Improving 3D NAND Flash Memory Lifetime by Tolerating Early Retention Loss and Process Variation

Yixin Luo[†] Saugata Ghose[†] Yu Cai[†] Erich F. Haratsch[‡] Onur Mutlu[§†]

[†]Carnegie Mellon University [‡]Seagate Technology [§]ETH Zürich

ABSTRACT

Compared to planar NAND flash memory, 3D NAND flash memory uses a new flash cell design, and vertically stacks dozens of silicon layers in a single chip. This allows 3D NAND flash memory to increase storage density using a much less aggressive manufacturing process technology than planar NAND. The circuit-level and structural changes in 3D NAND flash memory significantly alter how different error sources affect the reliability of the memory. Our goal is to (1) identify and understand these new error characteristics of 3D NAND flash memory, and (2) develop new techniques to mitigate prevailing 3D NAND flash errors.

In this paper, we perform a rigorous experimental characterization of real, state-of-the-art 3D NAND flash memory chips, and identify *three new error characteristics* that were not previously observed in planar NAND flash memory, but are fundamental to the new architecture of 3D NAND flash memory.

(1) 3D NAND flash memory exhibits *layer-to-layer process variation*, a new phenomenon specific to the 3D nature of the device, where the average error rate of each 3D-stacked layer in a chip is significantly different. We are the *first* to provide detailed experimental characterization results of layer-to-layer process variation in real flash devices in open literature. Our results show that the raw bit error rate in the middle layer can be 6× the error rate in the top layer.

(2) 3D NAND flash memory experiences *early retention loss*, a new phenomenon where the number of errors due to charge leakage increases *quickly within several hours* after programming, but then increases at a much slower rate. We are the *first* to perform an extended-duration observation of early retention loss over the course of 24 days. Our results show that the retention error rate in a 3D NAND flash memory block quickly increases by an order of magnitude within ∼3 hours after programming.

(3) 3D NAND flash memory experiences *retention interference*, a new phenomenon where the rate at which charge leaks from a flash cell is dependent on the amount of charge stored in neighboring flash cells. Our results show that charge leaks at a lower rate (i.e., the retention loss speed is slower) when the neighboring cell is in a state that holds more charge (i.e., a higher-voltage state).

Our experimental observations indicate that we must revisit the error models and error mitigation mechanisms devised for planar NAND flash, as they are no longer accurate for 3D NAND flash behavior. To this end, we develop *new analytical models* of (1) the layer-to-layer process variation in 3D NAND flash memory, and (2) retention loss in 3D NAND flash memory. Our models estimate the raw bit error rate (RBER), threshold voltage distribution, and the *optimal read reference voltage* (i.e., the voltage at which RBER is minimized when applied during a read operation) for each flash page. Both models are useful for developing techniques to mitigate raw bit errors in 3D NAND flash memory.

Motivated by our new findings and models, we develop four new techniques to mitigate process variation and early retention loss in 3D NAND flash memory. Our first technique, LaVAR, reduces process variation by fine-tuning the read reference voltage independently for each layer. Our second technique, LI-RAID, improves reliability by changing how pages are grouped under the RAID (Redundant Array of Independent Disks) error recovery technique, using information about layer-to-layer process variation to reduce the likelihood that the RAID recovery of a group could fail significantly earlier during the flash lifetime than recovery of other groups. Our third technique, ReMAR, reduces retention errors in 3D NAND flash memory by tracking the retention age of the data using our retention model and adapting the read reference voltage to data age. Our fourth technique, ReNAC, adapts the read reference voltage to the amount of retention interference to re-read the data after a read operation fails. These four techniques are complementary, and can be combined together to significantly improve flash memory reliability. Compared to a state-of-the-art baseline, our techniques, when combined, improve flash memory lifetime by 1.85×. Alternatively, if a NAND flash manufacturer wants to keep the lifetime of the 3D NAND flash memory device constant, our techniques reduce the storage overhead required to hold error correction information by 78.9%.

For more information on our new experimental characterization of modern 3D NAND flash memory chips and our proposed models and techniques, please refer to the full version of our paper [1].

KEYWORDS

3D NAND flash memory; error correction; fault tolerance; reliability; solid-state drives; storage systems

ACM Reference Format:

Y. Luo et al. 2018. Improving 3D NAND Flash Memory Lifetime by Tolerating Early Retention Loss and Process Variation. In *SIGMETRICS '18 Abstracts: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems Abstracts, June 18–22, 2018, Irvine, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3219617.3219659

SIGMETRICS '18 Abstracts, June 18–22, 2018, Irvine, CA, USA

ACM ISBN 978-1-4503-5846-0/18/06.
https://doi.org/10.1145/3219617.3219659

REFERENCES

[1] 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," *POMACS*, 2018.