Read-Voltage Optimization for Finite Code Length in MLC NAND Flash Memory

Kang Wei[1], Jun Li[1], Lingjun Kong[2], Feng Shu[1] and Yonghui Li[3]
[1]Nanjing University of Science and Technology, Nanjing, China
[2]Nanjing University of Posts and Telecommunications, Nanjing, China
[3]University of Sydney, Sydney, Australia
{kang.wei, jun.li, shufeng}@njust.edu.cn, ljkong@njupt.edu.cn, yonghui.li@sydney.edu.au

***Abstract*—In this paper, we propose an effective read-voltage optimization method for multi-level-cell (MLC) NAND flash memory to improve the performance of error correcting codes (ECCs) with finite blocklength. Specifically, we first obtain the maximal channel coding rate achievable at a given blocklength and error probability of quantized channel. Based on this finite-blocklength channel-coding rate (FCR), we convert the optimization problem into minimizing the error probability instead of the channel coding rate. Then, we develop a cross iterative search (CIS) method and the genetic algorithm to solve this optimization problem. In our simulations, for a well-designed LDPC code, our read-voltage optimization method improves program-and-erase (PE) endurance up to about 900 and 600 cycles against the maximizing the mutual information (MMI) and entropy-based optimization methods, respectively, at a frame-error-rate (FER) of 2×10^{-4}.**

I. Introduction

NAND flash memory is widely used over the past decade, which has low power consumption and large storing capacity. The original NAND flash memory cell can only store one bit with two levels, which called single-level-cell (SLC). Using the multi-level-cell (MLC) or triple-level cell (TLC) technique [1], [2], It can store multiple bits over a single memory cell. Due to the increasing number of levels in one memory cell, serious scaling challenges loom up in NAND flash memory, which has a large effect on the reliability. These challenges are originated from flash device characteristics and can be seen as several noise models, such as programming noise (PN), cell-to-cell interference (CCI), random telegraph noise (RTN) and retention noise (RN) [3]. Consequently, the large number of program-and-erase (PE) cycles which the flash experiences limits the operational lifetime of flash memory.

To overcome the large degradation of reliability and error performance, error correcting codes (ECCs) for flash memory are becoming more important, such as Bose-Chaudhri-Hocquenghem (BCH) [4] and low-density parity-check (LDPC) [5], [6] codes. The belief-propagation (BP) algorithm is one of the probability-based iterative decoding algorithms with excellent performance [7]–[10].

It is well known that LDPC codes have typically been decoded with soft information. Using an one-bit A/D converter and a higher-precision A/D converter are two ways to obtain the soft information for a codeword in flash memory Mathematical formulations are derived to approximately model the threshold-voltage distribution of memory cells in the presence of CCI, based on which the calculation of log-likelihood-ratios (LLRs) is formulated [3]. A non-uniform memory sensing strategy is introduced to reduce the memory sensing precision and sensing latency with maintaining good error-correction performance [5]. [11] provides an enhanced precision scheme for multiple reads of the same flash memory cell, which selects the word-line voltages by maximizing the mutual information (MMI) between the input and output of the equivalent discrete read channel. A voltage entropy-based quantization scheme for reading the MLC NAND flash memory cells is introduced, where the voltage erasure regions are controlled to produce the lowest frame error probability [12]. In practice, there is an obvious gap between the actual channel coding rate (CCR) and capacity of the flash memory for the finite code length. Above mentioned works select read voltages with MMI or entropy-based scheme without considering the finite blocklength. However, for the discrete memoryless channels (DMCs), there are no exact formulas for the maximal rate as a function of blocklength and error probability.

In this paper, we are interested in the read-voltage optimization method for the finite length code in MLC NAND flash memory. We first consider the maximal CCR achievable of the quantized flash memory. Then, we formulate the read-voltage optimization problem by minimizing the error probability basing on the maximal CCR achievable, which considers the finite blocklength. Next, we investigate a nonlinear optimization based genetic algorithm and a cross iterative searching (CIS) method to solve this problem. For a well-designed LDPC code, our read-voltage method increases the useful lifetime of the flash memory obviously. In our simulations, our proposed method can improve the PE cycles by up to 900 and 600 cycles against to the MMI and entropy-based quantization, respectively.

II. System model

For MLC flash memories, each cell can store 2 bits and they are corresponding to four data symbols ‘11’, ‘10’, ‘00’, ‘01’. In general, the final threshold voltage distribution functions are calculated by the convolution integral of initial voltage distribution functions with various noise functions and denoted

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:27:52 UTC from IEEE Xplore. Restrictions apply.

as p_{s_0}, p_{s_1}, p_{s_2} and p_{s_3} , given by

$$p_{s_0}(v) = \frac{1}{\sigma_{s_0}\sqrt{2\pi}} e^{-\frac{(v-(\tilde{V}_{\min}-\mu_{r_{s_0}}))^2}{2\sigma_{s_0}^2}},$$
$$p_{s_i}(v) = \frac{1}{\sqrt{\pi}\Delta V_{pp}} \int_{\frac{V_i+\Delta V_{pp}-v-\mu_{r_{s_i}}}{\sqrt{2}\sigma_{s_i}}}^{\frac{V_i-v-\mu_{r_{s_i}}}{\sqrt{2}\sigma_{s_i}}} e^{-x^2} dx, \quad (1)$$

where $p_{s_i} \in \{p_{s_1}, p_{s_2}, p_{s_3}\}$, $V_i \in \{V_1, V_2, V_{\max}\}$, $\sigma_{s_i} \in \{\sigma_{s_1}, \sigma_{s_2}, \sigma_{s_3}\}$, $\widetilde{V}_{\min}$ is the shifted mean of the threshold voltage distribution of erased cells caused by CCI, ΔV_{pp} is a programming voltage step size using an iterative incremental step pulse programmed (ISPP) technique and v is the given threshold voltage. The final standard deviation σ_{s_i} are given by

$$\sigma_{s_0} = \sqrt{\sigma_e^2 + \sigma_{rtn}^2 + \sigma_{r_0}^2},$$
$$\sigma_{s_1} = \sigma_{s_2} = \sigma_{s_3} = \sqrt{\sigma_{pn}^2 + \sigma_{rtn}^2 + \sigma_{r_s}^2}, \quad (2)$$

where $\mu_{r_s} \in \{\mu_{r_{s_0}}, \mu_{r_{s_1}}, \mu_{r_{s_2}}, \mu_{r_{s_3}}\}$ is the mean of RN, σ_{pn}^2, σ_{rtn}^2 and $\sigma_{r_s}^2 \in \{\sigma_{r_{s_0}}^2, \sigma_{r_{s_1}}^2, \sigma_{r_{s_2}}^2, \sigma_{r_{s_3}}^2\}$ are the noise variance parameters of PN, RTN and RN, respectively. The mean and variance parameters σ_{r_s} are given by

$$\mu_{r_s} = (V_i - x_0) \cdot [X_t \cdot (PE)^{\alpha_0} + Y_t \cdot (PE)^{\alpha_1}] \cdot \log(1+T),$$
$$\sigma_{r_s} = 0.4|\mu_{r_s}|, \quad (3)$$

where $V_i \in \{V_{\min}, V_1, V_2, V_{\max}\}$, x_0, α_0, α_1, X_t and Y_t are constants, T is the data retention time and PE is the number of PE cycles.

In our study, the parameters of MLC NAND flash memory are set as: $\widetilde{V}_{\min} = 1.6429$, $\sigma_e = 0.35$, $V_{\max} = 3.93$, $\sigma_{pn} = 0.05$, $\sigma_{rtn} = 0.00025(PE)^{0.62}$, $x_0 = 1.4$, $X_t = 0.000055$, $Y_t = 0.000235$, $\alpha_0 = 0.62$, $\alpha_1 = 0.32$. These values and the intermediate write-voltage levels (V_1 and V_2) are assigned according to [12]. Fig.1 shows the voltage distributions of four states and a 6-level read scheme for the MLC NAND flash memory.

To obtain the soft information, reading the same sense-amp comparator multiple times with different word-line voltages is a useful approach. With J-level read, the threshold voltages of memory cells are quantized into J+1 regions. In this scenario, a continuous output of voltages for the flash memory channel can be seen as an equivalent discrete output. We use $s_i \in \{s_0, s_1, s_2, s_3\}$ and $r_j \in \{r_0, r_1, \ldots, r_J\}$ to denote the four input states and J+1 output regions, respectively. p_{r_j,s_i} represents the probability that the memory cell was programmed to s_i and appeared into region r_j, which can also be seen as a conditional probability mass function from s_i to r_j in the equivalent discrete memoryless channel.

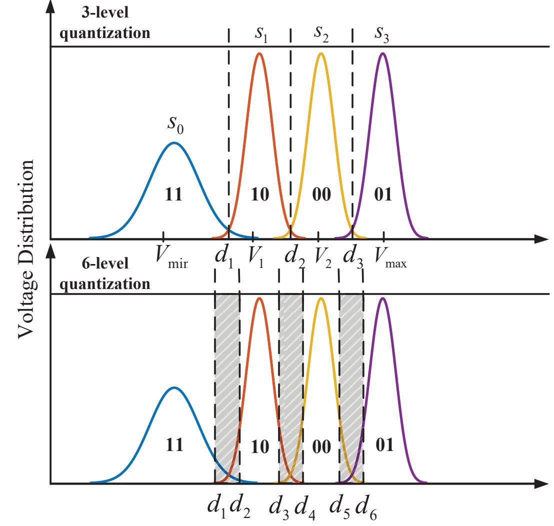

Fig. 1. Illustration of the 6-level read scheme for the MLC NAND flash memory.

Let $U_{k,n} = (u_{k,1}, u_{k,2}, \ldots, u_{k,n}), k = 1, \ldots, K$ be the encoded input of K page for the flash memory. Every K bits $(u_{1,j}, u_{2,j}, \ldots, u_{K,j})$ are mapped to a uniform distribution and K is corresponding to the number of stored bits of a memory cell. Then, the encoded codeword U will be written to n^* memory cells with the corresponding voltages connected to different states S. When we want to acquire the information of this codeword, the voltages of n^* memory cells will be read and quantized into various regions R, which are connected to several LLR values. The decoded codeword of k page for NAND flash memory can be denoted by $Y_{k,n} = (y_{k,1}, y_{k,2}, \ldots, y_{k,n}), k = 1, \ldots, K$. The case of $K = 2$ represents a 2-bit per cell flash channel model. Note that the conventional flash memory coded system is shown in Fig. 2.

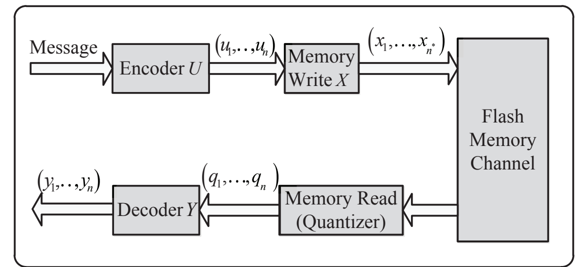

Fig. 2. System diagram for the MLC NAND flash memory.

III. Read-Voltage Design for the Flash Memory Channel

The problem of finding good channel quantizers is reported in [11], [12]. However, they are based on shannon capacity or entropy-based function, which can't consider the effect of code blocklength. In this section, we derive the better read-voltage design with a particular code for the flash memory.

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:27:52 UTC from IEEE Xplore. Restrictions apply.

A. Channel-Coding Rate based Read-Voltage Design

As [13] mentioned, for a finite blocklength code, the achievable CCR $\mathcal{R}(n,\epsilon)$ with a given error probability is given by

$$\mathcal{R}(n,\epsilon)=\frac{\log \mathcal{M}(n,\epsilon)}{n}, \tag{4}$$

$$\log \mathcal{M}(n,\epsilon) \geq nI(P,W)-\sqrt{nU(P,W)}Q^{-1}(\epsilon) + O(1), \tag{5}$$

where $\mathcal{M}(n,\epsilon)$ is the number of codewords in a codebook, ϵ is the average probability of error, n is the blocklength, the maximum likelihood (ML) decoder is considered in this equation and $Q(x)$ is the tail of the standard Gaussian random variable, given by

$$Q(x)=\int_x^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt. \tag{6}$$

In (5), $I(P,W)$ is the mutual information between input and output, given by

$$I(P,W)=\sum_{s_i\in S}\sum_{r_j\in R}p_{s_i}p_{r_j,s_i}\log_2\frac{p_{r_j,s_i}}{pw_{r_j}}, \tag{7}$$

and $U(P,W)$ is the unconditional information variance, given by

$$U(P,W)=\sum_{s_i\in S}\sum_{r_j\in R}p_{s_i}p_{r_j,s_i}\left(\log_2\frac{p_{r_j,s_i}}{pw_{r_j}}\right)^2-[I(P,W)]^2, \tag{8}$$

where p_{s_i} is the probability of input s_i, p_{r_j,s_i} is a conditional probability mass function from s_i to r_j. The conditional probability mass function W is determined by the read voltages, given by

$$p_{r_j,s_i}=\int_{d_{j^*-1}}^{d_{j^*}}p_{s_i}(v)dv \tag{9}$$

where d_{j^*-1} and d_{j^*} are the lower bound and upper bound of the region r_j. In (7) and (8), the quantity pw_{r_j} represents the output distribution, given by

$$pw_{r_j}=\sum_{r_j\in R}p_{s_i}p_{r_j,s_i}. \tag{10}$$

Considering the constraint of the practical code, the loss of the error performance is existing. We use A to represent the loss caused by the code and B to compensate for the inequality. Combining the above calculation with (4) and (5), for $\delta=B-A<\log n$, we have

$$\mathcal{R}(n,\epsilon)+\frac{A}{n}=I(P,W)-\sqrt{\frac{U(P,W)}{n}}Q^{-1}(\epsilon) + \frac{B}{n}, \tag{11}$$

$$\mathcal{R}(n,\epsilon)=I(P,W)-\sqrt{\frac{U(P,W)}{n}}Q^{-1}(\epsilon)+\frac{\delta}{n}, \tag{12}$$

It is easy to see, the error probability ϵ of a practical code can be denoted as

$$\epsilon(n,\mathcal{R})=Q\left\{\left[I(P,W)-\mathcal{R}+\frac{\delta}{n}\right]\sqrt{\frac{n}{U(P,W)}}\right\}, \tag{13}$$

where $\mathcal{R}$ is the code rate.

When we optimize the read voltages, $d_1,d_2,\ldots,d_J$, the conditional probability function p_{r_j,s_i} and the quantity pw_{r_j} will be changed. According to (13), this optimization problem can be seen as finding the $d_1,d_2,\ldots,d_J$ to minimize the ϵ. Note that this read-voltage design scheme for the flash memory is dynamic. The optimization function of this model is given by

$$(d_1^*,d_2^*,\ldots,d_J^*)=\min_{d_1,d_2,\ldots,d_J}\epsilon(d_1,d_2,\ldots,d_J,n,R^*), \tag{14}$$

where $d_1^*,d_2^*,\ldots,d_J^*$ is the optimization result. Note that the monotonicity of the function Q is decreasing, so the optimization function can be modified as

$$(d_1^*,d_2^*,\ldots,d_J^*)=\max_{d_1,d_2,\ldots,d_J}\left\{\left[I(P,W)-\mathcal{R}+\frac{\delta}{n}\right]\sqrt{\frac{n}{U(P,W)}}\right\}. \tag{15}$$

Considering the voltages of memory cells, the $d_1,d_2,\ldots,d_J$ must satisfy the following constraint, given by

$$0<d_1<d_2<\cdots<d_J<V_{max}. \tag{16}$$

B. 6-level Read for Mlc Flash Memory

The channel capacity with 6-level read scheme registers a big jump from 3-level quantization [12], so we strive to optimize the 6-level read scheme. Assume U is equally likely to be 0 or 1 and the read level is 6 in our model. Considering the case of $K=2$, combining the above calculation with (9) and (10), we have

$$pw_{r_j}=\frac{1}{4}p_{r_j,s_0}+\frac{1}{4}p_{r_j,s_1}+\frac{1}{4}p_{r_j,s_2}+\frac{1}{4}p_{r_j,s_3}, \tag{17}$$

where $j=0,1\ldots,6$. The mutual information of 2-bits per cell of flash memory can be represented as

$$I(P,W)=C=I(S;R), \tag{18}$$

where $I(S;R)$ is the mutual information between input symbol $S=\{s_0,s_1,s_2,s_3\}$ and output symbol $R=\{r_0,\ldots,r_6\}$. The $H(R)$ and $H(R|S)$ represent the entropy and conditional

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:27:52 UTC from IEEE Xplore. Restrictions apply.

entropy of output R, respectively, and given as

$$\begin{aligned}
&I(P,W) = C = H(R) - H(R|S)\\
&= H\left(\frac{p_{r_0,s_0}+p_{r_0,s_1}+p_{r_0,s_2}+p_{r_0,s_3}}{4},\right.\\
&\frac{p_{r_1,s_0}+p_{r_1,s_1}+p_{r_1,s_2}+p_{r_1,s_2}}{4},\\
&\frac{p_{r_2,s_0}+p_{r_2,s_1}+p_{r_2,s_2}+p_{r_2,s_3}}{4},\\
&\frac{p_{r_3,s_0}+p_{r_3,s_1}+p_{r_3,s_2}+p_{r_3,s_3}}{4},\\
&\frac{p_{r_4,s_0}+p_{r_4,s_1}+p_{r_4,s_2}+p_{r_4,s_3}}{4},\\
&\frac{p_{r_5,s_0}+p_{r_5,s_1}+p_{r_5,s_2}+p_{r_5,s_3}}{4},\\
&\left.\frac{p_{r_6,s_0}+p_{r_6,s_1}+p_{r_6,s_2}+p_{r_6,s_3}}{4}\right)\\
&-\frac{1}{4}H\left(p_{r_0,s_0},p_{r_1,s_0},p_{r_2,s_0},p_{r_3,s_0},p_{r_4,s_0},p_{r_5,s_0},p_{r_6,s_0}\right)\\
&-\frac{1}{4}H\left(p_{r_0,s_1},p_{r_1,s_1},p_{r_2,s_1},p_{r_3,s_1},p_{r_4,s_1},p_{r_5,s_1},p_{r_6,s_1}\right)\\
&-\frac{1}{4}H\left(p_{r_0,s_2},p_{r_1,s_2},p_{r_2,s_2},p_{r_3,s_2},p_{r_4,s_2},p_{r_5,s_2},p_{r_6,s_2}\right)\\
&-\frac{1}{4}H\left(p_{r_0,s_3},p_{r_1,s_3},p_{r_2,s_3},p_{r_3,s_3},p_{r_4,s_3},p_{r_5,s_3},p_{r_6,s_3}\right).
\end{aligned} \tag{19}$$

To show the CCR under different read levels and blocklength, in Fig. 3, we plot the cell storage capacity versus flash memory PE cycles. The Shannon capacity of the flash memory is using (19). The CCR in the finite blocklength code regime is using (4) and (5). As can be seen, the loss of channel capacity caused by the blocklength is obvious when the blocklength is small.

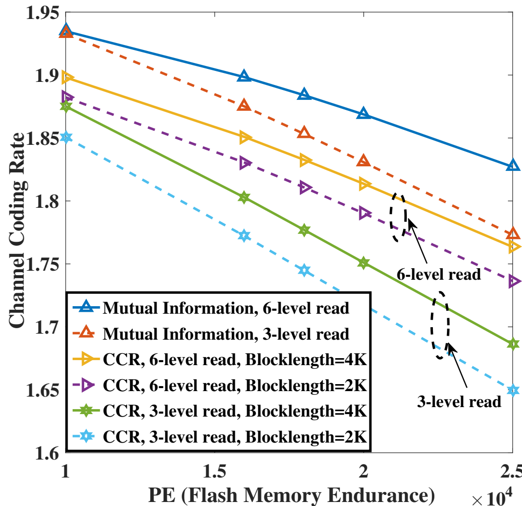

Fig. 3. Cell channel coding rate plotted against memory PE cycles for 2-bit per cell NAND flash memory with 3-level and 6-level of memory sensing.

IV. Genetic Algorithm and Cross Iterative Searching Algorithm

A specified constant ratio is used to select thresholds so that the three erasure regions have the same size in the simple symmetric Gaussian model [5]. In [11], the bisection search or other quasi-convex optimization technique is used to compute the MMI quantization levels. Several intelligence algorithms can be utilized to obtain the optimized quantization levels in our model and also be extended for various read-levels.

A. Nonlinear Optimization based Genetic Algorithm

Since (15) are continuous and smooth function and we can select appropriate quantization levels. Conventional nonlinear optimization often has high local searching ability but low global searching ability. The genetic algorithm is an optimization algorithm which mimics the biological evolution. This algorithm uses survival of the fittest as a method to achieve a good solution for the optimization problem. In this algorithm, the evolution is achieved using set of stochastic genetic operator which mimics the natural process of reproduction and mutation. Combining the genetic algorithm and nonlinear optimization, we can escape from local minima and obtain excellent results in our model.

B. Cross Iterative Searching Algorithm

Note that the optimization problem in (15) is a continuous convex-optimization problem. All the read voltages are constrained to the equation (16). The cross iterative searching (CIS) algorithm requires the optimization function and is used to obtain the read voltages in the flash memory.

Algorithm 1: Cross Iterative Searching Algorithm

Data: the optimization function $\epsilon(d_1, d_2, \ldots, d_J, n, \mathcal{R})$, maximum iterations I_{max}, threshold θ, blocklength n, code rate $\mathcal{R}$

Result: the read voltages $d_1, d_2, \ldots, d_J$

1 Initialization: $i \leftarrow 0$;
2 **while** $|\epsilon^{(i)} - \epsilon^{(i-1)}| > \theta$ *and* $i < I_{max}$ **do**
3 $\quad i \leftarrow i + 1$;
4 $\quad$ **Optimize the** $d_j^{(i)}$ **serially**;
5 $\quad$ **while** $j \leq J$ **do**
6 $\quad\quad$ Search for the optimal $d_j^{(i)}$ using the
7 $\quad\quad$ optimization function:
8 $\quad\quad (d_j^{(i)}) = \min_{d_j^{(i)}} \epsilon(d_1^{(i)}, \ldots, d_j^{(i-1)}, \ldots, d_J^{(i-1)}, n, \mathcal{R})$;
9 $\quad\quad j \leftarrow j + 1$
10 $\quad$ **Perform**;
11 $\quad$ calculate the newest value:
$\quad\epsilon^{(i)}(d_1^{(i)}, d_2^{(i)}, \ldots, d_J^{(i)}, n, \mathcal{R})$;

To begin with, the CIS determines the value ranges of these objective read voltages, which can reduce the searching space. Note that the iterative sequence among these read voltages will not affect the optimized results. Hence, we optimize these

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:27:52 UTC from IEEE Xplore. Restrictions apply.

read voltages in a serial sequence. When a decision boundary is optimizing by the searching method, other read voltages are fixed. After all the boundaries have been updated, we will compare the difference value $|\epsilon^{(i)} - \epsilon^{(i-1)}|$. When it is up to the end condition, we stop the iteration and obtain the results. The detailed steps of the CIS algorithm are outlined in Algorithm 1.

V. Simulation Results

To compare the error-performance of finite-blocklength channel-coding rate method (FCR) with conventional read methods, we simulate two binary QC-LDPC codes, referred to as $2K$-*code* and $4K$-*code*. In the $4K$-*code*, each entry of a small base matrix H_B (7×71) is replaced by either a circulant shift of a 64×64 identify matrix or an all-zero 64×64 matrix. The block-length of this code is 4544 bits and code rate is set as 0.9. This irregular code has column-weight $d_v = 5$ and row-weight $d_c = 50$ or 51. The $2K$-*code* is chosen as a QC-LDPC code with uniform column-weight of 4 and row-weight of 40 or 41. The block-length and code rate of this code are the same as $4K$-*code*.

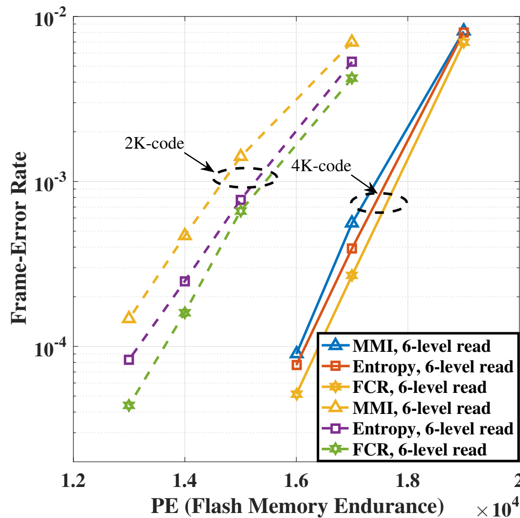

Fig. 4. Frame-error-rate (FER) performance of LDPC $2K$-*code* and $4K$-*code* plotted versus maximum decoding iterations 25 at different PE cycles.

In Fig.4, we plot the FER curves over different PE cycles while employing the proposed read method and other methods using the $2K$-*code* and $4K$-*code* above. The decoding algorithm and optimization algorithm we used are sum-product algorithm and CIS or genetic algorithm. We assume that the PE count at PE = 19K, 17K, 16K, 15K, 14K and 13K, while setting the retention time to zero (T=0, representing the early retention time). We can notice that the flash channel tolerance against the PE-cycles-induced errors is significantly improved using the proposed scheme. For instance, we assume that the required performance of frame-error rate (FER) is at 2×10^{-4}. The MMI scheme and entropy-based quantization can approximately endure 13300 and 13600 PE cycles when we use the $2K$-*code*, respectively. In contrast, for the same FER, the endurance limit of the PE cycles is extended to 14200 using the the proposed scheme. With the $4K$-*code*, the proposed scheme also can improve the PE cycles by up to 300 and 500 cycles against to the MMI and entropy-based quantization, respectively.

VI. Conclusions

Considering the CCR of a finite blocklength code, we have proposed an effective read-voltage method for MLC NAND flash memory. The loss of the finite blocklength and limit of cell storage capacity have been considered in our proposed method. In previous work, the research has been carried over simplified Gaussian channel. However, we have analyzed the asymmetric models using FCR method to select read voltages. To obtain read voltages of this complicated model, the genetic algorithm and a CIS method is employed in this paper. With our proposed FCR method, error-performance of the flash memory can be further improved at the different PE cycles. Additional lifetime extension is expected for a general flash memory channel model.

References

[1] K. Kim, "Future memory technology: Challenges and opportunities," in *Proc. Int. Symp. VLSI Technol. Syst. Appl.*, San Jose, CA, USA, Apr. 2008, pp. 5–9.

[2] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis," in *Proc. Des. Autom. Test Eur. (DATE'12)*, San Jose, CA, USA, Mar. 2012, pp. 521–526.

[3] Q. Li, A. Jiang, and E. F. Haratsch, "Noise modeling and capacity analysis for NAND flash memories," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun. 2014, pp. 2262–2266.

[4] S. G. Cho, D. Kim, J. Choi, and J. Ha, "Block-wise concatenated BCH codes for NAND flash memories," *IEEE Trans. Commun.*, vol. 62, no. 4, pp. 1164–1177, Apr. 2014.

[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, Reg. Papers*, vol. 58, no. 2, pp. 429–439, Feb. 2011.

[6] C. A. Aslam, Y. L. Guan, and K. Cai, "Non-binary LDPC code with multiple memory reads for multi-level-cell (MLC) flash," in *Proc. Int. Conf. APSIPA*, Dec. 2014, pp. 1–9.

[7] R. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, Jan. 1963.

[8] G. Han, Y. L. Guan, and X. Huang, "Check node reliability-based scheduling for BP decoding of non-binary LDPC codes," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 877–885, Mar. 2013.

[9] H. Xiao and A. H. Banihashemi, "Graph-based message-passing schedules for decoding LDPC codes," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2098–2105, Dec. 2004.

[10] E. Sharon, S. Litsyn, and J. Goldberger, "Efficient serial message-passing schedules for LDPC decoding," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4076–4091, Nov. 2007.

[11] 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.

[12] 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.

[13] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:27:52 UTC from IEEE Xplore. Restrictions apply.