This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2900677, IEEE Communications Letters

Cell-State-Distribution-Assisted Threshold Voltage Detector for NAND Flash Memory

Zhengqin Fan, Guofa Cai, Guojun Han, *Senior Member, IEEE,* Wenjie Liu, Yi Fang, *Member, IEEE*

***Abstract*—The process scaling as well as the multi-level cell (MLC) technology greatly increase the storage density of NAND flash memory. Unfortunately, the high-density flash memory suffers from severer noise such that its performance will be dramatically deteriorated. In this paper, we propose a cell-state-distribution-assisted threshold voltage detection (CSD-TVD) to optimize read reference voltages. Specifically, the proposed detection method exploits the variation of number of cells within each sub-window of the overlap region to detect the voltage shift and analyze its distribution. Afterwards, the mean of the voltage-shift distribution is viewed as the optimal voltage shift, which is used to boost the accuracy of read reference voltage. Based on the retention characteristics, we also conceive a novel low-latency (LL) CSD-TVD to reduce the detection range of the CSD-TVD. Experimental results demonstrate that the proposed CSD-TVD and LL-CSD-TVD achieve better error performance than the state-of-the-art retention-optimized-reading (ROR) and nonuniform detection methods. In addition, the LL-CSD-TVD significantly reduces the read latency with respect to the CSD-TVD and ROR.**

***Index Terms*—NAND flash memory, retention noise, read reference voltage, error performance, read latency.**

I. Introduction

NAND flash memory is widely used in mainstream electronic products due to its high performance, low power consumption, non-volatility and high storage capacity. With the gradual reduction in feature size and more bits stored in each cell, NAND flash memory cells become more vulnerable to various channel noises, including data retention, cell-to-cell interference (CCI), program/erase (P/E) cycles and read disturb [1]–[3]. The above noises significantly degrades the data-storage reliability. As a result, the data-storage lifetime is dependent on retention time and P/E cycles. Retention and CCI are recognized as the main sources of noise, which lead to significant performance deterioration of flash memory systems. To address this issue, powerful error-correcting codes, such as low-density parity-check (LDPC) codes with soft-decision decoding [4], were applied to improve the reliability of flash memory.

The flash controller reads the stored data by sensing the region of threshold voltage. To optimize the read reference voltage, a variety of methods were proposed in recent years [5]–[9]. More specifically, a nonuniform quantization strategy in the voltage overlap region was conceived to enhance the system performance [5]. In [6], [7], authors developed a scheme to optimize the read reference voltage by maximizing mutual information. Consider the retention time and P/E cycles, an entropy-function-aided optimization scheme was presented for the read reference voltage in [8]. Nonetheless, the optimization methods in [5]–[8] must estimate the threshold-voltage distribution in the presence of channel noise. Another related read-voltage design, referred to as *retention-optimized-reading (ROR)*, was introduced in [9]. The ROR optimizes the read reference voltage with the lowest raw BER. However, it does not substantially take the *a-priori* information of the threshold voltages (*i.e.*, the distribution characteristics of the voltage shifts on different wordlines) into consideration, and thus cannot achieve desirable BER performance.

With an aim to acquire high-reliability log-likelihood-ratio (LLR) for LDPC decoder, this paper proposes a threshold voltage detection (TVD) method based on cell-state distribution (CSD). In the proposed CSD-TVD, the overlap region of threshold voltage is divided into several sub-windows with a width of Δ, and the sub-window with the fewest number of cells is used to estimate the shift of threshold voltage. Furthermore, we conceive a latency-reduction approach to reduce the number of read operations in the CSD-TVD.

II. Background

A. Main Channel Noises

1) Erase and Program Operation: In NAND flash memory channel, each memory cell can be classified into erased state and programmed state. The threshold voltage of the erased state follows a normal distribution [10], which can be modeled as

$$p_e\left(x\right) = \frac{1}{\sigma_e\sqrt{2\pi}} e^{-\frac{(x-\mu_e)^2}{2\sigma_e^2}}, \tag{1}$$

where μ_e and σ_e are the mean and standard deviation, respectively.

To program each memory cell, an incremental step pulse program (ISPP) technique is usually applied [11]. The threshold voltage of the k-th programmed state follows a uniform distribution over $[V_p^k, V_p^k + \Delta V_{pp}]$, which can be written as

$$p_p^{(k)}\left(x\right) = \begin{cases} \frac{1}{\Delta V_{pp}} & \text{if } V_p^k \leq x \leq V_p^k + \Delta V_{pp}, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where V_p^k and ΔV_{pp} denotes the verify voltage of the k-th programmed state and the incremental programmed step voltage, respectively.

The authors are with the school of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China. (e-mail: fanzhengqin1992@163.com, caiguofa2006@126.com, gjhan@gdut.edu.cn, wenjie_liu@yeah.net, fangyi@gdut.edu.cn), and Y. Fang is also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. G. Han is the corresponding author.

This work was partially supported by the NSF of China (Nos. 61871136, 61471131, 61771149), the Project of the Education Department of Guangdong Province under Grants 2017KZDXM028 and 2017KTSCX060, and the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2018D02.

1089-7798 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2900677, IEEE Communications Letters

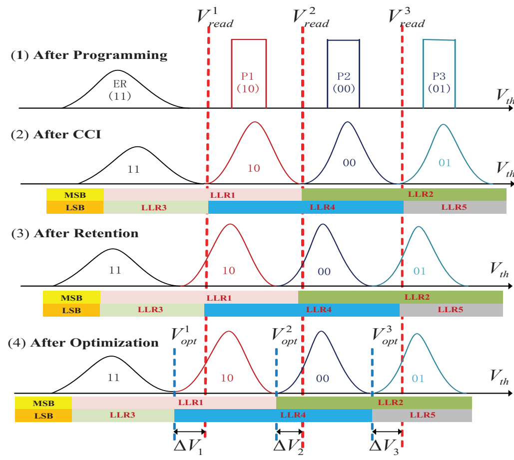

Fig. 1. Voltage distribution for 2-bit/cell NAND flash after programming, CCI, retention, and optimization.

2) *CCI:* Due to the parasitic capacitance-coupling effect [5], the threshold voltage shift F can be estimated as

$$F = \sum \Delta V_t^k \gamma^k, \tag{3}$$

where ΔV_t^k denotes the threshold-voltage shift of the k-th interfering cell during programming operation, and γ^k denotes the coupling ratio between the victim cell and interfering cell.

3) *Retention Noise:* Retention noise, arising from charge leakage over time, results in the threshold-voltage shift to the left. These noises follow a normal distribution $\mathcal{N}\left(\mu_t, \sigma_t^2\right)$ according to [12], where the mean and standard deviation are respectively given by

$$\mu_t = (x - x_0) \cdot \left[A_t\,(PE)^{\alpha_i} + B_t\,(PE)^{\alpha_o}\right] \cdot \log\left(1 + t\right), \tag{4}$$

$$\sigma_t = 0.3 \cdot |\mu_t|, \tag{5}$$

where x_0, A_t, B_t, α_i, and α_o are constant parameters, and t and PE are the retention time and P/E cycles, respectively.

B. LLR Calculation for LDPC Decoder

To improve the accuracy of the LLR values for LDPC decoder, a higher sensing accuracy is required in the overlap region. For an MLC flash memory channel, the threshold voltage can be divided into four regions. The four regions corresponds to four different states, i.e., the ER (11), P1 (10), P2 (00), and P3 (01) states. The probability density function (PDF) $p^k(v)$ of each state in the presence of CCI can be obtained in [5], where v denotes the read-back voltage. The initial LLR corresponding to the i-th bit can be calculated as

$$L(b_i) = \log \frac{\int_{R_l}^{R_r} \sum_{k \in S_i} p^k(v)\, dv}{\int_{R_l}^{R_r} \sum_k p^k(v)\, dv - \int_{R_l}^{R_r} \sum_{k \in S_i} p^k(v)\, dv}, \tag{6}$$

where R_l and R_r are two adjacent read reference voltages, and S_i denotes the set of states whose i-th bit is 1.

C. Effect of Read Reference Voltage

Fig. 1 shows the threshold-voltage distribution for 2-bit/cell NAND flash after programming, CCI, and retention, where V_{read}^1, V_{read}^2, and V_{read}^3 denote the three default read reference voltages. We assume that the LLR values are pre-calculated and stored in a look-up table. It can be seen that the threshold voltage of each programmed state moves from right to the left with retention noise. The principle of read operation is to sense the region of threshold voltage by exploiting read reference voltage.

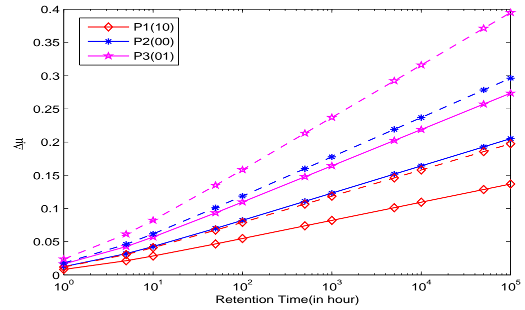

Fig. 2. Variation of the means over retention time at PE = 5K (solid lines) and PE = 10K (dotted lines).

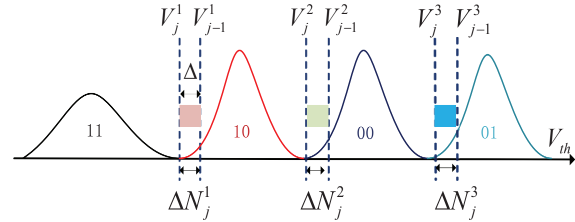

Fig. 3. Distribution of the threshold voltage in NAND flash memory.

In general, the read reference voltage can be optimized to mitigate the negative effect of noise. The optimized read reference voltage located in the overlap region between ER and P1 states (i.e., V_{opt}^1) and that between P2 and P3 states (i.e., V_{opt}^3) are used to read the least-significant-bit (LSB) page, while optimized read reference voltage located in the overlap region between P1 and P2 state (i.e., V_{opt}^2) is used to read the most-significant-bit (MSB) page.

III. Proposed Threshold-Voltage Detection Method

A. Effect of Retention Noise

To analyze the effect of the retention noise, we simulate the variation of the means for P1, P2 and P3 states in the presence of retention noise and show the results in Fig. 2. It is observed that the mean of the erased state remains the same. Moreover, a higher programmed state shifts faster than a lower programmed state, and thus $\Delta V_1 < \Delta V_2 < \Delta V_3$. The reason is that the threshold-voltage shift is directly proportional to the variation of the number of electrons stored in the cell's floating-gate.

B. CSD-TVD Method

Through the above analysis, a CSD-TVD is proposed to mitigate the effect of random noise and improve the system reliability. For convenience, Fig. 3 shows the distribution of the threshold voltage in NAND flash memory. In this figure, Δ denotes the width of the voltage sub-windows, ΔN_j^i (i=1,2,3, j=1,2,$\ldots$,n) denotes the number of cells whose threshold voltages are located in $[V_j^i, V_{j-1}^i)$, $V_j^i = V_{read}^i - j \times \Delta$, and V_{read}^i denotes the default read reference voltage of the i-th overlap region. In addition, let ΔV_i^k denote the voltage shift of i-th overlap region on the k-th wordline. Subsequently, the proposed CSD-TVD can be divided into two steps, i.e., calculating the voltage shifts and evaluating the distribution of voltage shifts.

1089-7798 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2900677, IEEE Communications Letters

1) Step 1: When sliding the voltage sub-window from the default read voltage V_{read}^{i} to the left in the i-th overlap region, the number of cells in the voltage sub-window will be first decreased and then increased. Thus, we stop sliding the voltage sub-window once the number of cells is increased. Then, the voltage sub-window including the fewest number of cells can be obtained, and thus yielding the voltage shift ΔV_i. Compared with the existing ROR [9], the CSD-TVD possesses similar detection range, which leads to almost the same read latency.

2) Step 2: Consider a 2-bit/cell flash channel and assume the parameters as follows: the number of page in a block is 256, cell-to-cell coupling strength factor s is 1.4, retention time is 10^5 hours, P/E cycles is 20K, and the number of wordlines M is 128. Fig. 4 compares the PDFs of ΔV_i^k in the three overlap regions before and after parameter estimation. In particular, the PDF after parameter estimation follows a normal distribution $\mathcal{N}(\mu_{\Delta_V}, \sigma_{\Delta_V}^2)$, where μ_{Δ_V} and $\sigma_{\Delta_V}^2$ denotes the mean and variance of all voltage shifts on M wordlines. It can be seen that the actual PDF of the voltage shifts well matches with the idea normal distribution. Based on the above observations, we can conclude that the voltage shifts on different wordlines approximately follow a normal distribution and will use this conclusion from this point onwards. Now, we propose a CSD-TVD, which can substantially exploit the normal-distribution property to improve the accuracy of read reference voltage. The detailed steps of the proposed CSD-TVD are outlined in Algorithm 1.

Algorithm 1 CSD-TVD Method

Input: The width of the voltage sub-windows Δ, the number of wordline in a block M, and the default read reference voltage V_{read}^{i};

Output: For $0 < i < 2^{n_b}$, the optimal voltage shift of the i-th overlap region ΔV_{opt}^{i};

1: **for** $k = 1 : M$ **do**
2:     **for** $i = 1 : 2^{n_b} - 1$ **do**
3:         Set $j \leftarrow 0$;
4:         **while** $\Delta N_{j+1}^{i} < \Delta N_{j}^{i}$ **do**
5:             $j \leftarrow j + 1$;
6:         **end while**
7:         $\Delta V_i^k = j \times \Delta$;
8:     **end for**
9: **end for**
10: Compute $\Delta V_{opt}^{i} = \overline{\Delta V_i^k}$;
11: End

C. Low-Latency CSD-TVD Method

As mentioned in [9], the flash read latency is proportional to the number of read operations, and the number of read operations of the i-th overlap region can be defined as $N_i = \frac{(V_{read}^{i} - V_{opt}^{i})}{\Delta} + 1$. According to the retention characteristics, the voltage of the $(k+1)$-th programmed state shifts faster than that of the k-th programmed state as the retention age increases. The default read reference voltage of the $(i+1)$-th overlap region can be optimized by utilizing the following equation: $V_{read}^{i+1} = V_{read}^{i+1} - \Delta V_i$. Based on above optimization method, we develop a low-latency CSD-TVD (LL-CSD-TVD), in which the number of read operations in the $(i+1)$-th $(i > 0)$ overlap region is reduced by $\frac{\Delta V_i}{\Delta}$, to reduce the read latency of the proposed CSD-TVD. For a n_b-bit/cell flash channel, the detailed steps of the proposed LL-CSD-TVD are outlined in Algorithm 2.

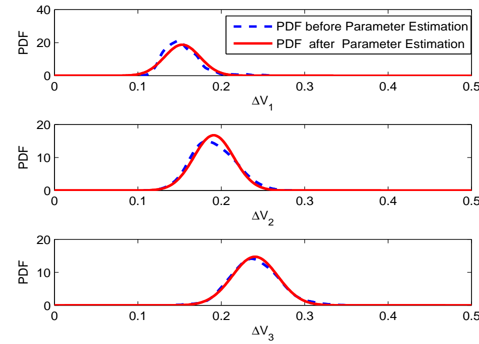

Fig. 4. PDFs of ΔV_i^k in the three overlap regions before and after parameter estimation.

Algorithm 2 LL-CSD-TVD Method

Input: The width of the voltage sub-windows Δ, the number of wordline in a block M, and the default read reference voltage V_{read}^{i};

Output: For $0 < i < 2^{n_b}$, the optimal voltage shift of the i-th dominating overlap region ΔV_{opt}^{i};

1: **for** $k = 1 : M$ **do**
2:     Set $j \leftarrow 0$;
3:     **for** $i = 1 : 2^{n_b} - 1$ **do**
4:         $j \leftarrow j$;
5:         **while** $\Delta N_{j+1}^{i} < \Delta N_{j}^{i}$ **do**
6:             $j \leftarrow j + 1$;
7:         **end while**
8:         $\Delta V_i^k = j \times \Delta$;
9:     **end for**
10: **end for**
11: Compute $\Delta V_{opt}^{i} = \overline{\Delta V_i^k}$;
12: End

To elaborate a little further, an example is provided to show the principle of the LL-CSD-TVD. In an MLC flash memory channel, there are three overlap regions in total. We can first obtain ΔV_1 by sliding the voltage sub-window from V_{read}^{1}. Second, we further get ΔV_2 by sliding the voltage sub-window from $(V_{read}^{2} - \Delta V_1)$. Finally, we can get ΔV_3 by sliding the voltage window from $(V_{read}^{3} - \Delta V_2)$. Therefore, by exploiting the above method, the number of read operations is reduced by $\frac{\Delta V_1 + \Delta V_2}{\Delta}$ compared with the original CSD-TVD.

D. Calculation of Optimal Voltage Shift

By subtracting the ΔV_{opt}^{i}, the noise of the flash channel can be compensated without estimating the threshold-voltage distribution. For a soft-decision reference voltage, ΔV_{opt}^{i} is subtracted from the read reference voltages located in the i-th overlap region. For instance, if the optimal voltage shift and original read reference voltage are given by $\Delta V_{opt}^{i} = \{0.101, 0.150, 0.186\}$ and $V_{read} = \{2.3, 2.4, 2.5, 2.9, 3.0, 3.1, 3.5, 3.6, 3.7\}$, respectively, then the optimized read reference voltage can be formulated as V_{opt}=$\{2.199, 2.299, 2.399, 2.75, 2.85, 2.95, 3.314, 3.414, 3.514\}$.

1089-7798 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2900677, IEEE Communications Letters

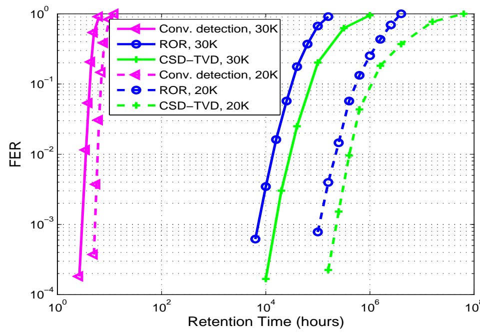

Fig. 5. FER performance of the proposed CSD-TVD, ROR, and conventional detection methods over an NAND flash memory channel.

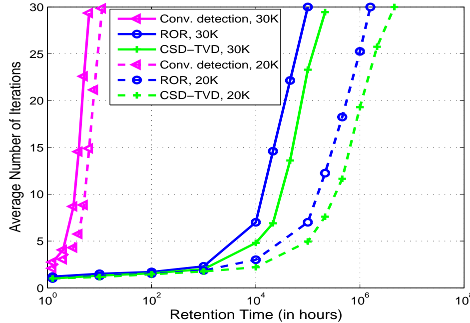

Fig. 6. The average number of iterations required by the proposed CSD-TVD, ROR, and conventional detection methods over an NAND flash memory channel.

IV. Simulation Results

In our simulations, the parameters of the MLC flash channel are set as $\mu_e = 1.4$, $\sigma_e = 0.35$, $V_p^k \in \{2.6, 3.2, 3.8\}$, $\Delta V_{pp} = 0.2$, $A_t = 0.000035$, $B_t = 0.000235$, $\alpha_i = 0.62$, $\alpha_o = 0.30$. The CCI coupling strength factor s is set to 1.4. A rate-0.9014 (4544, 4096) QC-LDPC code and the min-sum (MS) algorithm are used for coding and decoding, respectively. The maximum number of iterations I_{max} is set to 30. The conventional detection method in [5] and the ROR in [9] are considered as benchmarks.

Fig. 5 shows the frame-error-rate (FER) versus the retention time for the proposed CSD-TVD and the two existing methods over a flash memory channel.[1] Referring to this figure, the proposed CSD-TVD significantly outperforms the ROR and conventional detection methods. For instance, at $\mathrm{FER} = 10^{-3}$ and PE = 20K, the flash memory only has a retention age of about 700 hours by using the ROR, while it can achieve a retention age of about 1500 by using the proposed CSD-TVD. Furthermore, for a given retention time, the proposed CSD-TVD has a one-order-of-magnitude gain over the ROR.

Fig. 6 presents the average number of iterations for MS algorithm required by the above three detection methods. It can be observed that the average number of iterations increases as either the P/E cycles or retention time becomes larger. Moreover, the proposed CSD-TVD has better convergence performance (i.e., faster convergence speed) compared with the two existing methods, especially for the conventional detection method.

Fig. 7 compares the number of read operations required by the LL-CSD-TVD, CSD-TVD, and ROR over a flash memory channel. We can see that the CSD-TVD and ROR have almost

[1]The FER and average number of iterations of LL-CSD-TVD are not included in Figs. 5 and 6 because they are identical as those of CSD-TVD.

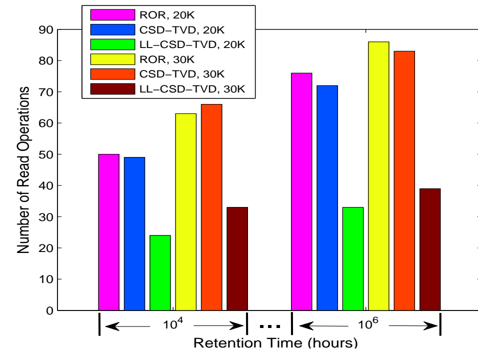

Fig. 7. The number of read operations on each wordline required by the LL-CSD-TVD, CSD-TVD, and ROR over an NAND flash memory channel.

the same number of read operations. Moreover, the number of the read operations required by the LL-CSD-TVD is only a half of those required by the other two methods.

V. Conclusions

In this paper, a dynamic CSD-TVD has been proposed to obtain high-reliability LLRs for LDPC decoder in NAND flash memory systems. The proposed CSD-TVD can optimize the voltage shift by analyzing the distribution characteristics of the voltage shifts on different wordlines. We have further proposed a LL-CSD-TVD, which can significantly reduce the number of read operations. With respect to the state-of-the-art ROR, both proposed detection methods can achieve one-order-of-magnitude performance improvement. Moreover, the LL-CSD-TVD is a-menable to much lower read latency.

References

[1] Y. Cai, S. Ghose, E. F. Haratsch, Y. Luo, and O. Mutlu, "Error characterization, mitigation, and recovery in flash memory based solid-state drives," *IEEE Proc.*, vol. 105, no. 9, pp. 1666–1704, Sep. 2017.
[2] Y. Cai, Y. Luo, S. Ghose, O. Mutlu, "Read disturb errors in MLC NAND flash memory: characterization, mitigation, and recovery," *DSN*, 2015.
[3] X. Wang, G. Dong, L. Pan, and R. Zhou, "Error correction codes and signal processing in flash memory," *Flash memories. InTech*, Croatia: InTech, 2011.
[4] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, Aug. 1963.
[5] G. Dong, N. Xie, and T. Zhang, "On the use of soft-decision error-correction codes in NAND flash memory," *IEEE Trans. Circuits Syst. I*, vol. 58, no. 2, pp. 429–439, Feb. 2011.
[6] J. Wang, T. Courtade, H. Shankar, and R. Wesel, "Soft information for ldpc decoding in flash: Mutual-information optimized quantization," in *Global Telecommun. Conf. (GLOBECOM)*, 2011, pp. 1–6.
[7] J. Wang, *et al.*, "Enhanced precision through multiple reads for LDPC decoding in flash memories," *IEEE J. Sel. Areas Commun*, vol. 32, no. 5, pp. 880–891, May. 2014.
[8] C. A. Aslam, Y. L. Guan, and K. Cai, "Read and write voltage signal optimization for multi-level-cell (MLC) NAND flash memory," *IEEE Trans. Commun*, vol. 64, no. 4, pp. 1613 –1623, Apr. 2016.
[9] Y. Cai, Y. Luo, E. F. Haratsch, K. Mai, and O. Mutlu, "Data retention in MLC NAND flash memory: characterization, optimization, and recovery," in *Proc. IEEE 21st Int. Symp. High Perform. Comput. Archit.*, San Francisco, CA, USA, Feb. 2015. pp. 551–563.
[10] K. Takeuchi, T. Tanaka, and H. Nakamura, "A double-level-V_{th} select gate array architecture for multilevel NAND flash memories," *IEEE J. Solid-State Circuits*, vol. 31, no. 4, pp. 602–609, Apr. 1996.
[11] K. D. Suh, *et al.*, "A 3.3V 32 Mb NAND flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.
[12] G. Dong, N. Xie, and T. Zhang, "Enabling NAND flash memory use soft-decision error correction codes at minimal read latency overhead," *IEEE Trans. Circuits Syst. I*, vol. 60, no. 9, pp. 2412–2421, Sep. 2013.

1089-7798 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.