Intrablock Wear Leveling to Counter Layer-to-Layer Endurance Variation of 3-D NAND Flash Memory

Md Raquibuzzaman, *Member, IEEE*, Aleksandar Milenkovic, *Senior Member, IEEE*, and Biswajit Ray, *Senior Member, IEEE*

***Abstract*—A shift to 3-D NAND technology has resulted in flash memory blocks that include many pages, leading to "big-block" management issues in storage systems. This article experimentally explores endurance variability in 3-D NAND flash memory blocks and finds that pages in the bottom and top layers exhibit lower endurance than pages in the middle layers. We find that erase threshold voltage (V_t) variation between the layers is the root cause for the observed endurance variation. This variation in endurance among pages can cause severe underutilization of flash memory. To counter these effects and improve the overall utilization of 3-D NAND flash memories, we propose an intrablock wear-leveling algorithm based on dynamic, incremental, and layer-aware downsizing of flash memory blocks.**

***Index Terms*—3-D NAND flash, endurance, layer variation.**

I. Introduction

TRADITIONAL 2-D NAND flash technology reached the end of scaling with sub-15-nm technology nodes [1], [2], hampered by an increasing cost of lithography and fundamental cell reliability issues such as random telegraph noise, program noise, and cell-to-cell interference. The flash memory industry has responded by transitioning to monolithic 3-D NAND flash fabrication processes. A shift to 3-D NAND technology opened up a new "scaling" dimension—the number of vertical layers—enabling further increases in the bit density for a given die area [3], [4]. Continual advances in this technology resulted in several generations of 3-D flash memory chips, each having a larger number of stacked layers, from early 32- to contemporary 178-layer designs, as shown in Fig. 1. These advances promise to extend an incredible growth of bit density over the next decade [1], [5], [6]. However, these trends increase the number of pages in a memory block leading to large-sized memory blocks. Managing such blocks in the flash translation layer (FTL) presents a new set of challenges for data mapping and wear leveling of storage media [7], [8], [9].

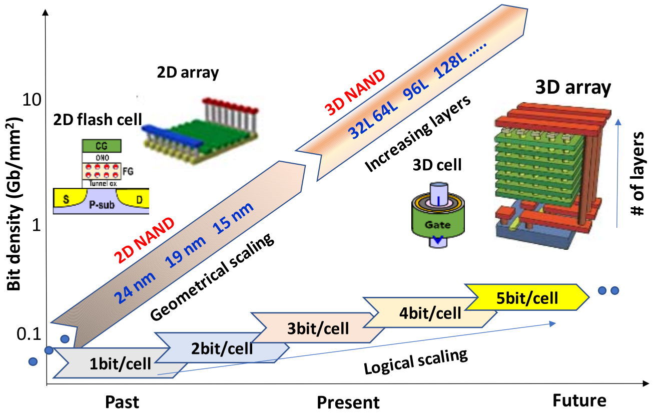

Fig. 1. Trends of bit density of planer and vertical layers NAND flash memory technology.

The FTL wear-leveling algorithms ensure that all blocks experience an equal share of program/erase (PE) cycles. Since the maximum number of PE cycles a block can endure is determined by the endurance of the weakest page within the block, the entire block is marked as "bad" when the FTL determines that a page can no longer ensure data integrity. This typically occurs when the number of bit errors exceeds the capability of error-correction codes when reading a page or when the controller fails to erase the block. This approach may lead to underutilization of storage resources, especially when there is significant variability in endurance among pages within a memory block. Hence, understanding the endurance variability among pages in a 3-D NAND memory block is critical for efficiently utilizing storage resources.

In this article, we systematically analyze the endurance of individual pages with respect to their physical location in the 3-D array. We demonstrate that pages located at the edge layers of the 3-D stack degrade faster than the pages located in the middle layers. We analyze the root cause for such

Manuscript received 16 September 2022; accepted 16 November 2022. Date of publication 29 November 2022; date of current version 3 January 2023. This work was supported by the National Science Foundation under Grant 1929099 and Grant 2007403. The review of this article was arranged by Editor T.-H. Kim. *(Corresponding author: Md Raquibuzzaman.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Alabama in Huntsville, Huntsville, AL 35899 USA (e-mail: mr0068@uah.edu; milenka@uah.edu; biswajit.ray@uah.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2022.3224420.

Digital Object Identifier 10.1109/TED.2022.3224420

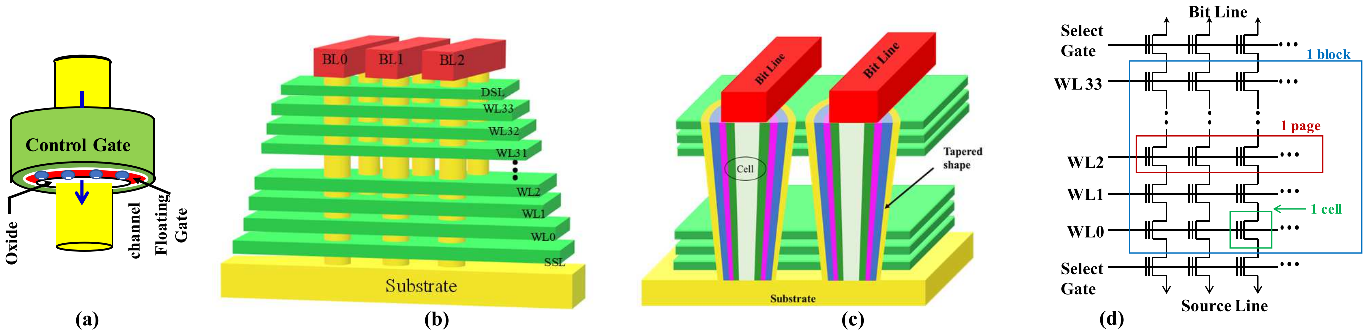

Fig. 2. (a) 3-D NAND flash memory cell. (b) Physical organization of a 3-D flash memory array. (c) Vertical cross section of the 3-D NAND memory array. (d) Transistor-level schematic of a flash memory block.

endurance variation and demonstrate that it is fundamentally related to the layer-to-layer erase-state threshold voltage (V_t) variation [10]. To counter the effects of uneven page endurance and improve the overall utilization of 3-D NAND flash memories, we propose an intrablock wear-leveling algorithm that is based on dynamic, incremental, and layer-aware downsizing of flash memory blocks. Our qualitative analysis shows that this approach can significantly improve the endurance of storage systems.

The rest of this article is organized as follows. Section II motivates this article and gives a brief background on 3-D flash memory technology. Section III describes the experimental setup and methodology used in this study. Section IV discusses the results of the endurance characterization, elucidates the root causes for the endurance variation and describes how an intrablock wear-leveling algorithm can leverage it to improve the overall block endurance. Section V provides the conclusion.

II. Background

Fig. 2(a) shows the device structure of a 3-D NAND flash memory cell, a floating-gate metal–oxide–semiconductor field-effect transistor (MOSFET) transistor with a gate-all-around cylindrical channel structure. Unlike the 2-D NAND array with a planar structure, the 3-D NAND has a layered structure consisting of alternate metal and insulating layers. Fig. 2(b) shows the physical structure of a 3-D NAND flash memory array with 34 layers. The metal layers are shown with green color, whereas intermediate insulating layers are kept transparent. Vertical yellow pillars are holes that contain flash memory cells. The holes through the 3-D stack are created using the reactive ion etching process (RIE). Each metal layer forms a word line (WL) of the memory array connecting multiple memory cells. The gate-stack of the memory cells is formed by sequentially depositing blocking oxide, charge trap layer or floating gate, and tunnel oxide along the sidewall of the vertical holes (also known as ONO—oxide–nitride–oxide). Finally, a thin layer of the poly-Si channel is deposited with a hollow core filled with oxide. Fig. 2(c) shows the vertical cross section of the 3-D array, illustrating different layers of the gate-stack. The tapered shape of the cylindrical memory hole is due to the high aspect ratio that makes the RIE inefficient at the bottom layers. Thus, memory cells in the bottom layers typically have a smaller diameter than the ones in the top layers. The unique tapered shape of the array structure leads to significant layer-to-layer variability, as reported by several recent articles [11], [12], [13], [14].

Fig. 2(d) shows the circuit diagram of the NAND flash memory array that corresponds to a single flash memory block. Each memory block consists of a fixed number of memory pages. The cells in each memory page are electrically connected through a metal WL that acts as their control gate. Each column (or string) of cells in a block is connected to a bitline (BL). Memory read and program operations are performed at the page-level granularity, whereas erase operations are performed at the block-level granularity.

III. Experimental Setup

Our experimental setup consists of a TSOP-48 socket that holds a flash memory chip, an FT2232H mini module from Future Technology Devices International (FTDI), and a workstation. The FT2232H module acts as a bridge between the workstation and the device, implementing an asynchronous 8-bit parallel interface to the device. A software package on the workstation executes the ONFI commands for sending data to the flash memory chip, erasing a block, writing a page, reading a page, or retrieving the data from the device. This hardware setup allows us to access raw memory bits without error correction.

We perform the experimental evaluation on several commercial-off-the-shelf (COTS) 3-D NAND MLC chips with the following properties: the chip capacity is 256 Gbits, the number of blocks is 2192, each block contains 1024 pages, and each page contains 18 592 bytes of data [16 384 bytes of user data with 2208 spare bytes reserved for storing out-of-band information such as error-correction codes (ECCs)]. The chip is manufactured using the 32-layer 3-D technology. A memory block consists of multiple identical subblock structures (the Micron chip has 16 subblocks). Each layer within a given subblock contains two shared pages, LSB and MSB pages. First, the LSB page is written, and then, the corresponding MSB page is written. The cell V_t distribution of the MLC memory has four different levels.

For the reliability characterization, we program the entire memory block with a random data pattern and readback the data immediately after programming. A raw bit error rate (RBER) is determined for each page in each layer of

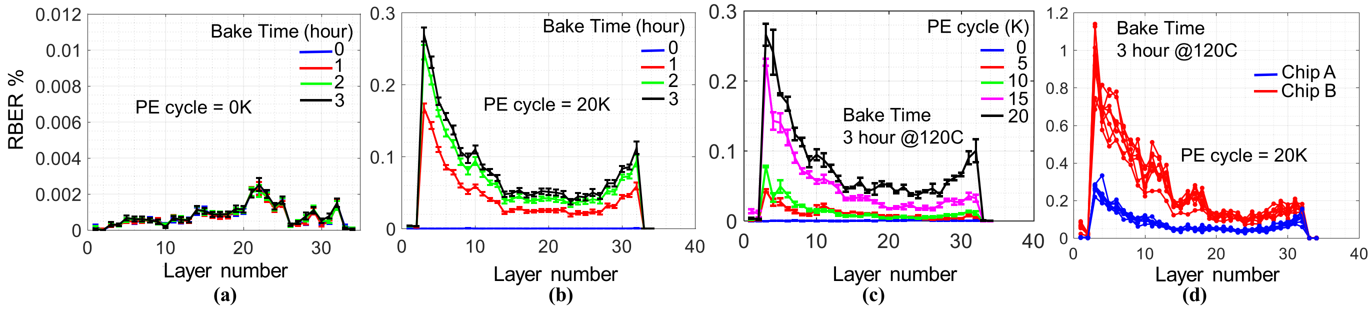

Fig. 3. RBER of pages in a block as a function of the layer number. (a) Fresh condition. (b) 20 K PE cycles. (c) Layer-dependent RBER for different PE cycle counts. (d) RBER from two different chips with similar PE cycle counts (PE = 20 K) and 3 h of bake time. Different lines represent different memory blocks of the chip.

the flash block. We take an average of all the pages in each layer to calculate the layer-specific RBER. Then, we analyze the variation of RBERs among different layers.

To understand the endurance variation among different layers as a function of the array's wear level, we perform repeated program–erase (PE) operations on different blocks and repeat the experiment to determine the layer-specific RBERs. The RBERs determined immediately after programming may not represent the actual endurance status as flash memory is expected to retain data for a long time. For more realistic endurance characterization, we perform accelerated data retention (DR) tests by baking the memory chip at a high temperature (120 °C) for 1, 2, and 3 h. We then perform read operations to determine RBERs on both fresh (PE = 0 K) and PE cycled flash memory blocks (PE = 20 K).

IV. Results and Discussion

A. Experimental Results

We first present the layer-dependent RBERs for the fresh and 20 K PE cycled blocks in Fig. 3(a) and (b), respectively. RBER is analyzed for four DR conditions (0–3 h bake time). Each point in Fig. 3 represents the average RBER of 16 pages located in that layer. The fresh blocks have a very low RBER, with the middle layers having slightly higher RBER than the edge layers, as shown in Fig. 3(a). Changing the DR conditions does not significantly impact the fresh block's RBERs. However, the RBERs of the PE cycled blocks show a very different behavior, as shown in Fig. 3(b). First, the RBERs remain very low without accelerated aging (0-h bake time). In contrast, the RBERs after 1, 2, and 3 h of baking show a characteristic profile. The RBERs of the bottom layers (L_2–L_{14}) are significantly higher than RBERs of all other layers. Next, the RBERs of pages close to the top (L_{28}–L_{31}) are also slightly higher than the RBERs of the pages located in the middle layers. Comparing the effects of DR = 3 h versus DR = 1 h, the increases in the RBERs on edge layers are significantly larger than the increases in RBERs of the middle layers.

Fig. 3(c) shows the layer-dependent RBERs for different PE cycle conditions (0, 5, 10, 15, and 20 K), while DR is fixed and set to 3 h of baking time. It is interesting to note that RBERs remain relatively low and flat for fresh memory blocks. The characteristic RBER profile becomes more prominent as we gradually increase the number of PE cycles. This result confirms that edge layers, especially bottom layers, degrade faster than the middle layers of the 3-D stack.

Fig. 3(d) profiles layer-dependent RBERs gathered from two different memory chips with the same part number. Data are collected from multiple memory blocks (~10 blocks per chip) that are exposed to the same PE cycle count and 3 h of baking. Even though the absolute RBER values vary from chip to chip, the layer-dependent RBER profile is found to be quite robust across different memory blocks and memory chips. In all cases, we find that pages in the bottom layers have the highest RBER. The RBERs gradually decrease as we move from the bottom to the upper layers; however, they slightly increase for a few top layers. In Section IV-B, we provide a root cause analysis for this characteristic RBER profile.

B. Root Cause Analysis

We postulate that layer-dependent endurance variation originates from the unique architecture of the 3-D NAND memory array, which fundamentally dictates uneven erase speed between different layers of the 3-D stack. Specifically, the root causes of endurance variation are as follows: 1) layer-dependent structural variation—uneven cell diameters and 2) layer-dependent electrical variation—uneven erase voltage distribution.

1) Uneven Cell Diameters: The monolithic 3-D fabrication process of 3-D NAND dictates a tapered shape of the cylindrical memory hole, as shown in Fig. 4(a). The tapering is caused by the memory hole's high aspect ratio, which makes the RIE inefficient at the bottom layers. Thus, memory cells at the bottom layers typically have a smaller diameter (D) than cells in the upper layers. The diameter difference (ΔD) between the top and bottom layer cells depends on the tapering angle (θ) and the height of the 3-D stack (H) as follows: $\Delta D \sim 2H \times \tan(\theta)$. The exact values for the tapering angle and the height of the 3-D stack are proprietary information. However, we can estimate ΔD by assuming typical values of array dimensions from the published literature [15]. Assuming the tapering angle, $\theta \sim 0.5°$, and height of the 3-D stack, $H \approx 32 \times 80$ nm $= 2.56$ μm, we can estimate the diameter difference $\Delta D \sim 45$ nm for 32-layer stack. Since the typical diameter of the memory

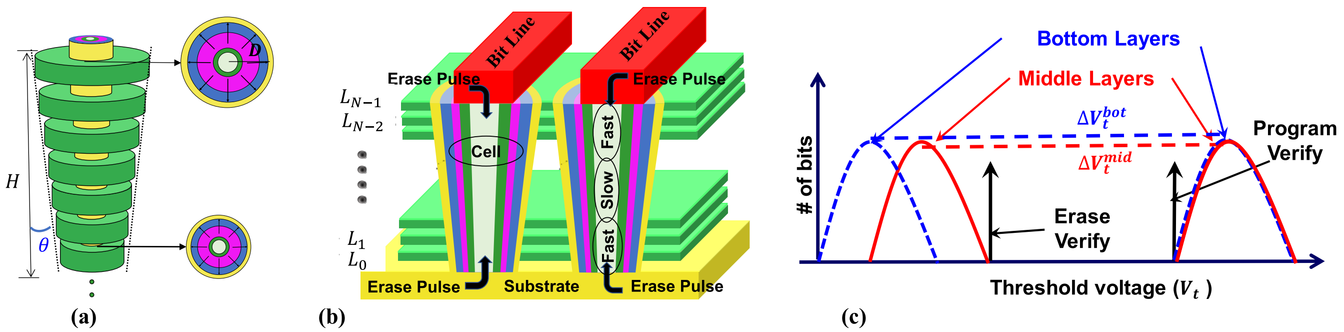

Fig. 4. (a) Tapered shape of the cylindrical memory hole. (b) Erase pulse delivery in 3-D array. (c) v_t distribution of edge (blue) and middle (red) layers. Higher erase depth in edge layers creates a higher shift in threshold voltage ($\Delta V_t^{\text{bot}} > \Delta V_t^{\text{mid}}$).

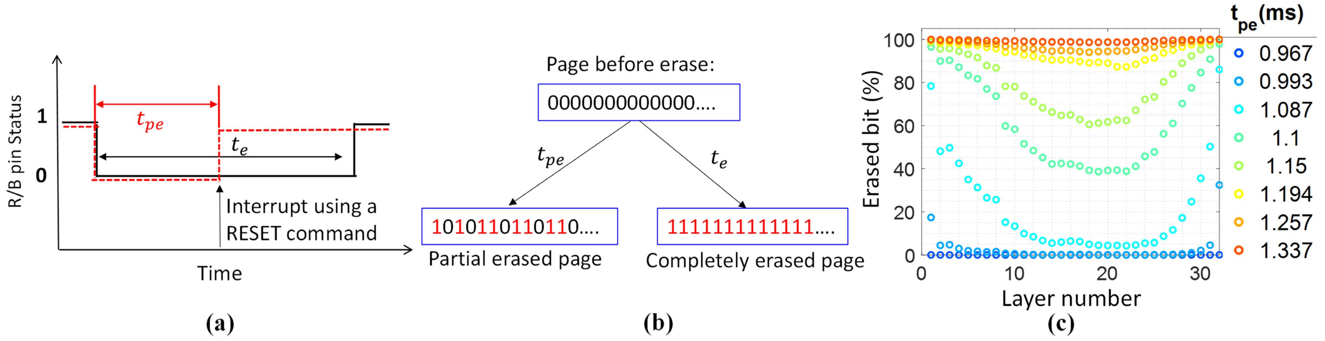

Fig. 5. (a) Block erase operation (black solid line) and partial erase operation (red dashed line). (b) Transitions of cells state through a partial erase operation. (c) Percentage of erased bits as a function of layer number for different partial erase times—“U” curves show that cells in edge layers are faster-to-erase than cells in middle layers.

hole is ∼90 nm, the estimated diameter difference is quite significant. It may result in roughly two times smaller diameter in the bottom layer cells compared to the top layer cells. Since the electric field in the oxide layers is inversely proportional to the cell diameter, cells in the bottom layer experience a significantly higher electric field during erase operation than cells in the middle and upper layers. Since the probability of damaging the oxide layer increases exponentially with the magnitude of the electrical field across the oxide [16], memory cells at the bottom layer experience faster cell degradation.

2) Uneven Erase Voltage Distribution: Erase operation takes place at a block level, meaning that all memory layers of a given block are erased together. The erase voltage is typically applied through the Si substrate and the bit lines, while all the WLs are grounded [Fig. 4(b)]. Since the erase voltage gradually propagates into the vertical memory channel through the edge layers (top and bottom), memory cells at the edge layers see high voltages during block erase, as shown in Fig. 4(b). As a result, their gate oxides are stressed a bit more relative to the cells in the middle. The degradation rate in the bottom layer cells is the highest due to the combined effects of relatively smaller diameter and higher erase voltage. The cells at the top layers, even though they have a larger diameter, experience higher erase voltage, which causes a slightly higher electric field in the gate oxide. Such uneven erase voltage in different layers is inevitable and will be amplified as more layers are stacked in future generations of 3-D NAND flash. The uneven erase voltage between different memory layers amplifies the layer-to-layer endurance difference as the memory block is PE cycled.

Fig. 4(c) shows the layer-dependent erase V_t variation using threshold voltage distributions. We choose single-level_cell (SLC) memory for simplicity, but the reasoning holds for multibit flash memories. Note that the ISPP scheme is used during page programming—multiple voltage pulses, each followed by a verify phase and program prohibit scheme. This mechanism ensures that cells in the programmed state have similar V_t distribution, irrespective of the layer number they reside in, as shown in Fig. 4(c). Since an erase operation takes place on all the layers simultaneously, the faster-to-erase layers (e.g., bottom layers) will experience a higher V_t shift compared to slow to erase layers (e.g., middle layers). In other words, the bottom layer will be deeply erased, whereas middle layers are shallowly erased. Since shallow erase improves endurance [17], middle layers exhibit better endurance compared to the bottom layers.

To validate the layer-dependent erase V_t hypothesis from above, ideally, we would measure the V_t distribution for cells in different layers after a block erase operation. However, the chip under test does not allow for such profiling. Instead, we validate this hypothesis by measuring the percentage of the erased bits as a function of the layer number by varying partial erase time. The experimental flow is given as follows. All cells in a block are programmed, and then, an erase operation is started and terminated prematurely. The black solid line in Fig. 5(a) represents the nominal block erase time t_e. We terminate the erase operation before it is completed

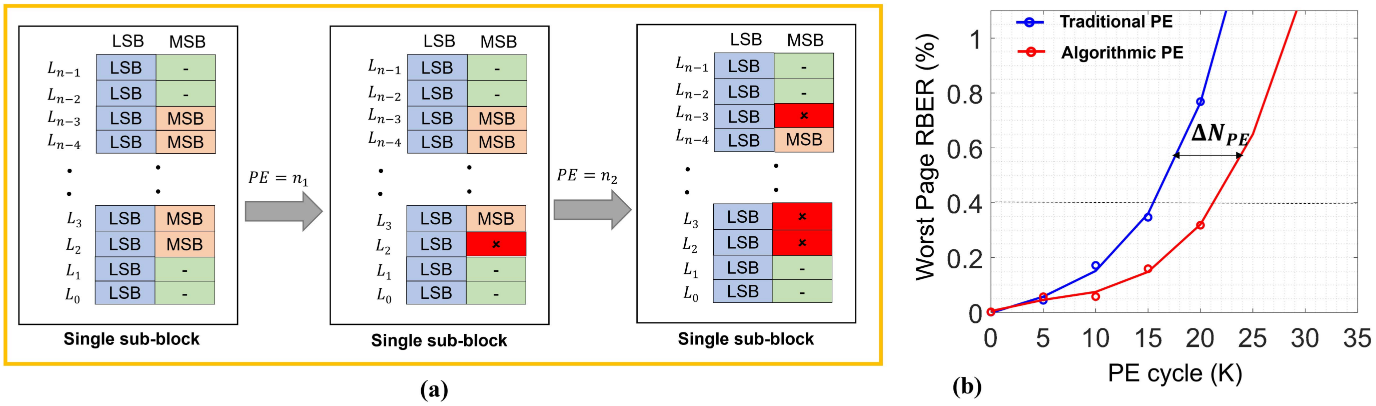

Fig. 6. (a) Proposed intrablock wear-leveling algorithm. (b) Comparison of worst case page RBER using traditional PE cycling and algorithm-based PE cycling. Symbols are from experimental evaluations and the solid lines are fit curves.

using a RESET command at a partial erase time t_{pe}, as shown by the red dashed line in Fig. 5(a). The state of the cells is determined by reading all pages in the block. Fig. 5(b) shows the state of a block after a particular partial erase time is used. Note that only faster-to-erase cells will change their state from $0 \rightarrow 1$. Slow-to-erase cells will remain at logic 0. We count the total number of erased bits in a page.

Fig. 5(c) shows the percentage of the erased cells as a function of the layer number. Different colors in Fig. 5(c) represent different partial erase times. For a given partial erase time, we observe a clear trend in the number of erased cells as a function of the layer number. Please note that the percentage of the erased bits correlates strongly with the erase speed in that layer, where a higher percentage of erased bits implies a faster erase speed. Thus, the partial erase data on the 3-D memory block illustrate that the middle layers are slow to erase compared to the top and upper layer cells.

C. Intrablock Wear Leveling

Layer-to-layer endurance variation of 3-D NAND memory may lead to severe underutilization of a memory block. For example, an entire flash block will be marked as bad even though just a few worn-out pages on the bottom layer may reach RBER that exceeds levels that can be corrected using ECCs. As the number of layers and thus pages in a block keeps increasing with each new generation of 3-D NAND flash chips, this uneven layer-to-layer endurance variation may be a limiting factor resulting in either underutilization of storage media and/or reduced data integrity. Reduced data integrity will require a more sophisticated ECC, limiting the throughput and increasing the latency of flash operations. To improve the utilization of storage media and its endurance, we propose an intrablock wear-leveling algorithm that exploits the observed endurance variation within 3-D flash arrays.

Fig. 6 shows the proposed algorithm. We explain our algorithm using an MLC memory block, but the proposed algorithm applies to all other multibit flash memories. Fig. 6 shows an n-layer 3-D stack corresponding to a subblock, where a WL is shared by corresponding LSB and MSB pages. The top and bottom two layers operate exclusively in the SLC mode. The remaining layers contain shared memory pages. MSB pages are usually more erroneous than LSB pages as they involve high V_t cells, which are more affected by retention loss. Thus, MSB pages of the bottom layer (L_2) will have the highest RBER for a given PE cycle condition. Thus, the endurance of a memory block will be limited by the MSB pages of $L_{2,}$ which will fail ECC after a certain number of PE cycles. Thus, we propose to keep using the memory block by skipping the programming of the MSB pages of L_2 after PE $= n_1$. This parameter is determined by measuring the RBER; when the RBER reaches a certain fraction of the maximum correctable RBER. Similarly, after n_2 PE cycles, we need to skip programming of MSB pages of the next layer from the bottom (L_3). In implementing the algorithm, one may skip a group of layers instead of just one to keep a safety margin for data integrity. Similarly, if the top layer also shows very high RBER, which may be the case for specific chips, we need to skip MSB pages of top layers as well, as shown in Fig. 6(a). A more sophisticated algorithm can involve a series of PE cycles (e.g., $n_2, n_3, \ldots$) that trigger the logical reshaping of the flash memory block.

The critical design parameters for the implementation of the proposed algorithm are the threshold PE cycle counts n_1, n_2, and n_3, as well as the number and location of pages to be skipped/excluded at each step. These parameters need to be predetermined for each family of chips that use a given technology process through a detailed characterization. This characterization is necessary because layer-dependent RBER may vary among chips from different families and/or manufacturers. It will be conducted once by either the chip manufacturer or a system integrator. The flash memory controller is in charge of implementing logical reshaping of flash memory blocks, considering these parameters and the current state of each flash block.

In order to illustrate the improvement in endurance achieved by the proposed algorithm, we perform the following experiment. A flash memory block is exposed to traditional block management. All pages of the block are stressed by repeated PE cycling. Once the RBER in any page of the block reaches the threshold level set to 0.2% in this experiment, the block is marked as bad, and the corresponding PE cycle count is reported.

Another flash memory block is managed by a simplified version of the proposed algorithm. All pages of the block are repeatedly stressed until we reach $n_1 = 5$ K PE cycles. Please note that worst case RBER after 5 K PE cycles is only 0.1%, which is 50% below the threshold RBER level of 0.2%. We made this choice of $n_1 = 5$ K in order to keep extra RBER margin to account for the block-to-block or chip-to-chip variation. At this point, the flash memory block is reshaped, and the MSB pages in a group of bottom layers ($L_2 L_6$) are excluded for further use. The next logical reshaping of the flash memory block takes place when the number of PE cycles reaches PE $= n_2 = 10$ K. At this moment, the MSB pages residing in a group of top layers ($L_{29} L_{31}$) are excluded from further use. The remaining pages are further stressed until the total number of PE cycles reaches PE $= 20$ K. At this moment, the experiment is concluded.

Fig. 6(b) shows the results of the experimental evaluation. We plot the worst case page RBER from a memory block as a function of PE cycle count for the traditional case (blue line) and for the simplified block management described above (red line). Using the traditional block management, the threshold RBER of 0.2% is reached when PE $= 11$ K. The proposed simplified two-step reshaping block management reaches the threshold RBER when PE $= 16$ K. With the proposed implementation, we find that the maximum PE cycle of a block can be enhanced by ~45% (from 11 to 16 K). The tradeoff for this endurance enhancement is a reduced block size (12.5% reduction in the current implementation) at its end of life. The improvement of endurance will be significantly higher if the algorithm is implemented with finer granularity or with more than two reshaping steps.

V. Conclusion

We observe that pages within a 3-D NAND flash block show a large variation in their endurance. These variations are correlated with the geometrical location of the pages within a vertical 3-D structure. We show that the endurance of memory pages at the top and bottom layers are significantly lower than that of the pages in the middle layers. Our findings suggest that more sophisticated block management schemes can be developed that will consider layer-to-layer wear-out differences within a flash memory block. Such management schemes will improve the endurance and utilization of storage media.

References

[1] C. M. Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaita, and A. Visconti, “Reviewing the evolution of the NAND flash technology,” *Proc. IEEE*, vol. 105, no. 9, pp. 1609–1633, Sep. 2017, doi: 10.1109/JPROC.2017.2665781.

[2] A. Goda and K. Parat, “Scaling directions for 2D and 3D NAND cells,” in *IEDM Tech. Dig.*, Dec. 2012, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM.2012.6478961.

[3] A. Goda, “3-D NAND technology achievements and future scaling perspectives,” *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1373–1381, Apr. 2020, doi: 10.1109/TED.2020.2968079.

[4] K. Parat and A. Goda, “Scaling trends in NAND flash,” in *IEDM Tech. Dig.*, Dec. 2018, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM.2018.8614694.

[5] R. Micheloni, S. Aritome, and L. Crippa, “Array architectures for 3-D NAND flash memories,” *Proc. IEEE*, vol. 105, no. 9, pp. 1634–1649, Sep. 2017, doi: 10.1109/JPROC.2017.2697000.

[6] D. Resnati, A. Goda, G. Nicosia, C. Miccoli, A. S. Spinelli, and C. M. Compagnoni, “Temperature effects in NAND flash memories: A comparison between 2-D and 3-D arrays,” *IEEE Electron Device Lett.*, vol. 38, no. 4, pp. 461–464, Apr. 2017, doi: 10.1109/LED.2017.2675160.

[7] M.-C. Yang, Y.-M. Chang, C.-W. Tsao, P.-C. Huang, Y.-H. Chang, and T.-W. Kuo, “Garbage collection and wear leveling for flash memory: Past and future,” in *Proc. Int. Conf. Smart Comput.*, Nov. 2014, pp. 66–73, doi: 10.1109/SMARTCOMP.2014.7043841.

[8] C. Liu, J. Kotra, M. Jung, and M. Kandemir, “PEN: Design and evaluation of partial-erase for 3D NAND-based high density SSDs,” in *Proc. 16th USENIX Conf. File Storage Technol. (FAST)*, 2018, pp. 67–82. Accessed: Jul. 07, 2021. [Online]. Available: https://www.usenix.org/conference/fast18/presentation/liu

[9] S. Wang, F. Wu, C. Yang, J. Zhou, C. Xie, and J. Wan, “WAS: Wear aware superblock management for prolonging SSD lifetime,” in *Proc. 56th Annu. Design Autom. Conf.*, Jun. 2019, pp. 1–6.

[10] M. Raquibuzzaman, M. M. Hasan, A. Milenkovic, and B. Ray, “Layer-to-layer endurance variation of 3D NAND flash memory,” in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2022, pp. 1–5, doi: 10.1109/IRPS48227.2022.9764441.

[11] C.-H. Hung et al., “Layer-aware program-and-read schemes for 3D stackable vertical-gate BE-SONOS NAND flash against cross-layer process variations,” *IEEE J. Solid-State Circuits*, vol. 50, no. 6, pp. 1491–1501, Jun. 2015, doi: 10.1109/JSSC.2015.2413841.

[12] K. T. Kim, S. W. An, H. S. Jung, K.-H. Yoo, and T. W. Kim, “The effects of taper-angle on the electrical characteristics of vertical NAND flash memories,” *IEEE Electron Device Lett.*, vol. 38, no. 10, pp. 1375–1378, Oct. 2017, doi: 10.1109/LED.2017.2747631.

[13] U. M. Bhatt, S. K. Manhas, K. Nayyar, A. Kumar, and M. Pakala, “Mitigating pillar-to-pillar variability of ground select transistor in 3-D nand flash memory,” *IEEE Trans. Electron Devices*, vol. 67, no. 10, pp. 4152–4157, Oct. 2020, doi: 10.1109/TED.2020.3012927.

[14] P. Kumari, S. Huang, M. Wasiolek, K. Hattar, and B. Ray, “Layer-dependent bit error variation in 3-D NAND flash under ionizing radiation,” *IEEE Trans. Nucl. Sci.*, vol. 67, no. 9, pp. 2021–2027, Sep. 2020, doi: 10.1109/TNS.2020.3014261.

[15] A. J. Walker, “A rigorous 3-D NAND flash cost analysis,” *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 619–625, Nov. 2013, doi: 10.1109/TSM.2013.2283274.

[16] K. F. Schuegraf and C. Hu, “Effects of temperature and defects on breakdown lifetime of thin SiO_2 at very low voltages,” *IEEE Trans. Electron Devices*, vol. 41, no. 7, pp. 1227–1232, Jul. 1994, doi: 10.1109/16.293352.

[17] J. Jeong, S. S. Hahn, S. Lee, and J. Kim, “Lifetime improvement of NAND flash-based storage systems using dynamic program and erase scaling,” in *Proc. 12th USENIX Conf. File Storage Technol.*, Feb. 2014, pp. 61–74.