Received 4 November 2019; revised 19 December 2019; accepted 5 January 2020. Date of publication 10 January 2020; date of current version 24 January 2020.
The review of this article was arranged by Editor M. Liu.

Digital Object Identifier 10.1109/JEDS.2020.2965648

First Evidence of Temporary Read Errors in TLC 3D-NAND Flash Memories Exiting From an Idle State

CRISTIAN ZAMBELLI [1] (Member, IEEE), RINO MICHELONI [2] (Senior Member, IEEE), SALVATRICE SCOMMEGNA [2], AND PIERO OLIVO [1]

1 Dipartimento di Ingegneria, Università di Ferrara, 44122 Ferrara, Italy
2 Flash Signal Processing Labs, Microsemi Corporation (a Microchip Company), 20871 Vimercate, Italy

CORRESPONDING AUTHOR: C. ZAMBELLI (e-mail: cristian.zambelli@unife.it)

This work was supported in part by the Università degli Studi di Ferrara (UniFE)-Camera di Commercio, Industria, Artigianato e Agricoltura (CCIAA) 2018 grant "Bando 2018 per Progetti di Ricerca Finanziati con il Contributo Della Camera di Commercio, Industria, Artigianato e Agricoltura."

ABSTRACT This paper presents a new reliability threat that affects 3D-NAND Flash memories when a read operation is performed exiting from an idle state. In particular, a temporary large increase of the fail bits count is reported for the layers read as first after a sequence of program/verify and a idle retention phase. The phenomenon, hereafter called Temporary Read Errors (TRE), is not due to a permanent change of cell threshold voltage between the program verify and the following read operations, but to its transient instability occurring during the idle phase and the first read operations performed on a block. The experimental analysis has been performed on off-the-shelf gigabit-array products to characterize the dependence on the memory operating conditions. The TRE is found to be strongly dependent on the page read, on the read temperature and on the time delay between the first and the second read after the idle state. To emphasize its negative impact at system-level, we have evaluated the induced performance drop on Solid State Drives architectures.

INDEX TERMS 3D-NAND flash, characterization, fail bits, read errors, reliability, solid-state drives (SSD).

I. INTRODUCTION

The 3D-NAND Flash memory technology is now a data storage stronghold in mobile and Solid State Drives (SSD) applications. So far, several integration options for 3D-NAND reached a fabrication yield and a maturity level good enough for mass production, outmatching at the same time planar technologies in terms of bit density and performance [1], [2], [3]. Concerning reliability, several benchmarks showed that 3D-NAND Flash are better than state-of-the-art 2D NAND Flash [4], [5], [6], although all the issues concerning the cells threshold voltage (V_T) instabilities like Random Telegraph Noise (RTN), cycling and retention failures are inherited to a large extent [7].

Besides those intrinsic mechanisms, there are additional reliability detractors related to the structure of the 3D-NAND cells. The use of a poly-silicon (poly-Si) material for the cells' channel and their vertical integration in relatively long pillars exposes the V_T distributions to unwanted broadening and shifts [1], [8], [9], [10], [11], [12].

In this work, for the first time in 3D-NAND Flash gigabit array products, we show the presence of a reversible phenomenon, hereafter denoted as Temporary Read Errors (TRE). Such effect occurs during the read operation that impacts on the first group of cells read in a memory block when exiting from an idle state after a program/verify operation. The TRE issue is due to a transient instability of cell V_T occurring during the idle phase and the first read operations performed on a block.

II. EXPERIMENTAL SETUP AND METHODS

The electrical characterization of the TRE has been performed on off-the-shelf Triple Level Cell (TLC) 3D-NAND Flash products featuring M layers ($M \geq 32$) with the array architecture and V_T distribution coding mapped in three TLC pages (i.e., Lower Significant Bit - LSB, Central Significant

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

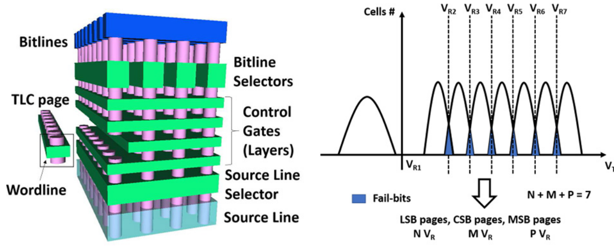


FIGURE 1. TLC 3D-NAND Flash architecture considered in this work (left) and definition of fail bits according to their V_T distribution coding (right).

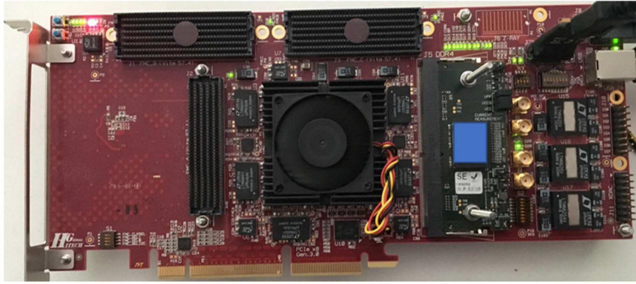


FIGURE 2. Automated test equipment exploited for fail bits extraction (3D-NAND Flash highlighted in blue).

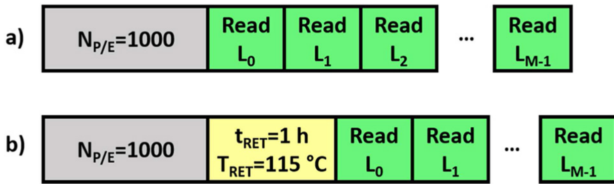


FIGURE 3. Test schemes exploited for TRE assessment.

Bit - CSB, and Most Significant Bit - MSB) as described in Fig. 1. The number of layers actually represents the number of cells stacked in the vertical direction along the NAND string. To consider the typical manufacturing variability, the tests have been repeated on several memory blocks and dies of the same product. The measurements were taken by a custom automated test equipment that allows extracting the number of fail bits after a read operation in every memory location (see Fig. 2). To rule out topological artifacts and interference on the TRE, we have programmed the samples with TLC random patterns and we disabled any error correction functionality of the chip. No modifications were applied to the standard working voltages of the devices and no special test modes were exploited to filter the fail bits count distribution.

The fail bits extraction has been performed for all physical layers (L_i) in the two following conditions (see Fig. 3): *a)* after 1000 program/erase cycles ($N_{P/E}$) and reading immediately at the end of the program operation starting from L_0 and sequentially reaching L_{M-1}; *b)* after $N_{P/E} = 1000$ and an idle state (characterized by a retention time, $t_{RET} = 1$ h, and temperature $T_{RET} = 115$ °C). The values of t_{RET} and T_{RET}

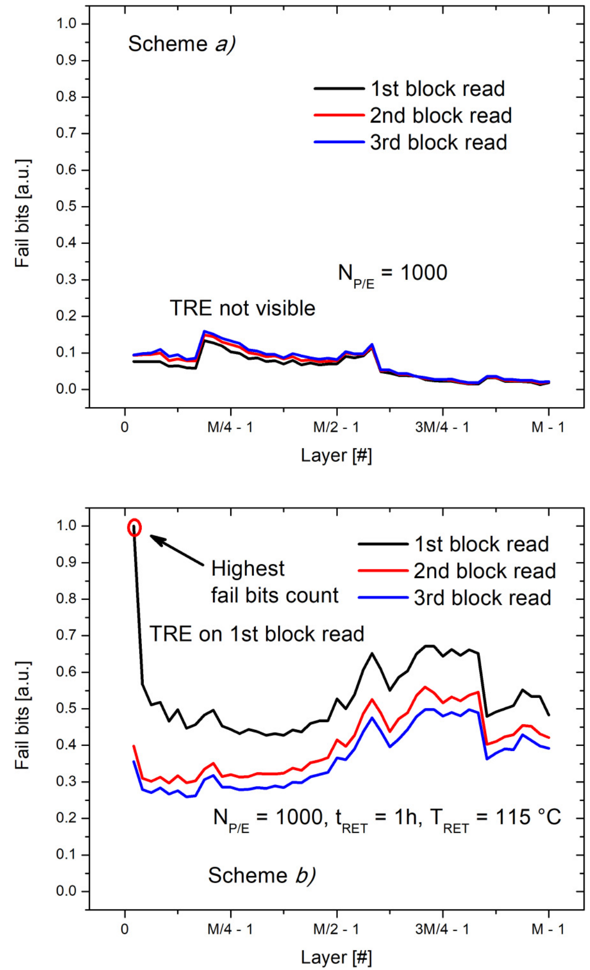


FIGURE 4. Fail bits retrieved on three consecutive block reads. In scheme *a)* the TRE is not appreciable, whereas in scheme *b)* it is shown that the first block read is the most affected.

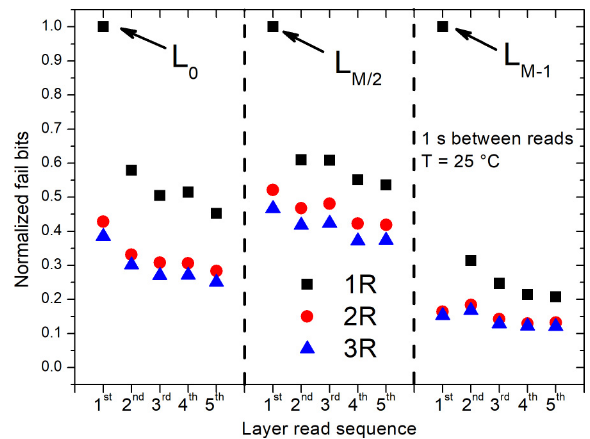


FIGURE 5. Normalized fail bits considering different topological locations of the layer read as first in a 3D-NAND Flash block. When read starts from L_{M-1}, the consecutive reads begin from L_0.

are chosen accordingly to the JEDEC tests specifications for the NAND Flash enterprise qualification procedure [13]. Same standards are adopted for P/E cycling which resulted

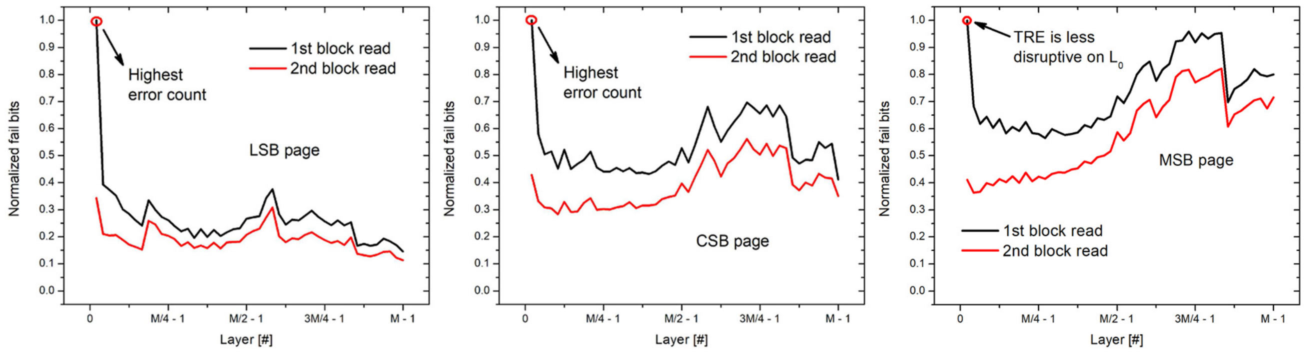


FIGURE 6. TRE dependency on the TLC page (i.e., LSB, CSB, and MSB page) and on the layer position in the block. Test *b)* has been considered for this analysis.

in an equivalent 500 hours cycle time at an 85 °C temperature for all tests. Particular attention has been paid to avoid any cross-temperature artifacts that could erroneously increase the fail bit count [14].

III. EXPERIMENTAL CHARACTERIZATION RESULTS

Fig. 4 compares the results of the test *a)* with test *b)* by evidencing the fail bits count retrieved for each physical layer in a block after three consecutive block reads. In the former test there is no perceivable difference in terms of fail bits between the block reads, whereas in the latter we appreciate a high fail bit count in the first layer of the block, with a general trend exposing the first block read as the one with the highest fail bits count. This is an evidence of the TRE phenomenon in 3D-NAND Flash memories if the read operation is performed after a reasonable amount of time in which the block has been in an idle state.

To exclude any topological dependency and to demonstrate a general presence of the phenomenon in all layers of 3D-NAND Flash blocks we repeated the readout of five adjacent pages in the vertical direction for three times considering the test case *b)* (i.e., after an idle state). The experiment is performed on different read starting positions (i.e., L_0, $L_{M/2-1}$, and L_{M-1}, respectively). The number of fail bits is normalized with respect to the maximum count retrieved for each experiment. A large difference is perceived between the first and the second read (1R, 2R), but not between the second and the third (2R, 3R) for all cases (see Fig. 5). This confirms that the TRE is visible mainly on the first read operation thus exposing its temporary nature, while proving its independence on the topological location of the first physical layer read. Fig. 6 shows that TRE is appreciable in all TLC pages with a larger impact evidenced for the LSB. The normalization of the fail bits is performed on the maximum count retrieved per page.

Fig. 7 shows, for the cases *a)* and *b)*, the boxplot of the fail bits count distribution per layer normalized to the maximum number retrieved in the test case. Each boxplot refers to a data population of about 400 kbits spanning over all TLC pages on multiple tested blocks and dies. The figure highlights two fundamental contributions in the explanation

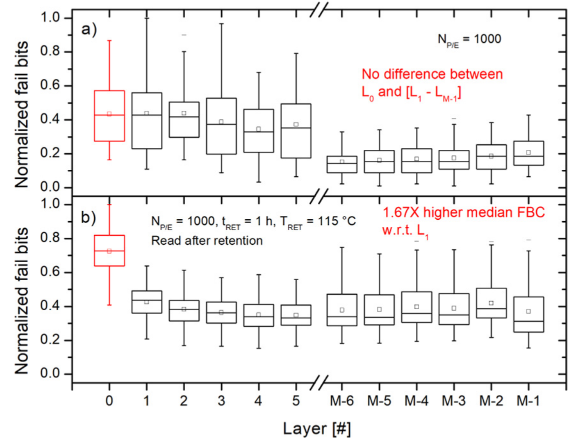


FIGURE 7. Boxplots of the fail bits count distribution (several dies and blocks) per layer normalized to the maximum number retrieved in each test case. The fail bits count distribution of the first layer read is highlighted.

of the TRE: *i)* if the read operation is performed after a reasonable amount of time in which the block has been in an idle state (case *b*), we observe an increase of the fail bits count for the first layer read within a block (up to 1.67 times, considering the median value of the distribution, with respect to the next layer in read sequence), gradually assessing to a more stable condition; *ii)* the fail bits count in the first layer read in the sequence is comparable to the following ones if the read operation is performed immediately after the last program verify operation (case *a*, where the difference in the normalized fail bits count for group $L_0 - L_5$ and group $L_{M-6} - L_{M-1}$ is an artifact introduced by the extremely low fail bits statistics).

Fig. 8 confirms that the TRE depends on the working conditions of the memory. By comparing the average fail bits count on the layer read as first and that on the rest of the block we note again that in case *a)* no difference is found, whereas in case *b)*, two effects are evidenced: the TRE effects in the layer read as first and a general fail bits count increase ascribed to unavoidable V_T degradation induced by retention effects [15], [16], [17].

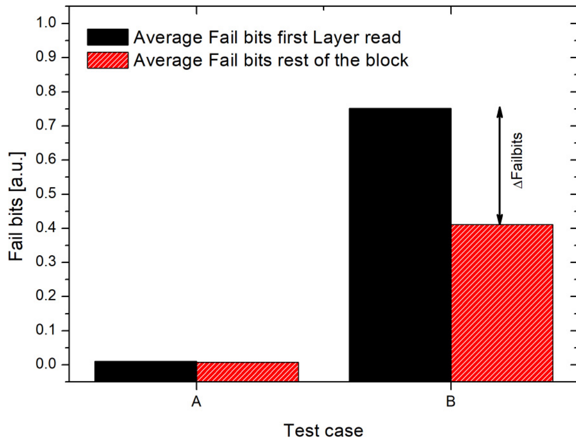


FIGURE 8. **Fail bits count for the four cases of Fig. 3. Full pattern shows the average calculated on the first layer read, whereas the dashed pattern shows the average of all the other layers in the block. The difference is highlighted as Δ Fail bits.**

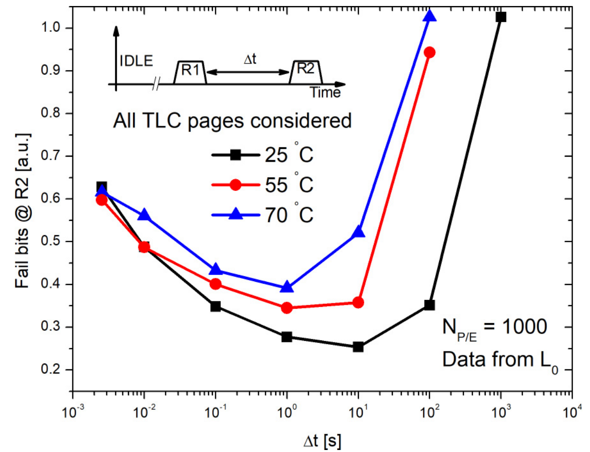


FIGURE 9. **Fail bits at second read (*R2*) as a function of the elapsed time (Δ*t*) after a first read (*R1*) for three different temperatures.**

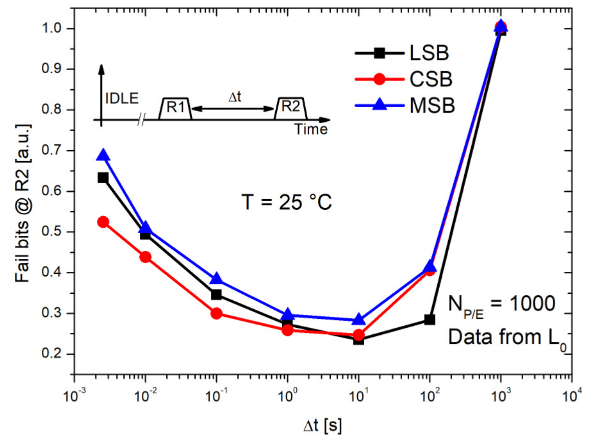


FIGURE 10. **Fail bits at second read (*R2*) as a function of the elapsed time (Δ*t*) after a first read (*R1*) for three different TLC pages.**

Finally, to monitor the TRE relationship with time and temperature, we performed an additional experiment on a single physical layer consisting of a set of two consecutive reads separated by a Δt delay. Three different read temperatures were considered: 25 °C, 55 °C, and 70 °C. Each point of Fig. 9 represents the fail bits count retrieved after the second read ($R2$) normalized with respect to that retrieved after the first read ($R1$), as a function of the Δt delay between the two reads. For each experimental point, $R1$ is performed after an idle state ($t_{IDLE} = 2$ h) ensuring similar starting conditions for each couple of $R1$ and $R2$ reads. The results evidence that the $R2$ fail bits count decreases as Δt increases, reaching a temperature dependent minimum in the $[1s; 10s]$ region. We must note that, as Δt increases above 100s, the number of fail bits retrieved after $R2$ increases again since there is a re-occurrence of the TRE because $R2$ is performed in similar conditions of $R1$ (i.e., after a sufficiently long idle state). The dependence on the TLC page is also exposed in Fig. 10. To consider the typical manufacturing variability, the tests have been repeated on several memory blocks and dies of the same product with similar outcome.

IV. DISCUSSION ON THE TRE NATURE

The characterization results evidenced that the phenomenon is unlikely ascribed only to a general V_T instability that occurs in presence of stochastic charge trapping/de-trapping events in the tunnel-oxide/interface like for RTN [7], [18], [19] since the fail bits increase is localized in the first group of cells read (see Fig. 5). Indeed, the RTN is an intrinsic characteristic common to all the cells in the array and therefore it is a fundamental contributor of the fail bits count distribution for every layer. The TRE reversibility and reproducibility also allows ruling out any modification of the cells' trapped charge like that caused by the early retention loss [20] or by program interference [21]. Given those assumptions, we suggest that TRE is due to a transient phenomenon occurring during the idle phase and the first read operations performed on a block and altering cell V_T.

Such mismatch could be attributed to several physical mechanisms already debated in literature [10], [11], [12] related to the peculiar behavior of the poly-Si channel. Although all the tests in those works were performed either on single cells/test element groups or on TCAD simulations, we found some evident similarities especially in the transient current instabilities phenomenology. The results obtained by the characterizations in [11], considering a cell structure similar to the one integrated in our 3D NAND Flash samples (i.e., poly-Si channel and a floating body), show the presence of a current transient instability originated by grain-boundary (GB) traps in poly-Si. During the program operation, there is a strong GB trapping before the program verify that will lead to a different discharge transient compared to what normally happens during read. In this case, an I_{BL} mismatch is produced during sensing that affects V_T distributions and the fail bits count. In [10], a similar explanation of the phenomenon is given except for the inclusion in the I_{BL} instability culprits of the blocking dielectric traps. However, even if we cannot rule out the impact of those contributors, our

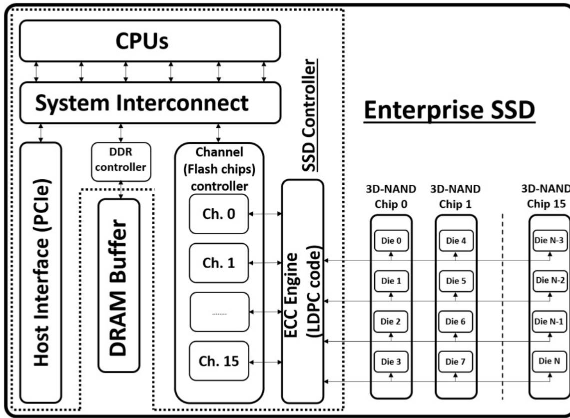


FIGURE 11. SSD architecture simulated in this work for system-level assessment of TRE impact.

experimental results are comparable to the conditions where their measured instability is dominated by GB traps in the poly-Si channel. Another potential source of I_{BL} mismatch is discussed in [12], where the down-coupling phenomenon triggered by a temporary floating-state of the channel during the program verify is evidenced. Additional investigations on large 3D-NAND Flash array structures are to be performed to verify whether, besides those physical explanations, there are other extrinsic causes related to the time constants necessary to charge/discharge long poly-Si pillars when exiting from idle state.

V. SYSTEM-LEVEL IMPACT OF TRE

To assess the impact of the TRE at system-level we have simulated an entepise-class TLC 3D-NAND Flash-based SSD architecture (see Fig. 11) with the emulator described in [22]. The main features of emulated drive are summarized in Table 1. The SSD controller implements a powerful Low-Density Parity Check (LDPC) Error Correction Code (ECC) that corrects up to 120 fail bits over 1KB sized codewords with soft-decoding capabilities. This latter option is enabled when the number of fail bits is higher than the ECC correction strength. We must recall that soft-decoding should be avoided as it is extremely time consuming then causing significant performance drops on the drive [23]. This calls for the evaluation of the soft-decoding trigger rate. Fig. 12 shows that without accounting for the presence of the TRE on the 3D-NAND Flash memories inside the SSD there would be a 3% soft-decoding trigger rate. However, that turns into an underestimation since the presence of the TRE will materialize in an increase up to 11%. Such result indicates both a reliability and a possible performance loss of the drive.

To assess that, we evaluated the impact on the SSD latency figure of merit at different percentiles of the cumulative distribution function (CDF) retrieved after 1 million read and write accesses to the drive (see Fig. 13). If the 99th percentile of the CDF is considered, we can see a 2 ms latency difference whether or not the TRE is accounted in the simulations. In this case, we can expose that TRE will represent a severe limiter for fast storage applications and will pose serious challenges for device and system mitigation in order to keep the SSD quality-of-service to an acceptable level.

TABLE 1. SSD features considered in the simulations assessing the TRE system-level impact.

Capacity	2 TB
Host interface	PCIe gen3 X4
Queue Depth (QD)	32 (high bandwidth)
Channels	16
NAND Flash technology	TLC 3D-NAND
NAND Flash page	16 KB+parity (4 sectors 4320B)
DRAM cache	2 GB
ECC	LDPC up to 120 bits/1kB
Advanced ECC protection	1 bit soft decoding
Over-Provisioning	30%
Write Amplification	2.4
Workload	4kB Random 50% Write - 50% Read

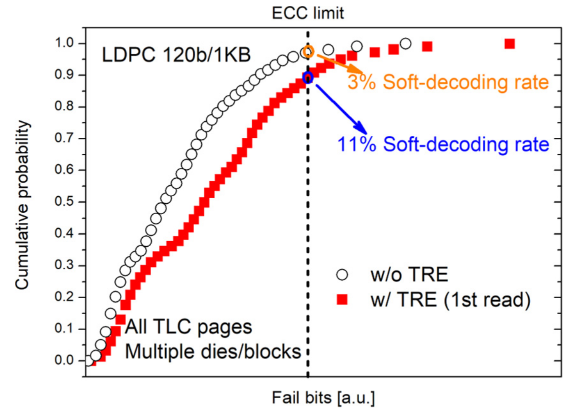


FIGURE 12. Soft-decoding trigger rate increase induced by TRE. The considered ECC for soft-rate assessment is a 120b/1KB LDPC.

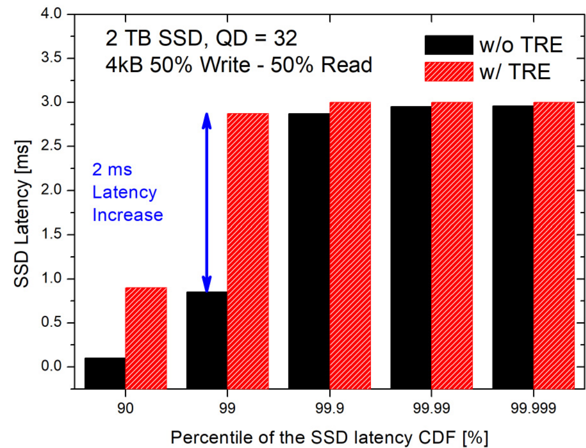


FIGURE 13. Impact on the SSD latency if the TRE would not be mitigated at system-level.

VI. CONCLUSION

In this work we exposed the TRE phenomenon as a potential reliability issue in 3D-NAND Flash technology through an experimental characterization of the fail bits count statistics in different working corners. Its occurrence is attributed to a transient phenomenon occurring during the idle phase and the

first read operations performed on a block and altering cell V_T. The time and temperature dependency of TRE suggests possible mitigation to be applied either at device or system level. As an example, by repeating after few seconds the first read operation before considering the data as valid.

REFERENCES

[1] Y. Fukuzumi *et al.*, "Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2007, pp. 449–452, doi: 10.1109/IEDM.2007.4418970.

[2] D. Kang *et al.*, "256 Gb 3 b/cell V-Nand flash memory with 48 stacked WL layers," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 210–217, Jan. 2017, doi: 10.1109/JSSC.2016.2604297.

[3] T. Tanaka *et al.*, "7.7 A 768Gb 3b/cell 3D-floating-gate NAND flash memory," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Jan. 2016, pp. 142–144, doi: 10.1109/ISSCC.2016.7417947.

[4] A. Goda and K. Parat, "Scaling directions for 2D and 3D NAND cells," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2012, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM.2012.6478961.

[5] 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.

[6] N. Righetti and G. Puzzilli, "2D vs 3D NAND technology: Reliability benchmark," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, Oct. 2017, pp. 1–6, doi: 10.1109/IIRW.2017.8361235.

[7] G. Nicosia, A. Goda, A. S. Spinelli, and C. M. Compagnoni, "Impact of cycling on random telegraph noise in 3-D NAND flash arrays," *IEEE Electron Device Lett.*, vol. 39, no. 8, pp. 1175–1178, Aug. 2018, doi: 10.1109/LED.2018.2847341.

[8] M. Ishiduki *et al.*, "Optimal device structure for pipe-shaped BiCS flash memory for ultra high density storage device with excellent performance and reliability," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2009, pp. 1–4, doi: 10.1109/IEDM.2009.5424261.

[9] E.-S. Choi and S.-K. Park, "Device considerations for high density and highly reliable 3D NAND flash cell in near future," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2012, pp. 9.4.1–9.4.4, doi: 10.1109/IEDM.2012.6479011.

[10] H.-J. Kang *et al.*, "Effect of traps on transient bit-line current behavior in word-line stacked nand flash memory with poly-Si body," in *Symp. VLSI Technol. (VLSI Technol.) Dig. Tech. Papers*, Jun. 2014, pp. 1–2, doi: 10.1109/VLSIT.2014.6894348.

[11] W.-J. Tsai *et al.*, "Polycrystalline-silicon channel trap induced transient read instability in a 3D NAND flash cell string," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2016, pp. 11.3.1–11.3.4, doi: 10.1109/IEDM.2016.7838395.

[12] Y. Kim and M. Kang, "Down-coupling phenomenon of floating channel in 3D NAND flash memory," *IEEE Electron Device Lett.*, vol. 37, no. 12, pp. 1566–1569, Dec. 2016, doi: 10.1109/LED.2016.2619903.

[13] *JESD22-A117 Document*, JEDEC, Arlington, VA, USA, Aug. 2018.

[14] C. Zambelli, L. Crippa, R. Micheloni, and P. Olivo, "Cross-temperature effects of program and read operations in 2D and 3D NAND flash memories," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, Oct. 2018, pp. 1–4, doi: 10.1109/IIRW.2018.8727102.

[15] N. Mielke *et al.*, "Bit error rate in NAND flash memories," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2008, pp. 9–19, doi: 10.1109/RELPHY.2008.4558857.

[16] T. Parnell, N. Papandreou, T. Mittelholzer, and H. Pozidis, "Modelling of the threshold voltage distributions of sub-20nm NAND flash memory," in *Proc. IEEE Glob. Commun. Conf.*, Dec. 2014, pp. 2351–2356, doi: 10.1109/GLOCOM.2014.7037159.

[17] K. Mizoguchi, T. Takahashi, S. Aritome, and K. Takeuchi, "Data-retention characteristics comparison of 2D and 3D TLC NAND flash memories," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2017, pp. 1–4, doi: 10.1109/IMW.2017.7939077.

[18] G. M. Paolucci, C. Miccoli, C. M. Compagnoni, L. Crespi, A. S. Spinelli, and A. L. Lacaita, "Investigation of cycling-induced VT instabilities in NAND flash cells via compact modeling," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2012, pp. 1–4, doi: 10.1109/IMW.2012.6213684.

[19] M. Jeong *et al.*, "Characterization of traps in 3-D stacked NAND flash memory devices with tube-type poly-Si channel structure," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2012, pp. 9.3.1–9.3.4, doi: 10.1109/IEDM.2012.6479010.

[20] B. Choi *et al.*, "Comprehensive evaluation of early retention (fast charge loss within a few seconds) characteristics in tube-type 3-D NAND flash memory," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2016, pp. 1–2, doi: 10.1109/VLSIT.2016.7573385.

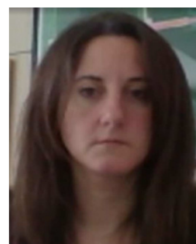
[21] K.-S. Shim *et al.*, "Inherent issues and challenges of program disturbance of 3D NAND flash cell," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2012, pp. 1–4, doi: 10.1109/IMW.2012.6213659.

[22] L. Zuolo *et al.*, "SSDExplorer: A virtual platform for performance/reliability-oriented fine-grained design space exploration of solid state drives," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1627–1638, Oct. 2015, doi: 10.1109/TCAD.2015.2422834.

[23] L. Zuolo, C. Zambelli, R. Micheloni, and P. Olivo, "Solid-state drives: Memory driven design methodologies for optimal performance," *Proc. IEEE*, vol. 105, no. 9, pp. 1589–1608, Sep. 2017, doi: 10.1109/JPROC.2017.2733621.

CRISTIAN ZAMBELLI (Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from the University of Ferrara, Ferrara, Italy, in 2008 and 2012, respectively, where he has been an Assistant Professor with the Department of Engineering since 2015. His current research interests include the characterization and modeling of nonvolatile memories and solid-state drives reliability.

RINO MICHELONI (Senior Member, IEEE) received the Ph.D. degree in engineering science from the University of Ferrara in 2018. He is a VP and a Fellow with Microsemi Corporation where he currently runs the Flash Signal Processing Labs. Early in his career, he led Flash design teams with STMicroelectronics, Hynix, and Infineon; during this time, he developed the industrys first MLC NOR device with embedded ECC technology and the industrys first MLC NAND with embedded BCH. He holds 278 patents worldwide (including 131 U.S. Patents) and he has published the following books, including *Solid-State-Drives (SSDs) Modeling* (Springer, 2017), *3D Flash Memories* (Springer, 2016), *Inside Solid-State Drives* (Springer, 2013), *Inside NAND Flash Memories* (Springer, 2010), *Error Correction Codes for Nonvolatile Memories* (Springer, 2008), *Memories in Wireless Systems* (Springer, 2008), and *VLSI-Design of Nonvolatile Memories* (Springer, 2005).



SALVATRICE SCOMMEGNA was born in Vimercate, Italy, in 1973. She received the High School Diploma in electronics in 1992. From 1993 to 2008, she worked in STMicroelecronics, Agrate, Italy, as a Product Engineer in the Flash Memory Division. From 2008 to 2015, she worked in Micron Technology as a NAND Product Engineer and a QA Engineer. In June 2015, she joined the PMC-Sierra/Microsemi that now is Microchip, as a Technical Staff Product Engineer within the Data Center Solution Group.

PIERO OLIVO received the Ph.D. degree in electronic engineering from the University of Bologna, Bologna, Italy, in 1987. He has been a Full Professor of electronics with the University of Ferrara, Ferrara, Italy, since 1994. His research interests include the physics, the reliability and the experimental characterization of innovative nonvolatile memorys and solid-state drive reliability.