
Modern NAND flash controllers typically employ methods to periodically adjust the read voltages in order to improve the RBER as the blocks undergo different types of stress. We refer to these methods as read voltage calibration in the rest of this manuscript. Although read voltage calibration is effective in combating different types of the memory device stress, it requires that a large amount of metadata is kept by the controller for each block. This problem has become worse with QLC, as there are more voltages to calibrate and more pages per block. To reduce the amount of metadata per block, a typical approach is to group multiple pages and calibrate a common set of voltages that are used across all pages in the group. The criterion used to group pages together is that they exhibit similar characteristics.

The increased block size in QLC NAND has additional collateral effects. For example, for workloads that are not write intensive, it is not uncommon that a block is not programmed fully in one go, but only partially at one time and then completed at some later time. As a result, the pages programmed in the first pass undergo data retention or read-disturb stress and thus exhibit different characteristics compared to those pages programmed at later passes. This phenomenon may affect the read voltage calibration process and thus the device reliability if it happens that programming is suspended before all pages in a page group have been written.

In this paper we present open block characterization and read voltage calibration results of state-of-the-art 3D QLC NAND devices. We discuss the reliability issues that may arise by program suspension in open blocks and analyze the performance and robustness of different read voltage calibration algorithms.

II. QLC Bit-Error Characteristics

Fig. 1 presents characterization results of (a) P/E cycling followed by (b) data retention from state-of-the-art QLC and TLC 3D NAND flash. The results show the maximum page RBER measured throughout the stress-test and averaged across multiple blocks of the QLC and TLC devices under test. The endurance measurements were collected at fixed intervals of P/E cycles (shown in normalized number of cycles), whereas the data retention measurements correspond to a total equivalent time of 3 months at 40 °C. The overall testing sequence was performed at elevated temperature to accelerate retention and charge-loss recovery effects. For each measurement we collected read data by using a read-voltage sweep, which



(a) Program-erase cycling stress.

(b) Data retantion stress.

Fig. 1. Maximum RBER as a function of (a) P/E cycles followed by (b) 12 weeks data retention. Optimized read voltages offer endurance and data retention gains for both QLC and TLC NAND.

enables calculation of the optimal offset values that results in the minimum number of bit-errors at each data point of the test sequence. Each graph in Fig. 1 shows the maximum RBER curves that correspond to the default (in blue) or optimized (in red) read voltages.

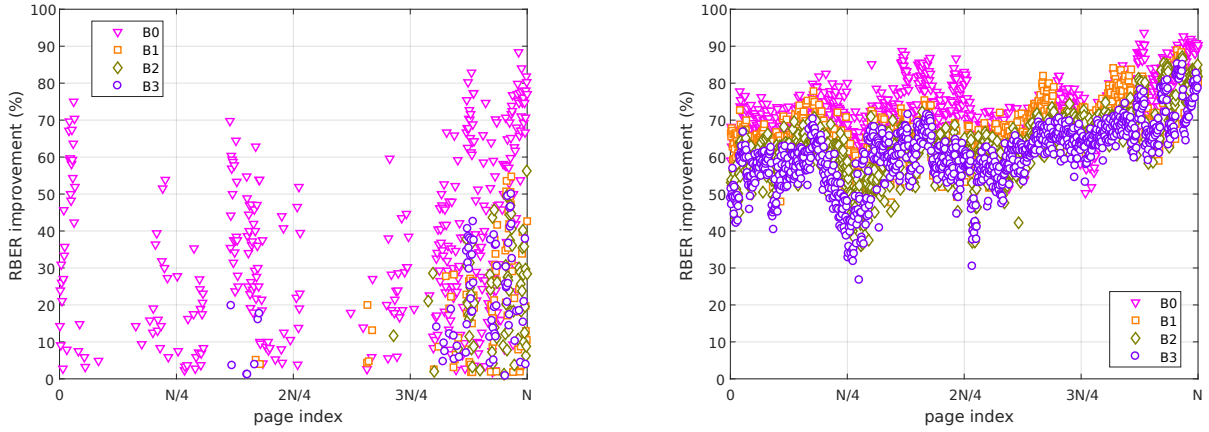
QLC NAND shows higher RBER from the onset of the cycling due to the more and denser packed V_{TH} distributions compared to TLC NAND. The increased error level is more pronounced for QLC NAND at 1X P/E cycles, whereas TLC NAND shows lower RBER after 3X P/E cycles. During data retention, the bit errors increase significantly for both QLC and TLC NAND. The RBER shows an abrupt increase during early retention, a behavior that was also reported for 2 and 3 bits/cell 3D NAND in [7]–[10]. Use of optimized read voltages is essential to lower the RBER and thus improve the device endurance and retention. The gains are significant in particular for data retention: the V_{TH} distributions move to the left due to charge loss and thus the memory controller needs to apply a set of negative offsets to reduce the RBER below 10^{-2}. Note that the default voltages are preset from the manufacturers based on different endurance and retention targets for the QLC and TLC devices. This explains the higher RBER for TLC NAND with default voltages in Fig. 1(b). On the other hand, the lowest RBER is achieved with optimized read voltages and thus the device lifetime can be extended.

Focusing on the QLC characterization data, Fig. 2 shows the amount of RBER improvement that is achieved when the controller uses read voltages optimized for the different stress conditions, e.g., cycling or data retention, compared to the default ones. The two graphs correspond to selected measurement points of the test in Fig. 1. Specifically, the results in Fig. 2(a) correspond to the end of the P/E cycling phase, whereas the results in Fig. 2(b) correspond to the end of the data retention phase. The amount of RBER improvement as a function of the page index is calculated for each of the QLC page types. We denote as B0, B1, B2 and B3 the four different pages that share the same word-line (WL) in a QLC NAND block, where B0 and B3 refer to the pages that store the least-significant bits (LSB) and most-significant bits (MSB), respectively, whereas B1 and B2 refer to the pages that store the two intermediate bits.

Fig. 2(a) shows that the default read voltages are effective for many of the pages at 1X P/E cycling. Moreover, we observe that the B0 pages are the ones that benefit the most from optimized read voltages. On the other hand, Fig. 2(b) shows that, during data retention, almost every page (and page type) achieves significant gains above 50% with optimized read voltages. Finally, both graphs show that the amount of RBER improvement is different between the various pages and page types, which can be attributed to the different effect of each stress mechanism, namely P/E cycling and data retention, on each page type, and is also related to process variations between different layers and areas in the block. The latter is expected to become more pronounced with the increase in the number of layers in next generation 3D NAND flash. From a technology point of view, reducing the variability between layers requires further advancements in the manufacturing process. From the memory controller perspective, the increase of page variability imposes additional workload to the calibration engine. The variability of error characteristics across pages and page types is captured in Figs. 3(a) and 3(b), which show the range of bit errors within a QLC block for the measurement points of Figs. 2(a) and 2(b), respectively. Note that in both graphs, the sorted page index (x-axis) is different for each page type.

III. Open Block Reliability Issues

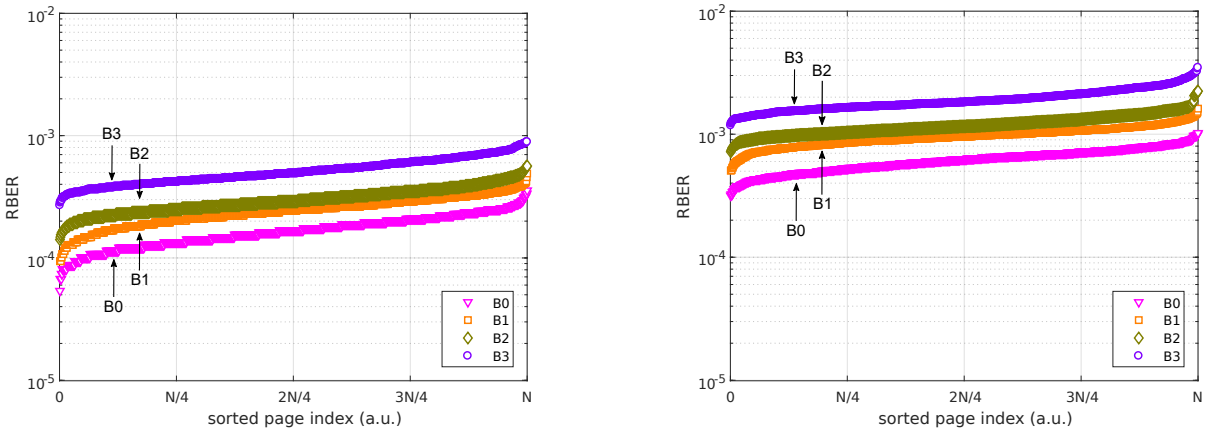
The reliability results presented in Figs. 1-3 highlight the importance of read voltage optimization for modern 3D QLC NAND. As it was also demonstrated for 3D TLC NAND devices in [7], the optimal read voltages may change significantly between different device stress conditions, e.g., retention or



(a) 1X program-erase cycles.

(b) 1X program-erase cyckes + 12 weeks data retention.

Fig. 2. RBER improvement using optimized read voltages compared to the default ones at (a) 1X P/E cycles and (b) 1X P/E cycles followed by 12 weeks data retention. The default read voltages are optimal for many of the pages at 1X P/E cycles. However, during retention, the majority of the pages have an improvement that exceeds 50% with optimized read voltages. Both graphs show results from one of the QLC blocks under test.



(a) 1X program-erase cycles.

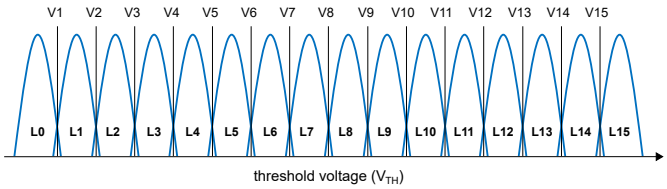
(b) 1X program-erase cyckes + 12 weeks data retention.

Fig. 3. Variability of RBER across pages and page types (B0-B3) at (a) 1X P/E cycles and (b) 1X P/E cycles followed by 12 weeks of data retention. The RBER values are displayed in ascending order and the sorted page index is different for each page type. Both graphs show RBER results that correspond to the optimized read voltages for each stress condition.

read-disturb, and at the same time, the optimal offset values may be different for pages located in different areas of the block. Therefore, proactive background calibration of read voltages triggered by a variety of metrics, such as retention time, P/E cycling and read-disturb counts, or other workload-related metrics, is an effective means to adjust the optimal offset values and thus avoid read-retry operations that cost in read latency [11].

Calibrating the read voltages for every page in a block, however, incurs a significant overhead in the amount of background reads and metadata required to keep track of the various offset values. Fig. 4(a) illustrates the 16 V_{TH} levels (L0-L15) in QLC NAND. A total of 15 read voltages (V1-V15) are used to distinguish between the adjacent levels. Depending on the 4-bit encoding, which is usually vendor dependent, a subset of the read voltages are used to decode the bits of each page type (B0-B3). The table in Fig. 4(b) shows the overhead increase for basic calibration parameters, such as the number of pages and read voltage offsets per block, going from 64-layer TLC NAND to 96-layer QLC NAND.

In [7], a number of approaches that aim to reduce the overall calibration overhead were discussed: (a) use of semi-optimized read voltages; such algorithms can reduce the complexity of finding the global optimum solution by using a common offset for some of the voltages; (b) grouping of pages with similar error and read voltage characteristics; such a technique can significantly reduce the amount of metadata under the condition that the pages in each group maintain similar characteristics; (c) use of a single set of read offset values for all pages in a page group; this method is applicable to pages of the same page type and can further decrease the overall metadata requirements.



(a) Illustration of the 16 V_{TH} levels (L0-L15) and corresponding read voltages (V1-V15) in QLC NAND with 4 bits/cell.

3D TLC NAND	3D QLC NAND	Relative increase of calibration parameters
64-layer, 3 bits/cell	96-layer, 4 bits/cell	100% in number of pages per block
7 read voltages per WL	15 read voltages per WL	220% in number of read voltage offsets

(b) Increase of relative calibration parameters from TLC to QLC NAND.

Fig. 4. Calibration overhead as a result of increased block size and number of read voltages from 64-layer TLC to 96-layer QLC NAND flash.

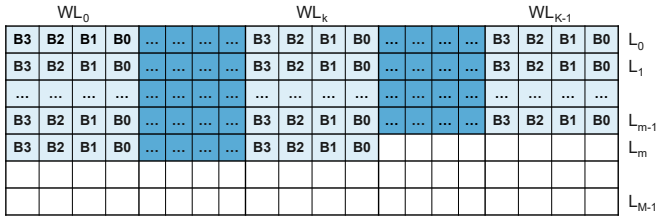


Fig. 5. Illustration of open block scenario. Pages are programmed sequentially up to word-line WL_k at layer L_m. The rest of the word-lines in the block are either incomplete, i.e., some pages may be programmed in intermediate states (not shown) or they are in the erase state.

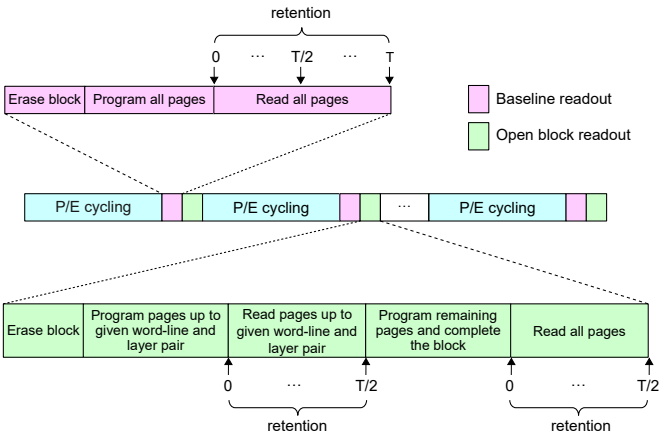


Fig. 6. Illustration of open block testing procedure. (a) Baseline readout: erase the block, program all pages, read all pages at regular intervals for a period T; (b) Open block readout: erase the block, program all pages up to a specific WL and layer; read the programmed pages at regular intervals for a period $T/2$; program the remaining pages and complete the block; read all pages for an additional period $T/2$.

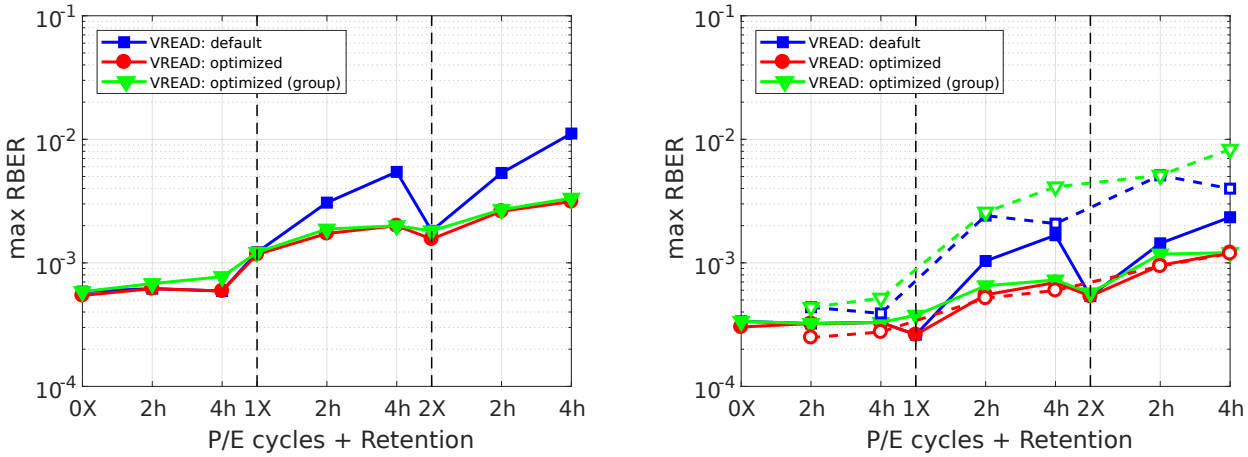
However, the key requirement for effective page grouping and low-complexity calibration schemes, namely the uniformity of page characteristics within a group, may be violated in workload-dependent scenarios, where the NAND memory controller suspends writing the block and continues programming after some time. Such a scenario is illustrated in Fig. 5, where a block of M layers and K word-lines per layer is depicted. Following the device page programming order, the memory controller writes data up to a point where programming stopped at layer L_m and word-line WL_k. We denote this layer as boundary layer, where only part of the pages have been programmed. Afterwards, the block stays open for some time and then data writing resumes by continuing the programming of the remaining pages until the entire block is completed. While the block stays open, i.e., the block is partially filled up to the boundary layer, the programmed pages may be read and thus they are subjected to read-disturb in addition to other retention effects.

To characterize the effect of open blocks in the page characteristics, a special set of experiments shown in Fig. 6 was designed. The QLC NAND block under test is subjected to a sequence of P/E cycles. At regular intervals we collect measurements from two different types of readouts. (a) The *baseline* readout consists of the following steps: erase the block, program all pages, read all pages at regular intervals for a period T; (b) The *open block* readout consists of the following steps: erase the block, program all pages up to a specific WL and layer; read the programmed pages up to the given WL and layer at regular intervals for a period $T/2$; program the remaining pages and complete the block; read all pages for an additional period $T/2$. Each readout is performed using a read-voltage sweep, which allows us to evaluate the performance of different calibration schemes as well as to extract the V_{TH} distributions.

We assume that the pages in one or more adjacent layers are grouped together and the calibration engine calculates a set of read voltages that is applied to all pages in the group. This assumption is justified by the fact that typically pages in neighboring layers have a similar bit-error profile. Moreover, each group consists of pages of the same type (B0-B3). Therefore, the pages of each group that contains the boundary layer, where programming stops in the testing procedure of Fig. 6, are divided into two subsets that have different retention exposure. The first subset experiences a total retention time of T, similar to the overall retention time of the *baseline* readout, whereas the second subset experiences a total retention time of $T/2$.

IV. Analysis of Characterization Results

Fig. 7 shows the RBER measurements according to the test sequence in Fig. 6. The QLC block under test is subjected to mixed phases of P/E cycles and data retention, i.e., successive intervals of 1X P/E cycles followed by 4 hours retention at 65 °C. Each graph compares the maximum page RBER with default read voltages (in blue), with read voltages optimized for each page separately (in red), and with read voltages optimized for each page group, where a single set of offset values are applied when the controller attempts to read any page in the same group (in green).



(a) All pages (baseline readout).

(b) Pages at the boundary layer (baseline vs. open block readouts).

Fig. 7. (a) Maximum RBER across all pages in the block for the baseline readout. Default read voltages result in high RBER even within 4 hours at 65 °C. Optimized read voltages show significant gains during retention. (b) Maximum RBER across the pages in the boundary layer for the baseline (solid lines) and open block (dashed lines) readouts. Selecting a random page in the boundary layer as reference for calibration may cause significant RBER increase due to the different retention history between the pages programmed before and after program suspension.

Fig. 7(a) presents the baseline measurements where the memory controller programs all pages in the block before any retention takes place. The graph shows the maximum RBER across all pages in the block under test. We observe that the default read voltages result in elevated RBER even within 4 hours at 65 °C. On the other hand, the optimized read voltages provide significant RBER improvements during retention. Moreover, both calibration schemes, namely read voltage optimization per page (in red) and per page group (in green), perform similarly. The latter shows a slightly elevated RBER. Typically, despite how good the page grouping formation is, there may be a mismatch between the single set of offset values of the group and the individual values of each member page. Therefore, some pages in the group may exhibit a higher RBER compared to the case where the read voltages were optimized for each page separately.

Fig. 7(b) shows the open block RBER measurements where programming stops at a given layer and continues after some time has elapsed. This graph shows the maximum RBER between the pages located at the boundary layer only. As described in Section III, these pages belong to groups that contain the boundary layer and, in an open block scenario, they are divided into two subsets with different retention history. The first subset (before controller paused page programming) has experienced a total of $2+2=4$ hours at 65 °C retention, whereas the second subset (after controller continued page programming) has experienced a total of 2 hours at 65 °C retention. For comparison, we show the maximum RBER of the selected pages for both baseline (solid lines) and open block (dashed lines) readouts. We observe that the error level after page programming resume depends heavily on the read voltage optimization algorithm. If each page is calibrated separately the RBER is improved significantly and is similar to that of the baseline readout. On the contrary, if a single set of voltages is used for each page group in the boundary layer of the open block then the RBER increases substantially. To understand this behavior, we analyze the V_{TH} distributions in the boundary layer.

Fig. 8 shows the V_{TH} distributions of the WLs in the boundary layer after 2X P/E cycles and 4 hours retention. The first subset (in blue) corresponds to the word-lines that were programmed before the controller stopped programming, and the second subset (in red) corresponds to the remaining word-lines in the boundary layer that were programmed when the controller resumed programming after a 2 hours pause. We observe that the two subsets of V_{TH} distributions exhibit a relative shift due to the different time at which they were programmed and the different retention history that they experienced. It is also clear that the optimal read voltages are different for each subset. Therefore, when a set of read offsets from a page belonging to one of the subsets is applied to all pages in the group, the RBER may increase since, for QLC in particular, small deviations from the optimal offsets can cause a high number of bit errors. This behavior explains the large RBER increase in Fig. 7(b) for the case of optimized read voltages per group.

We consider a similar experiment to the one presented in Fig. 6, but instead of introducing idle time between program suspend and program resume, we apply read-disturb cycles to the programmed pages. Fig. 9 shows the RBER measurements where programming stops at a given layer and then the programmed pages are subjected to an equivalent of 1K "block" read cycles (a block read cycle corresponds to reading so many pages as the size of the block). Afterwards, page programming continuous and the block is subjected to another 1K block read cycles. The graph shows the maximum RBER between the pages located at the boundary layer only, which again consist of two subsets with different read-disturb history. The first

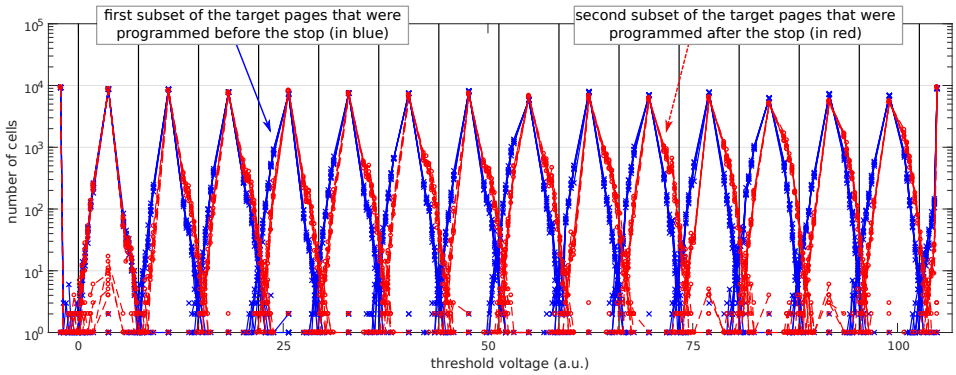


Fig. 8. Measured V_{TH} distributions of the word-lines at the boundary layer after 2X P/E cycles and 4 hours retention. The distributions in blue correspond to the first subset of pages that were programmed before the stop and have experienced 4 hours retention. The ones in red correspond to the second subset of pages that were programmed with a difference of 2 hours compared to the first subset and have experienced 2 hours retention. The solid vertical lines indicate the default read voltages.

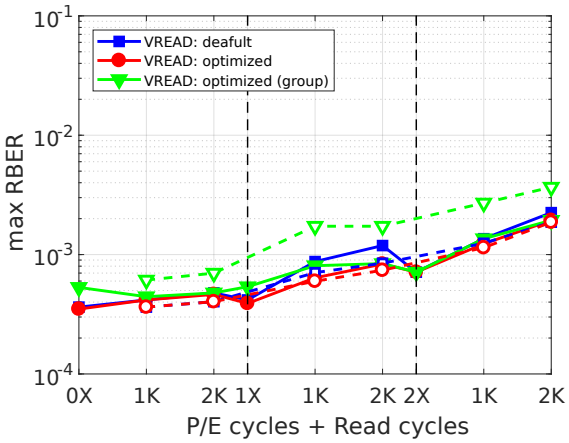


Fig. 9. Maximum RBER across pages in the boundary layer for the baseline (solid lines) and open block (dashed lines) readouts in a mixed P/E cycling and read-disturb experiment.

subset (before programming stopped) has experienced a total of $1K + 1K = 2K$ block reads, whereas the second subset (after programming continuous) has experienced a total of 1K block reads. We observe that the group-based calibration may result in high RBER, as a result of the different read-disturb history between the pages in the boundary layer.

The above analysis unveils the reliability issues with open blocks where the NAND flash controller uses a single set of voltage offsets for multiple pages that exhibit different characteristics. It is therefore beneficial for the controller to complete the block programming as soon as possible and to resort to more fine-grained calibration algorithms in such scenarios as the ones presented above.

V. Conclusions

In this paper we presented open block characterization and read voltage calibration results of 3D QLC NAND. We discussed the reliability issues with open blocks in preserving uniform characteristics within a page group at the boundary programmed layer and analyzed the performance of different calibration algorithms. We explained the increased error levels that can be observed in open block scenarios and we presented experimental measurements of RBER and V_{TH} distributions. These results can provide useful information for flash management operations in modern QLC NAND flash controllers.

References

[1] K. Parat and A. Goda, "Scaling Trends in NAND Flash," in *2018 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2018, pp. 2.1.1–2.1.4.

[2] S. Lee *et al.*, "A 1Tb 4b/Cell 64-Stacked-WL 3D NAND Flash Memory with 12MB/s Program Throughput," in *2018 IEEE International Solid-State Circuits Conference - (ISSCC)*, Feb. 2018, pp. 340–342.

[3] Y. Takai, M. Fukuchi, R. Kinoshita, C. Matsui, and K. Takeuchi, "Analysis on Heterogeneous SSD Configuration with Quadruple-Level Cell (QLC) NAND Flash Memory," in *2019 IEEE 11th International Memory Workshop (IMW)*, May 2019, pp. 1–4.

[4] C. Zambelli, R. Micheloni, and P. Olivo, "Reliability challenges in 3D NAND Flash memories," in *2019 IEEE 11th International Memory Workshop (IMW)*, May 2019, pp. 1–4.

[5] P. Breen, N. Papandreou, and G. Tressler, "Component-Level Characterization of 3D TLC, QLC, and Low-Latency NAND," in *2019 Flash Memory Summit (FMS)*, Aug. 2019.

[6] W. Liu, F. Wu, M. Zhang, Y. Wang, Z. Lu, X. Lu, and C. Xie, "Characterizing the Reliability and Threshold Voltage Shifting of 3D Charge Trap NAND Flash," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2019, pp. 312–315.

[7] N. Papandreou, H. Pozidis, T. Parnell, N. Ioannou, R. Pletka, S. Tomic, P. Breen, G. Tressler, A. Fry, and T. Fisher, "Characterization and Analysis of Bit Errors in 3D TLC NAND Flash Memory," in *2019 IEEE International Reliability Physics Symposium (IRPS)*, Mar. 2019, pp. 1–6.

[8] 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)*, May 2017, pp. 1–4.

[9] P. Breen, T. Griffin, N. Papandreou, T. Parnell, and G. Tressler, "3D NAND Assessment for Next Generation Flash Applications," in *2016 Flash Memory Summit (FMS)*, Aug. 2016.

[10] 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," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 2, no. 3, Dec. 2018.

[11] R. Pletka, I. Koltsidas, N. Ioannou, S. Tomic, N. Papandreou, T. Parnell, H. Pozidis, A. Fry, and T. Fisher, "Management of Next-Generation NAND Flash to Achieve Enterprise-Level Endurance and Latency Targets," *ACM Trans. Storage*, vol. 14, no. 4, Dec. 2018.