The Journal of Korean Institute of Communications and Information Sciences '24-06 Vol.49 No.06
https://doi.org/10.7840/kics.2024.49.6.816

A Joint Sensing and Decoding for Improving the Hard-Decision Lifetime of NAND Flash Memories

Seokju Han◆, Inayat Ali*, Jeongseok Ha°

ABSTRACT

Soft-decision decoding schemes are utilized for high data reliability of NAND flash memories. However, it requires excessive latency and power consumption compared to hard-decision (HD) decoding schemes. This work proposes a novel joint sensing and decoding scheme to extend the HD decoding lifetime. When the HD decoding fails, the proposed scheme re-reads HD channel outputs and judiciously combines the two consecutive HD readings utilizing reliability information from the failed HD decoding. Since the random telegraph noises impairing HD readings are statistically independent, the combining provides a diversity gain. Numerical results show that the proposed scheme significantly improves the HD lifetime.

Key Words : Channel coding, NAND flash memory, random telegraph noise, diversity combining

Ⅰ. Introduction

Digital storage devices using NAND flash memory, such as solid-state drives have almost replaced magnetic disk-based drives in the consumer storage market. The NAND flash memory technology has recently received extensive attention for providing scalability in both storage capacity and data throughput. This was made possible by using multi-level-cell (MLC) technology, which uses a single memory cell for storing multiple bits[1]. The sequence of storing and retrieving data in and out of NAND flash memory can be modeled as data communication over a noisy channel which is impaired by various noise sources such as cell-to-cell interference, retention problem, and stress-induced leakage[2]. For instance, the channel deteriorates as the number of program/erase (P/E) cycles grows and eventually reaches a point called the end-of-life, at which an error-control scheme used in a storage device cannot correct errors induced by the channel. Although the NAND flash memory channel is time-varying and continuous in nature, the data are collected in discrete form by instantaneous sensing of the memory cells.

The chip manufacturers of NAND flash memory are reducing the storage cost per bit by storing more bits in a memory cell. Recently, NAND flash memories with four bits per cell, i.e. quad-level cell (QLC), have become commercially available in the storage market. However, by increasing the number of bits per cell, the data reliability degrades and the device lifetime reduces significantly. To overcome these issues, there have been extensive studies[3-6] on developing powerful error-correcting codes (ECCs) for storage devices using NAND flash memories. In particular, soft-decision (SD) error-control schemes, such as low-density parity-check (LDPC) codes with belief propagation (BP) decoder, provide considerable coding gain by conducting multiple memory sensing

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2021R1A2B5B01002204).
◆ First Author : Samsung Electronics, vlrm1005@gmail.com, 정회원
° Corresponding Author : Korea Advanced Institute of Science and Technology, jsha@kaist.edu, 종신회원
* LG electronics, inayatali2261@gmail.com
논문번호 : KICS2023-11-156, Received November 27, 2023; Revised February 21, 2024; Accepted February 27, 2024

and taking soft channel outputs from NAND flash memories[4-6]. While the SD error-control schemes provide high data reliability, the multiple memory sensing operations for the soft channel outputs result in excessively long latency and high power consumption, which is not appropriate for energy-constrained mobile storage applications.

Sensing and decoding schemes that utilize the SD error-control scheme as a post-processor are proposed[6-9] to reduce the overall latency and complexity required for processing the data. A progressive data sensing and decoding scheme to minimize the use of high-precision SD sensing is proposed[6]. To reduce the latency, the data is first decoded by using the hard-decision (HD) channel outputs generated with a single memory sensing. When the HD decoder fails to retrieve the data, the SD decoding is performed with SD channel outputs obtained by performing additional memory sensing. The flash controller gradually increases the precision level of the SD channel outputs until the data are successfully decoded. To reduce the unnecessary read latency, a method for adaptively selecting the optimal read-level granularity is proposed[8-9]. The progressive data sensing and decoding scheme can achieve high data reliability. However, the existing works focus on designing SD error-control schemes which still require high latency and complexity to decode the data.

In this paper, we propose a novel joint sensing and decoding scheme that utilizes the statistical independence of random telegraph noise (RTN) in time [10]. The threshold voltage of the flash memory cell is disturbed by the RTN, which incurs errors in the HD channel outputs from the flash memory. The RTN is a time-varying electronic noise caused by the capture and emission of electrons from the interfacial traps in the oxide layer of a cell. Thus, the RTN noises in two HD channel outputs are statistically independent.

In this work, we assume that the proposed scheme employs an LDPC code with a low-complexity decoder (i.e., a gradient descent bit-flipping (GDBF) decoder[11]). The proposed scheme first performs HD decoding with a single memory sensing. When the decoder fails, it acquires another set of HD channel outputs by re-sensing the memory cells. The proposed scheme carefully combines the two sets of HD channel outputs and performs the GDBF decoding with the combined set. That is, for each coded bit, the proposed scheme takes the one with higher reliability from the two sets, which provides so-called the selection diversity gain. To judge which one is more reliable, we smartly exploit the results of the failed GDBF decoding. Note that the GDBF decoder builds up the reliability of each coded bit using a metric called the inversion function. By taking advantage of the values of the inversion function at the end of decoding, we can estimate the unreliable bit positions in the HD channel outputs from the first round of sensing. The HD decoding is repeated with new combined HD channel outputs obtained by replacing the unreliable channel outputs in the previous decoding round with the new channel outputs.

Note that the outputs of the GDBF decoder are estimates of the coded bits based on the HD channel outputs and a metric called the inversion function. Using the reliability of the bits calculated at the decoder by the inversion function. It will be demonstrated that the proposed scheme greatly improves the error-rate performance, which in turn significantly extends the HD lifetime of the flash memory, i.e., the lifetime with hard-decision decoding. Note that since the proposed scheme is utilized before SD decoding, the proposed scheme can be collaboratively used with recent work that optimizes the SD granularity[8-9]. While the proposed scheme requires one additional sensing, it significantly reduces the chance of activating the SD decoding with at least three sensings. To verify the claims, we carry out error-rate performance evaluations, the number of SD activations, and the average number of sensings.

Ⅱ. Preliminaries

2.1 Flash Memory Channel

In flash memories, data is stored in cells consisting of floating-gate transistors, with the threshold voltage required to turn on the transistors determined by the number of electrons stored in the floating gates. The

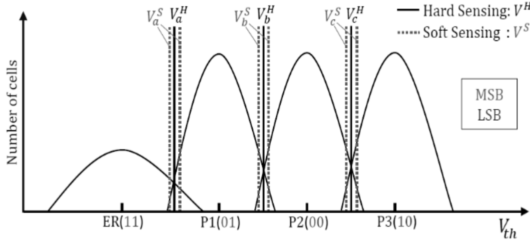

Fig. 1. Threshold voltage distribution of MLC NAND flash memory. Hard sensing V^H requires single data sensing per bit, and soft sensing V^s requires multiple (i.e., ≥ 2) data sensing per bit.

flash controller reads the data by checking the status of the transistors after applying a read reference voltage to the cells. During the programming operation of a cell, the flash controller gradually shifts the threshold voltage of a cell to a target write voltage level. An MLC NAND flash memory has four write voltage levels per cell, i.e., an erased state ER(11), and programmed states P1(01), P2(00), and P3(10) as shown in Fig. 1. In a flash memory channel, the threshold voltages in MLC are affected by different noise components such as the program-erase cycle (PE cycle), and retention time[1]. The initial threshold voltages in the erased cells $p_{\text{init}}^{\text{ER}}$ follow a Gaussian distribution, defined by $\mathcal{N}(V_{\min}, \sigma_e^2)$. The programmed states are generated with iterative incremental step pulse programming (ISPP). Furthermore, the initial threshold voltages in the programmed cells follow a uniform distribution after applying ISPP, defined as follows:

$$p_{\text{init}}^{PR}(x) = \begin{cases} \frac{1}{\Delta V_{pp}}, & \text{for } V_p \leq x \leq V_p + \Delta V_{pp} \\ 0, & \text{otherwise} \end{cases}$$

where $V_p \in \{V_1, V_2, V_3\}$ is the target programming voltage level and ΔV_{pp} is the ISPP step size.

The threshold voltage variations are mainly due to programming noise, retention noise, random telegraph noise (RTN), and cell-to-cell interference (CCI). The programming noise component affects the programmed cells with an additive white Gaussian noise with a distribution defined by $\mathcal{N}(0, \sigma_p^2)$. Retention noise is caused by the charge leakage through a floating gate. It is related to the PE cycles of the memory and data retention time. The retention noise[1] can be modeled as a noise component with a Gaussian distribution, $\mathcal{N}(\mu_r, \sigma_r^2)$. The mean μ_r and variance σ_r^2 of retention noise is given by:

$$\mu_r = (V_s - x_0)[A_t(N_P)^{\alpha_i} + B_t(N_P)^{\alpha_0}]\log(1+T),$$
$$\sigma_r = 0.4|\mu_r|$$

where T is the data retention time, N_p is the number of PE cycles, $V_s \in \{V_{min}, V_1, V_2, V_3\}$ and x_0, A_t, B_t, α_i, α_0 are constants described in [9].

The RTN is caused by the capture and emission of electrons from the interfacial traps in the oxide layer, which causes fluctuations in the threshold voltage of the flash memory cell. Due to the RTN, the threshold voltage changes even between consequent read operations. The RTN becomes more severe when N_p is increased, i.e., new traps are created in the oxide layer of the cell that makes RTN more susceptible. Similarly, the RTN is also modeled as a noise component with a Gaussian distribution defined by $\mathcal{N}(0, \sigma_{\text{RTN}}^2)$[1] where σ_{RTN} vary with PE cycles and can be calculated as $\sigma_{\text{RTN}} = 0.00025(N_p)^{0.62}$.

On the other hand, when a flash cell is programmed, the adjacent memory cells are affected by the parasitic capacitive coupling, which introduces CCI in the memory cells. However, the CCI can be removed from the programmed cells by using different pre-coding techniques[10]. For erased cells, the threshold voltage distribution with CCI is modeled as:

$$\tilde{V}_{\text{min,CCI}}^{\text{even}} = V_{\min} + \Delta V_{\text{avg}}\left(2\mu_{\gamma_x} + \mu_{\gamma_y} + 2\mu_{\gamma_{xy}}\right),$$
$$\tilde{V}_{\text{min,CCI}}^{\text{odd}} = V_{\min} + \Delta V_{\text{avg}}\left(\mu_{\gamma_y} + 2\mu_{\gamma_{xy}}\right)$$

where $\Delta V_{\text{avg}} = (V_{\min} + V_3)/2 - V_{\min}$, and $\mu_{\gamma x}$, $\mu_{\gamma y}$, $\mu_{\gamma xy}$ are capacitive coupling ratios which depend on the physical architecture of the cells in the flash memory.

2.2 Flash Memory Reading/Sensing

The flash memory controller reads data based on

a predetermined reference read voltage. As explained in Sec. 1.1, the distribution of the threshold voltages varies from the target voltage levels depending on the retention time and the PE cycle. The degradation of threshold voltage leads to a higher raw bit error rate (BER). In [6], an SD error control scheme is used as a post-processor to enhance the error-correcting performance as shown in Fig. 2. Firstly, the memory controller generates HD channel outputs based on a single data sensing per bit with the reference read voltage $V^H \in \{V_a^H, V_b^H, V_c^H\}$ shown in Fig. 2. Then, HD decoding is performed with HD channel outputs. If the decoder fails, the controller progressively reads additional reference voltage levels per cell (e.g., $V^S \in \{V_a^S, V_b^S, V_c^S\}$) to generate soft reliability information in the form of log-likelihood ratios (LLRs) by combining the sensed channel outputs. These LLRs are used for decoding the data using any soft-input/soft-output (SISO) decoder (such as sum-product, or min-sum decoder).

By using an SD decoder with soft reliability information, the decoding performance improves as compared to the single HD sensing and decoding. However, using an SD decoder has some drawbacks. Firstly, SD sensing requires generating additional SD information, which degrades the on-chip sensing and data transfer latency. Second, the SD decoder requires a large computational power due to its high decoding complexity. Unlike the HD decoder, the reliability messages used in the SD decoder are in the form of LLRs for multiple voltage levels in MLC. Thus, for practical applications, it is important to extend the HD decoding lifetime in the memory for minimizing the use of the SD decoder having higher latency and complexity.

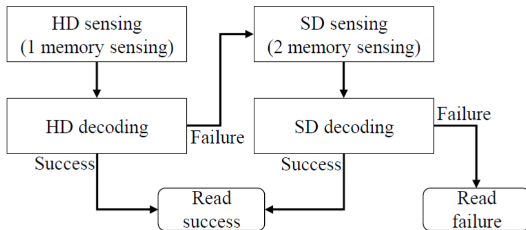

Fig. 2. Block diagram of conventional flash memory sensing and decoding scheme.

Ⅲ. Proposed Joint HD Sensing and Decoding

In this section, a novel joint decoding and memory sensing scheme is proposed, which utilizes the diversity provided by RTN for additional rounds of HD decoding. The random variations in the threshold voltage caused by the RTN provide diversity between the consecutively read data. The different reliability values between multiple signal inputs make it possible to improve the decoding performance through diversity combining, which is well-known in wireless communications[13]. There exist various diversity combining techniques, such as selection diversity, maximal-ratio diversity, equal gain diversity, etc. Selection diversity is a simple diversity combining technique that selects the input signal with the highest instantaneous reliability information. Since the HD sensed data is in discrete binary form, it is hard to select reliable and unreliable bits from the data. One possible way is to replace all previously sensed data by declaring them as unreliable when the decoder fails, hence new data is regenerated by re-sensing the memory for the next round of decoding. However, useful information is lost when discarding the previously sensed data that was not erroneous. In this paper, we propose an algorithm that utilizes the decoding result of the previous HD sensed data for improving the sensing and data transfer latency.

3.1 Measuring Data Reliability Using Inversion Function

In the SD decoding of the LDPC codes (such as the BP decoding), a hard decision over the reliability of the LLR values of each bit is used for the syndrome checking. Similarly, in the GDBF algorithm, a reliability measure called the *inversion function* is calculated for checking the amount of confidence over the bit decision value[11]. The inversion function Δ_k amounts to the reliability of the k-th variable node (VN) and is calculated as follows:

$$\Delta_k = y_k x_k + \sum_{i \in \mathcal{M}(k)} w_i$$

where y_k and $x_k \in \{-1, +1\}$ is the channel output for the k-th bit, and bi-polar bit decision value after decoding of the k-th VN, respectively. $\mathcal{M}(k)$ is a set of check nodes (CNs) connected to the k-th VN, and w_i is a reliability indicator value given by:

$$w_i = \begin{cases} +1, & \text{if } i-\text{th CN is satisfied,} \\ -1, & \text{otherwise} \end{cases}$$

The magnitude of Δ_k indicates the measure of confidence over the bit-decision value after decoding and is used for sorting the erroneous and reliable bits. For example, if Δ_k is small, this means that the majority of the CNs connected to the k-th VN are unsatisfied, consequently, the measure of confidence over the bit-decision value of the k-th VN after decoding is low. To show the significance of Δ_k in the sorting of the erroneous and reliable bits, the empirical behavior of the decoder is shown and compared with Δ_k in Fig. 3. The raw BER of the channel outputs is compared with Δ_k for (9216, 8192) quasi-cyclic LDPC (QC-LDPC) code at different N_p and T = 100 hours in Fig. 3a. We set the maximum number of iterations $I_{\max}$ as 50, and the decoder is run for $I_{\max}$ iterations to record these empirical results. We can observe that the raw BER is small for a large value of Δ_k (i.e., $\Delta_k = 5$), whereas, for the smaller value of Δ_k, the raw BER is notably high. Using the same simulation setup, the empirical distribution of the VNs having Δ_k value is depicted in Fig. 3b. We can observe that most of the bits have large Δ_k (i.e., the empirical distribution is high for $\Delta_k = 5$ in Fig. 3b), which means that we only need to sort a small number of VNs having unreliable channel outputs by Δ_k

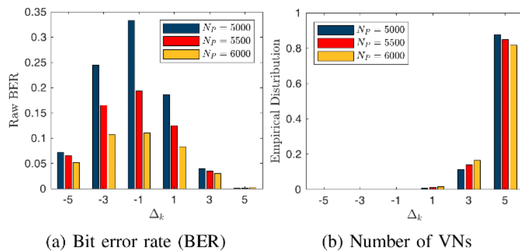

Fig. 3. Inversion function analysis with (9216, 8912) LDPC code when T = 100 hours.

3.2 Reliability-based Sensing and Combining

In this subsection, we propose an efficient joint sensing and decoding scheme in which HD data is re-sensed using the prior reliability information of the bits calculated from Δ_k. The proposed scheme is illustrated in Fig 4. Firstly, HD decoding is performed based on the HD sensed data, i.e., $\mathbf{r}^h = \{r_0^h, r_1^h, \ldots, r_{N-1}^h\}$, and if the decoder fails, it outputs the reliability values of the decoding result, i.e., $\mathbf{\Delta} = \{\Delta_0, \Delta_1, \ldots, \Delta_{N-1}\}$. Note that the bits with large Δ_k can make a sufficiently reliable decision based on the previous sensed data as observed in Section 3.1. Due to the time-varying nature of the RTN, new HD data for the unreliable bits are generated by re-sensing the memory cells. Therefore, only the unreliable data is re-generated in the next round of memory sensing, and we retain the reliable information from the previous sensing round.

We use Δ_k to define the set of unreliable bits called *combining set* (denoted by $\mathcal{A}$) that are updated with the new channel outputs. We define $\mathcal{A}$ as follows:

$$\mathcal{A} = \{k | \Delta_k \le \tau, 0 \le k \le N-1\} \quad (1)$$

where τ is a pre-determined reliability threshold value. All bits with Δ_k smaller than τ are included in $\mathcal{A}$. After determining the combining set, the new data $\mathbf{r}^h$ is sensed from the memory. Using $\mathbf{r}^h$ and $\mathcal{A}$, the HD channel output $\tilde{\mathbf{r}}^h$ for the current round of HD decoding is computed as follows:

$$\tilde{r}_i^h = \begin{cases} r_i^h, & \text{if } i \in \mathcal{A}, \\ r_i^{h-1}, & \text{otherwise.} \end{cases} \quad (2)$$

Algorithm 1: Proposed Joint HD Sensing and Decoding

Input: h_{max}
Initialization: Set $\mathbf{\Delta} \leftarrow \varnothing$, $\mathcal{A} \leftarrow \{i | 0 \le i \le N-1\}$

1 **for** $h \leftarrow 1$ *to* h_{max} **do**
2 HD Memory sensing $\rightarrow \mathbf{r}^h = \{r_0^h, r_1^h, \ldots, r_{N-1}^h\}$
3 **if** $h > 1$ **then**
4 Find combining set $\mathcal{A}$ using (1)
5 Compute $\tilde{\mathbf{r}}^h$ using (2)
6 Perform GDBF decoding with $\tilde{\mathbf{r}}^h$ and store $\{\hat{\mathbf{c}}, \mathbf{\Delta}\}$
7 **if** $\hat{\mathbf{c}}$ is a codeword (Successful Decoding) **then**
8 Return $\hat{\mathbf{c}}$

Fig. 4. Proposed joint HD sensing and decoding algorithm.

The HD decoder operates with re-generated HD channel outputs. Although the proposed algorithm describes a generalized decoder with h_{max} being any arbitrary number, we mainly focus on the case when $h_{max}=2$ in this paper. Since the latency and power consumption for data sensing is linearly proportional to the number of sensing, a smaller h_{max} is more desirable for practical considerations. A single round of SD sensing requires two times larger latency and power compared to a single round of the proposed scheme (i.e., when $h_{max}=2$). For some applications which require a lower target BER, the flash controller can proceed to SD sensing and decoding when the HD decoder fails. Furthermore, it will be shown in Section 4 that the proposed scheme can improve not only the HD decoding performance but also the latency and power requirements compared to the existing schemes.

Ⅳ. Numerical Results

In this section, the performances of the proposed and existing error-control schemes for NAND flash memory are compared in terms of word-error rate (WER) and the HD lifetime of the decoder. For evaluation, we use a (9216, 8192) QC-LDPC code having a code rate of 0.89. The code has VN and CN degrees of 4 and 36, respectively, and we use the GDBF algorithm[11] for decoding. To show the efficacy of the proposed decoder in terms of WER performance and complexity, we set $h_{max}=2$. The threshold value τ for finding $\mathcal{A}$ is fixed to 3. To determine each noise parameter used in numerical simulations, the reference read voltage for HD sensing is set as in [14]. We set the channel parameters according to [1] as listed in Table 1.

Table 1. NAND Flash Memory Channel Parameters

Parameter	Value	Parameter	Value
V_{min}	1.40	$\{V_1,\ V_2,\ V_3\}$	{2.6, 3.2, 3.9}
σ_e	0.35	σ_p	0.05
$\mu_{\gamma y}$	0.08	$\mu_{\gamma x}$	0.005
$\mu_{\gamma xy}$	4×10^{-3}	x_0	1.30
A_t	5.5×10^{-5}	B_t	2.4×10^{-4}
α_i	0.62	α_0	0.32

In Fig. 5, the WER performances of the proposed and existing algorithms are shown for different PE cycles, while the retention time is set to $T=100$ hours. For comparison, we plot the performances of the single HD sensing and decoding, HD re-sensing with full replacement with new sensing data, and SD sensing and decoding. Using the diversity of the RTN, we can observe that the WER of the flash memory significantly improves for both fully replaced and joint sensing and decoding algorithms, as compared with the single hard sensing and decoding scheme. Especially, the proposed algorithm shows more than 2 orders of coding gain for both the least significant bit (LSB) and the most significant bit (MSB) compared to the single hard sensing and decoding scheme.

To show the generality of the proposed algorithm, we also evaluate the performance for a TLC NAND flash memory model in [8]. In Fig. 6, the WER performances of the proposed and existing algorithms are depicted. For performance evaluation, the retention time is set to $T=40$ hours. It is observed that the proposed joint sensing and decoding algorithm improves WER performance of single HD decoding and fully replaced and shows more than one order of coding gain for all the LSB, central significant bit (CSB) and MSB.

In this paper, the efficacy of our proposed algorithm is shown in terms of the increased HD

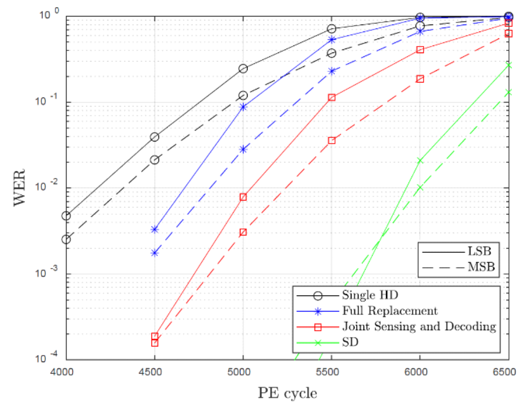

Fig. 5. WER results of different sensing and decoding algorithms used in an MLC NAND flash memory when $T=100$ hours.

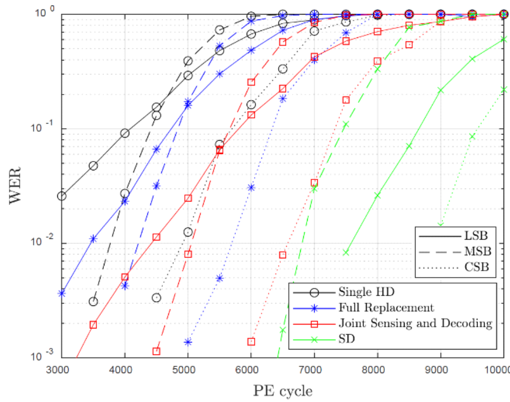

Fig. 6. WER results of different sensing and decoding algorithms used in a TLC NAND flash memory when T = 40 hours.

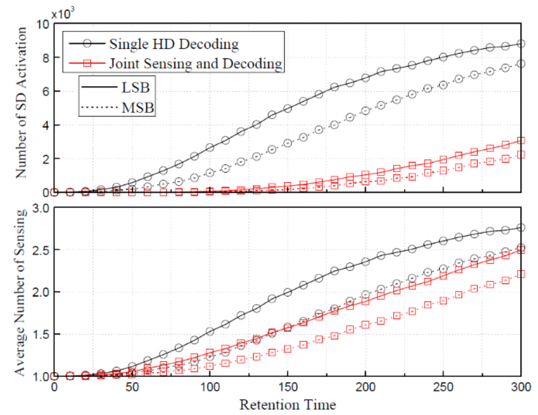

Fig. 7. Number of SD activation and average number of sensing for an MLC NAND flash memory when N_p = 5000.

lifetime of the data in memory cells, hence in numerical simulations, we did not perform SD decoding after the algorithm fails. In practice, the memory devices perform SD sensing and decoding when the flash memory cell degrades over time. Figure 7 shows the number of SD decoding activations and the average number of sensing compared to retention time when N_p = 5000. We define the average number of memory sensing for single HD decoding and the proposed algorithm as

$$U_s^{\mathrm{avg}} = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} U_k,$$

$$U_k = \begin{cases} 1 + 2P_1, & \text{for single HD decoding,} \\ 1 + P_1 + 2P_2, & \text{for joint sensing and decoding.} \end{cases}$$

where $\mathcal{N}$ is the number of samples from numerical simulation, and P_1 = 1 if the first HD decoding fails and it is 0 otherwise. Similarly, P_2 = 1 if the proposed algorithm fails to decode and it is 0 otherwise.

The number of SD activations and the average number of sensing are evaluated for $\mathcal{N}= 10^4$ samples of data read from the memory. The sensing and decoding complexity can be reduced and managed by reducing the number of times the algorithm activates SD sensing and decoding. By using the proposed scheme, we can observe that the number of SD decoding activations is significantly reduced compared to the single HD sensing and decoding scheme. Furthermore, the read latency of the flash memory is highly proportional to the number of times the memory is sensed. From Fig. 5, we can observe that the average number of memory sensing is greatly reduced by using the proposed scheme. Thus, the proposed scheme significantly improves the storage lifetime, with only a few additional HD sensing and low-complexity HD decoding.

Ⅴ. Conclusion

In this paper, we propose a decoder-aided flash memory read algorithm for NAND flash memories that utilizes the diversity of RTN. When HD decoding fails, the proposed scheme performs additional HD decoding by combining the consecutive HD sensed data using the reliability values of bits from the previous decoding round. By performing HD decoding in multiple rounds, the proposed algorithm significantly improves the data read latency and power efficiency of the NAND flash memories. The efficacy of the proposed scheme is shown by utilizing an LDPC code for numerical simulations, however, it can simply generalize to any type of ECC that output bit-wise reliability after HD decoding.

References

[1] 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, 2016. (https://doi.org/10.1109/TCOMM.2016.2533498)

[2] Y. Cai, S. Ghose, E. F. Haratsch, Y. Luo, and O. Mutlu, "Error characterization, mitigation, and recovery in flash-memory-based solid-state drives," in *Proc. IEEE*, vol. 105, no. 9, pp. 1666-1704, 2017. (https://doi.org/10.1109/JPROC.2017.2713127)

[3] J. Oh and J. Ha, "Error-correction codes for NAND flash memory," *Inform. & Commun. Mag.*, vol. 28, no. 9, pp. 58-68, 2011.

[4] J. Wang, T. Courtade, H. Shankar, and R. D. Wesel, "Soft information for LDPC decoding in flash: Mutual-information optimized quantization," in *IEEE GLOBECOM*, pp. 1-6, 2011. (https://doi.org/10.1109/GLOCOM.2011.6134417)

[5] J. Oh, S. Han, and J. Ha, "An improved symbol-flipping algorithm for nonbinary LDPC codes and its application to NAND flash memory," *IEEE Trans. Magnetics*, vol. 55, no. 9, pp. 1-13, 2019. (https://doi.org/10.1109/TMAG. 2019.2918985)

[6] G. Dong, N. Xie, and T. Zhang, "Enabling NAND flash memory use soft-decision error correction codes at minimal read latency overhead," *IEEE Trans. Cir. and Syst. I: Regular papers*, vol. 60, no. 9, pp. 2412-2421, 2013. (https://doi.org/10.1109/TCSI. 2013.2244361)

[7] M. Zhang, F. Wu, X. Chen, Y. Du, W. Liu, Y. Zhao, J. Wan, and C. Xie, "RBER aware multi-sensing for improving read performance of 3D MLC NAND flash memory," *IEEE Access*, vol. 6, pp. 61934-61947, 2018. (https://doi.org/10.1109/ACCESS.2018.2873081)

[8] B. Bao, Q. Li, W. Guan, Q. Wang, L. Liang, and X. Qiu, "Adaptive granularity progressive LDPC decoding for NAND flash memory," in *IEEE Trans. Computer-Aided Design of Integrated Cir. and Syst.*, Early Access. (https://doi.org/10.1109/TCAD.2023.3332829)

[9] Y. Du, S. Huang, Y. Zhou, and Q. Li, "Towards LDPC read performance of 3D Flash memories with layer-induced error characteristics," *ACM Trans. Design Automat. Electr. Syst.*, vol. 28, no. 3, pp. 1-25, 2023. (https://doi.org/10.1145/3585075)

[10] K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, and C. Hu, "Random telegraph noise in flash memories-model and technology scaling," in *2007 IEEE Int. Electron Devices Meeting*, pp. 169-172, 2007. (https://doi.org/10.1109/IEDM.2007.4418893)

[11] T. Wadayama, K. Nakamura, M. Yagita, Y. Funahashi, S. Usami, and I. Takumi, "Gradient descent bit flipping algorithms for decoding LDPC codes," in *2008 Int. Symp. Inform. Theory and Its Appl. IEEE*, pp. 1-6, 2008. (https://doi.org/10.1109/ISITA.2008.4895387)

[12] G. Dong, S. Li, and T. Zhang, "Using data post compensation and pre-distortion to tolerate cell-to-cell interference in MLC NAND flash memory," *IEEE Trans. Cir. and Syst. I: Regular Papers*, vol. 57, no. 10, pp. 2718-2728, 2010. (https://doi.org/10.1109/TCSI.2010.2046966)

[13] D. Brennan, "Linear diversity combining techniques," in *Proc. IEEE*, vol. 91, no. 2, pp. 331-356, 2003. (https://doi.org/10.1109/JPROC.2002.808163)

[14] G. Dong, N. Xie, and T. Zhang, "On the use of soft-decision error-correction codes in NAND flash memory," *IEEE Trans. Circuits and Syst. I: Regular Papers*, vol. 58, no. 2, pp. 429-439, 2011. (https://doi.org/10.1109/TCSI.2010.2071990)

Seokju Han

Feb. 2017 : B.S. in electrical engineering, KAIST
Feb. 2019 : M.S. in electrical engineering, KAIST
Feb. 2024 : Ph. D. in electrical engineering, KAIST
2024~current: Samsung Electronics
<Research Interest> communication theory, coding theory, error correction codes
[ORCID:0000-0001-9453-7033]

Jeongseok Ha

Feb. 1992 : B.S. in electronics, Kyungpook National University
Feb. 1994 : M.S. in electronic and electrical engineering, Pohang University of Science and Technology
2004~current : Professor, School of electrical engineering, KAIST
<Research Interest> wireless communication, coding theory, error correction codes
[ORCID:0000-0003-1262-151X]

Inayat Ali

Aug. 2009 : B.S. in electrical engineering, KAIST
Oct. 2011 : M.S. in telecommunication engineering, Hamdard University
Feb. 2019 : Ph. D. in electrical and computer engineering, Sungkyunkwan University
2024~current : LG Electronics
<Research Interest> coding theory, information theory, error correcting codes, digital communication, blockchain technology
[ORCID:0000-0002-0566-6405]