





Article

Channel Modeling and Quantization Design for 3D NAND Flash Memory

Cheng Wang ¹ , Zhen Mei ^{1,2,*} , Jun Li ^{1,*} , Feng Shu ³, Xuan He ⁴ and Lingjun Kong ⁵ 

¹ School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; cheng.wang@njjust.edu.cn

² National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

³ School of Information and Communication Engineering, Hainan University, Haikou 570228, China; shufeng@njjust.edu.cn

⁴ School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China; xhe@swjtu.edu.cn

⁵ Jinling Institute of Technology, Nanjing 211169, China; kong@jit.edu.cn

* Correspondence: meizhen@njjust.edu.cn (Z.M.); jun.li@njjust.edu.cn (J.L.)

Abstract: As the technology scales down, two-dimensional (2D) NAND flash memory has reached its bottleneck. Three-dimensional (3D) NAND flash memory was proposed to further increase the storage capacity by vertically stacking multiple layers. However, the new architecture of 3D flash memory leads to new sources of errors, which severely affects the reliability of the system. In this paper, for the first time, we derive the channel probability density function of 3D NAND flash memory by taking major sources of errors. Based on the derived channel probability density function, the mutual information (MI) for 3D flash memory with multiple layers is derived and used as a metric to design the quantization. Specifically, we propose a dynamic programming algorithm to jointly optimize read-voltage thresholds for all layers by maximizing the MI (MMI). To further reduce the complexity, we develop an MI derivative (MID)-based method to obtain read-voltage thresholds for hard-decision decoding (HDD) of error correction codes (ECCs). Simulation results show that the performance with jointly optimized read-voltage thresholds can closely approach that with read-voltage thresholds optimized for each layer, with much less read latency. Moreover, the MID-based MMI quantizer almost achieves the optimal performance for HDD of ECCs.

Keywords: 3D flash memory; information theory; channel modeling; read-voltage thresholds; quantization; LDPC codes



Citation: Wang, C.; Mei, Z.; Li, J.; Shu, F.; He, X.; Kong, L. Channel Modeling and Quantization Design for 3D NAND Flash Memory. *Entropy* **2023**, *25*, 965. <https://doi.org/10.3390/e25070965>

Academic Editors: Pingyi Fan, Qi Chen, Suihua Cai and Gangtao Xin

Received: 31 May 2023

Revised: 17 June 2023

Accepted: 19 June 2023

Published: 21 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

NAND flash memory has been widely used in consumer electronic devices over the past decade due to non-volatility, fast write and read speed, and low power consumption [1,2]. Moreover, as the development of process technology and multi-level-cell (MLC) technique, the storage capacity and density of NAND flash memory have been greatly increased. However, as the technology continues to scale down, the storage capacity of NAND flash memory has reached its bottleneck and the reliability of flash memory is severely affected by various types of errors, such as program/erase (P/E) cycling errors, cell-to-cell interference, retention interference, and program errors [3]. The reliability of flash memory is not only affected by P/E cycles and data retention time, but also influenced by the temperature and process variations [4–7].

Recently, three-dimensional (3D) stacking technology has been applied in NAND flash memory, in which several layers are stacked in the vertical direction [8]. With 3D stacking technology, the density and storage capacity of NAND flash memory are tremendously increased compared with two-dimensional (2D) flash memory. However, the new architecture of 3D NAND flash memory leads to new sources of errors [9–11]. For example, due to

process variations, the memory cells in different layers of 3D NAND flash memory have shown different error characteristics [6,7,12,13]. In addition, early retention loss occurs in the several hours after programming due to fast charge leakage [14]. As a result, the noise characteristics of 3D flash memory channel are significantly different from that of the 2D flash memory and it is worth investigating the channel modeling of 3D flash memory.

To improve the reliability of 3D flash memory system, advanced error correcting codes (ECCs) are essential. Recently, low-density parity-check (LDPC) codes [15] with hard and soft decision decoding have been considered for 3D flash memory [16–18]. For LDPC codes with hard-decision decoding (HDD), the performance mainly depends on the raw bit error rate of the flash memory channel [19,20]. The raw bit error rate of 3D flash memory increases when data retention time and P/E cycles increase. To further improve the error-correction performance, LDPC codes with soft-decision decoding can be employed and its performance heavily depends on the accuracy of channel log-likelihood ratios. Therefore, increasing the number of read-voltage thresholds (i.e., quantization levels) will effectively improve the accuracy of log-likelihood ratios [18,21,22]. However, for flash memory channel, high-precision quantization is not feasible due to the read latency restriction. It is essential to design a quantization scheme with limited read-voltage thresholds to achieve near-optimal performance.

1.1. Related Work

Many studies have examined the quantization design for 2D flash memory under limited number of read-voltage levels [2,18,21,23–25]. These prior works optimized the read-voltage thresholds with perfect knowledge of channel information, e.g., the number of P/E cycles and data retention time. Based on the well-established channel model, they adopted a non-uniform quantization strategy to reduce the number of read-voltage thresholds and improve the performance by using the maximizing the mutual information (MMI)-based, entropy-based, and finite blocklength-based quantization design methods, respectively.

To improve the system reliability for 3D flash memory, several studies have focused on dealing with new error sources (e.g., layer-to-layer variation) in 3D flash memory. For example, a sentinel-cell approach was proposed in [26] to utilize the error characteristics of memory cells for different layers to obtain optimal read-voltage thresholds. A layer variation-aware reading method was proposed to obtain optimal read-voltage thresholds by recording the voltage offset of each layer in a table in flash memory controller [9]. By exploiting asymmetric error characteristics of 3D flash memory, three asymmetric sensing schemes were proposed to reduce the number of sensing levels and maintain considerable performance [27]. A polynomial-based method was proposed to fit the voltage distribution of flash memory, and optimized read-voltage thresholds by a least square method [28]. A process-variation-aware strategy was proposed to reduce uncorrectable bit errors by storing important information in the flash memory block with higher reliability grades [6]. In addition, a voltage compensation strategy based on the reliability grades was proposed to tune the range of sensing voltages for unreliable pages [7]. However, due to the lack of an accurate 3D flash memory channel model, the above work only focused on proposing new strategies to mitigate the adverse effects of specific error sources. It is difficult to systematically design the read-voltage thresholds and analyze the performance theoretically.

In terms of channel modeling and error analysis for 3D flash memory, previous studies in [11,29] analyzed the common error sources in 3D flash memory, such as incremental step pulse programming noise, cell-to-cell interference, and data retention noise. They found new error sources in 3D flash memory (e.g., layer-to-layer variation and early retention loss) and modeled 3D flash memory channel by using Gaussian distribution. In addition, a neural network-based model was proposed to predict the voltage distribution [30]. However, these work failed to derive the joint probability distributions of different types of errors in 3D flash memory.

Therefore, it is essential to analyze the statistics of major errors of 3D flash memory, and derive a mathematical formulation for the channel model. Using this channel model, the mutual information (MI) of 3D flash memory can be derived, and the quantization can be designed theoretically.

1.2. Contributions

The main contributions of this paper are as follows:

- For the first time, we derive an analytic channel model for 3D NAND MLC flash memory, which considers major sources of errors.
- Rather than optimizing read-voltage thresholds for each layer, we jointly design the quantization for multiple layers of 3D MLC flash memory by MMI. By doing this, the number of total read thresholds for multiple layers can be greatly reduced, such that the storage cost and read latency can be significantly decreased.
- A dynamic programming (DP) algorithm is proposed to optimize read-voltage thresholds for 3D MLC flash memory by MMI of the joint channel.
- To further reduce the complexity of DP algorithm, the 3D MLC flash memory channel is simplified and a MI derivative (MID)-based method is developed to obtain read-voltage thresholds for ECCs with HDD.

The remainder of this paper is organized as follows. In Section 2, we derive the channel model for 3D MLC flash memory and further propose a quantized channel model. In Section 3, we formulate the read-voltage thresholds optimization problem of 3D MLC flash memory with multiple layers and utilize MMI-DP algorithm to solve this optimization problem. Then, to further reduce the complexity, we simplify 3D MLC flash memory channel and propose a MID-based quantization scheme. The simulation results are presented in Section 4, and we conclude this paper in Section 5.

2. Channel Modeling of 3D NAND MLC Flash Memory

2.1. 3D NAND Flash Memory Basics

The 3D NAND flash memory has a vertically stacked structure, which includes word-lines and bit lines located in different layers [8,31]. In 3D NAND flash memory, the information is stored as charges in memory cells [8]. As shown in Figure 1, for MLC flash memory, a memory cell can store two bits of information, which can be represented by four storage states (i.e., $\mathbf{S} = \{s_0, s_1, s_2, s_3\}$). To write data into a storage cell, the flash memory controller must erase the entire block that this cell belongs to [32]. After erasing, the voltage of this cell is programmed to a specific value by an incremental step pulse programming technique [33]. To read data from a storage cell, the flash memory controller applies read threshold voltage V_{th} (e.g., d_1, d_2, d_3) to the word-line. By comparing the voltage of storage cell with these read-voltage thresholds, we can determine the stored bits or obtain soft information for further ECC decoding [34].

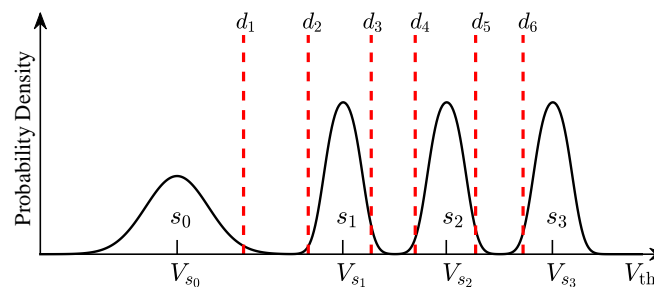


Figure 1. Illustration of voltage distributions and read-voltage thresholds for MLC flash memory.

2.2. Channel Modeling

To investigate the optimal read-voltage thresholds design, the channel modeling of 3D NAND MLC flash memory is essential. In this subsection, we propose a simplified

channel model for 3D NAND MLC flash memory based on the experimental data in [9] by considering major sources of errors.

2.2.1. Initial Threshold Voltage Distribution

With the reference to [9], the threshold voltage distribution of the memory cell at erased state s_0 is approximately modeled as a Gaussian distribution $p_{u_0}(v) \sim \mathcal{N}(V_{s_0}, \sigma_u^2)$ with mean V_{s_0} and standard deviation σ_u , respectively. Moreover, the voltages at programmed states (i.e., s_1, s_2, s_3) are generated by the incremental step pulse programming technique. Hence, the threshold voltage distribution of each programmed state $s_i, i = 1, 2, 3$, follows a uniform distribution

$$p_{u_i}(v) = \begin{cases} 1/V_p, & v \in [V_{s_i}, V_{s_i} + V_p) \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where V_p denotes the programming step voltage and V_{s_i} denotes the target programmed voltage of programmed state s_i . The programmed cells are also affected by the programming noise, which can also be modeled by a Gaussian distribution $n_{p_i}(v) \sim \mathcal{N}(0, \sigma_{p_i}^2)$, $i = 1, 2, 3$.

2.2.2. Program/Erase Cycling Errors

P/E cycling errors n_{pe} appear immediately after the programming and erasing operations. As the number of P/E cycles increases, more and more electrons are trapped in the transistor, which will reduce the tunneling efficiency and result in inaccurate charge transport. The P/E cycling errors can be approximately modeled as a Gaussian distribution $p_{n_{pe}}(v) \sim \mathcal{N}(\mu_{pe}, \sigma_{pe}^2)$, where μ_{pe} and σ_{pe} are the mean and standard deviation, respectively, [9].

2.2.3. Cell-to-Cell Interference

Due to the parasitic capacitance-coupling effect, one cell can affect the voltage distribution of its neighbor cells. The voltage of the victim cell is affected by that of its adjacent cells, leading to program interference errors. Note that, different from 2D MLC flash memory, the voltage of the victim cell of 3D MLC flash memory can also be affected by its neighbor layers. As mentioned in [29,35], the cell-to-cell interference can be estimated and mitigated by the pre-distortion or post-compensation techniques.

2.2.4. Early Retention Loss

As the data retention time increases, the charge loss will move the voltage distribution of flash memory to lower states [36]. Compared with 2D flash memory, the memory cells in 3D flash memory suffer from early retention loss, which presents a large amount of charge loss within a few minutes after being programmed. According to [9,11], the early retention noise can be modeled as a Gaussian distribution, $p_{n_r}(v) \sim \mathcal{N}(\mu_r, \sigma_r^2)$, where μ_r and σ_r are the mean and standard deviation, respectively.

2.2.5. Layer-to-Layer Process Variation

Due to process variation, the error characteristics of each layer are quite different. The layer-to-layer process variation of each layer can be modeled as a Gaussian distribution $p_{n_{l_k}}(v) \sim \mathcal{N}(\mu_{l_k}, \sigma_{l_k}^2)$ [9], where μ_{l_k} and σ_{l_k} are the mean and standard deviation of the data stored in the k -th layer, respectively. According to the experimental results in [9], it shows the mean and standard deviation of each state vary with the layer. To evaluate the influence of process variations for different layers, we set the first layer as the reference layer and fit the mean and standard deviation of other layers using a polynomial fitting method.

Finally, the overall voltage distribution can be calculated by the convolution integral of initial voltage distribution functions with other major noises [2,21]. As mentioned above, the erased state s_0 is not affected by programming noise, its voltage distribution for the k -th layer is given as

$$p_{s_{0,k}}(v) = p_{u_0} \otimes p_{n_{pe}} \otimes p_{n_r} \otimes p_{n_{l_k}} \\ = \frac{1}{\sigma_{s_{0,k}} \sqrt{2\pi}} e^{-\frac{(v-\mu_{s_{0,k}})^2}{2\sigma_{s_{0,k}}^2}}, \quad (2)$$

where \otimes denotes the convolution operation, $\mu_{s_{0,k}} = V_{s_0} + \mu_{pe} + \mu_r + \mu_{l_k}$, $\sigma_{s_{0,k}}^2 = \sigma_u^2 + \sigma_{pe}^2 + \sigma_r^2 + \sigma_{l_k}^2$. The voltage distributions of the programmed states s_i ($i = 1, 2, 3$) for the k -th layer are given as

$$p_{s_{i,k}}(v) = p_{u_i} \otimes p_{n_{p_i}} \otimes p_{n_{pe}} \otimes p_{n_r} \otimes p_{n_{l_k}} \\ = \frac{1}{2V_p} \left(\operatorname{erf} \left(\frac{V_{s_i} + V_p - v - \mu_{s_{i,k}}}{\sqrt{2}\sigma_{s_{i,k}}} \right) \right) - \\ \frac{1}{2V_p} \left(\operatorname{erf} \left(\frac{V_{s_i} - v - \mu_{s_{i,k}}}{\sqrt{2}\sigma_{s_{i,k}}} \right) \right), \quad i = 1, 2, 3, \quad (3)$$

where i denotes the index of the storage states, $\mu_{s_{i,k}} = \mu_{pe} + \mu_r + \mu_{l_k}$, $\sigma_{s_{i,k}}^2 = \sigma_{p_i}^2 + \sigma_{pe}^2 + \sigma_r^2 + \sigma_{l_k}^2$, and $\operatorname{erf}(\cdot)$ denotes the Gauss error function.

According to the experimental data of 3D NAND flash memory given in [9], the parameters of voltage distributions with layer-to-layer variation can be obtained by fitting the experimental data using the least square method, listed in Table 1. Similar to [9], the voltage values are normalized in this work. Finally, the parameters of 3D flash memory voltage distribution can be calculated by $\mu_{s_{i,k}} = \mu_{s_i}^* + \mu_{l_{i,k}}$, and $\sigma_{s_{i,k}}^2 = (\sigma_{s_i}^*)^2 + \sigma_{l_{i,k}}^2$.

Table 1. Parameters of 3D NAND Flash Memory with K layers.

Variable A	Variable A = $(\alpha \times PE + \beta) \times \log(t) + \gamma \times PE + \delta$				Variable B	Variable B = $\alpha \times k^3 + \beta \times k^2 + \gamma \times k$		
	α	β	γ	δ		α	β	γ
$\mu_{s_0}^*$	1.01×10^{-4}	0.74	4.2×10^{-3}	−67.27	$\mu_{l_{0,k}}$	0	−0.028	1.94
$\mu_{s_1}^*$	-1.94×10^{-5}	−0.4	5.14×10^{-4}	106.47	$\mu_{l_{1,k}}$	0	0	0.0075
$\mu_{s_2}^*$	-4.71×10^{-5}	−0.7	1.94×10^{-4}	183.58	$\mu_{l_{2,k}}$	0	0	−0.0447
$\mu_{s_3}^*$	-7.37×10^{-5}	−1.2	4.68×10^{-4}	252.85	$\mu_{l_{3,k}}$	0	0	−0.0308
$\sigma_{s_0}^*$	1.2×10^{-5}	−0.1	2.1×10^{-4}	14.01	$\sigma_{l_{0,k}}$	0	−0.0048	0.185
$\sigma_{s_1}^*$	-1.34×10^{-6}	0.0098	1.56×10^{-4}	8.2	$\sigma_{l_{1,k}}$	0	−0.0045	0.153
$\sigma_{s_2}^*$	-2.12×10^{-6}	0.0098	1.09×10^{-4}	9.65	$\sigma_{l_{2,k}}$	-1.8×10^{-5}	9.1×10^{-4}	−0.037
$\sigma_{s_3}^*$	2.87×10^{-6}	0.014	8.5×10^{-5}	9.83	$\sigma_{l_{3,k}}$	7.86×10^{-5}	−0.0034	0.0129

PE: number of program/erase cycles; t : data retention time/second; k : layer index.

2.3. Quantized Model of 3D NAND Flash Memory

In this paper, we assume that the 3D NAND MLC flash memory has a total of K layers. Since the read process transforms voltages into discrete values, the flash memory channel is quantized into a discrete memoryless channel. For MLC flash memory, each cell can store two bits of information and has four storage states, s_0, s_1, s_2 , and s_3 . Figure 1 shows an example of the voltage distribution with read-voltage thresholds (i.e., red dotted line). Let $\mathbf{S} = \{s_0, s_1, s_2, s_3\}$ denote the storage states of MLC flash memory. Given J thresholds, the flash memory channel is quantized into $J+1$ levels. Let $\mathbf{D}_k = \{d_{1,k}, d_{2,k}, \dots, d_{J,k}\}$ denote J read-voltage levels, and $\mathbf{R}_k = \{r_{0,k}, r_{1,k}, \dots, r_{J,k}\}$ denote $J+1$ channel outputs in the k -th layer, where $r_{j,k} \in [d_{j,k}, d_{j+1,k})$ with $d_{0,k} = -\infty$ and $d_{J+1,k} = +\infty$. In addition, $d_{1,k} < d_{2,k} < \dots < d_{J,k}$. As shown in Figure 2, in the k -th layer, read voltage quantization produces a discrete memoryless channel with inputs $\{s_0, s_1, s_2, s_3\}$ and outputs $\{r_{0,k}, r_{1,k}, \dots, r_{J,k}\}$. The transition probability of storing s_i and quantizing as $r_{j,k}$ in the k -th layer is given by

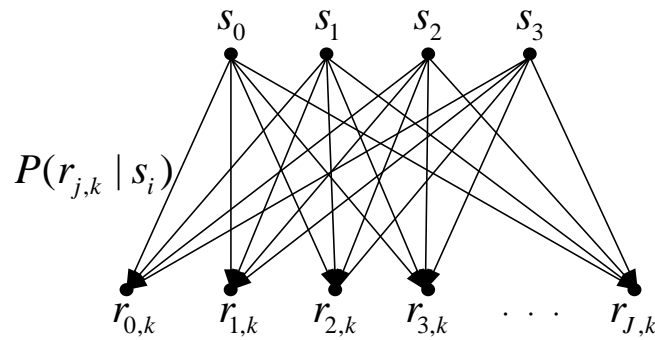


Figure 2. Equivalent discrete memoryless channel for 3D NAND MLC flash memory in the k -th layer.

$$P(r_{j,k} | s_i) = \int_{d_{j,k}}^{d_{j+1,k}} p_{s_i,k}(v) dv, \quad (4)$$

and

$$P(r_{j,k}) = \sum_{i=0}^3 P(s_i) P(r_{j,k} | s_i), \quad (5)$$

where $P(s_i)$ is the probability of the flash memory cell stores s_i . Due to the fact that the scrambler in the flash controller randomly perturbs data bits to ensure the stored data bits (1 or 0) are evenly distributed [20], the probability of input is $P(s_i) = \frac{1}{4}$. With this quantized model of 3D flash memory, the MI of the 3D MLC flash memory channel will be derived in the next section, and the read-voltage thresholds can be optimized by MMI.

3. Quantization Design for 3D NAND Flash Memory

In this section, the MI of 3D NAND MLC flash memory channel is derived based on the quantized channel model. Then, a MMI-DP algorithm is proposed to optimize read-voltage thresholds. Moreover, to reduce the complexity of DP-based MMI quantization algorithm, we propose a MID-based quantization read-voltage thresholds design for ECCs with HDD.

3.1. MI for 3D NAND MLC Flash Memory

The voltage distributions of different layers in 3D NAND MLC flash memory have different statistical properties. Conventional methods, such as the MMI-based quantization [18], entropy-based quantization [21], can still optimize read-voltage thresholds for each layer in 3D NAND flash memory. However, applying different read-voltage thresholds in different layers requires more read operations. Since the read operation is a time-consuming process in flash memory [37], the conventional methods of optimizing read-voltage thresholds for each layer are not desired for 3D NAND flash memory. Therefore, it is necessary to jointly optimize read-voltage thresholds for all layers to reduce the read overhead of 3D flash memory.

For 3D NAND flash memory channel with K layers, let \mathcal{S} and \mathcal{R} denote the inputs and outputs of this channel, respectively. Then, the MI of 3D NAND flash memory channel can be calculated by

$$\begin{aligned} I(\mathcal{S}; \mathcal{R}) &= I(\mathbf{S}_1 \mathbf{S}_2 \cdots \mathbf{S}_K; \mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_K) \\ &= H(\mathcal{R}) - H(\mathcal{R} | \mathcal{S}) = H(\mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_K) - \sum_{k=1}^K H(\mathbf{R}_k | \mathbf{S}_k), \end{aligned} \quad (6)$$

where $\mathbf{S}_k \in \mathcal{S}$ and $\mathbf{R}_k \in \mathcal{R}$, $H(\cdot)$ denotes the entropy.

The MI for the k -th layer 3D NAND flash memory channel can be calculated by Equation (7).

$$\begin{aligned}
 I(\mathbf{S}_k; \mathbf{R}_k) &= H(\mathbf{R}_k) - H(\mathbf{R}_k | \mathbf{S}_k) = H(\mathbf{S}_k) - H(\mathbf{S}_k | \mathbf{R}_k) \\
 &= \sum_{j=0}^J \sum_{i=0}^3 P(s_i) P(r_{j,k} | s_i) \log \frac{P(r_{j,k} | s_i)}{\sum_{i=1}^3 P(r_{j,k} | s_i) P(s_i)} \\
 &= \sum_{j=0}^J \sum_{i=0}^3 P(s_i) \int_{d_{j,k}}^{d_{j+1,k}} p_{s_{i,k}}(v) dv \cdot \log \frac{\int_{d_{j,k}}^{d_{j+1,k}} p_{s_{i,k}}(v) dv}{\sum_{i=0}^3 P(s_i) \int_{d_{j,k}}^{d_{j+1,k}} p_{s_{i,k}}(v) dv}, \quad k = 1, 2, \dots, K.
 \end{aligned} \tag{7}$$

The sum of MI for all layers is given by

$$\begin{aligned}
 \sum_{k=1}^K I(\mathbf{S}_k; \mathbf{R}_k) &= \sum_{k=1}^K [H(\mathbf{R}_k) - H(\mathbf{R}_k | \mathbf{S}_k)] \\
 &= \sum_{k=1}^K [H(\mathbf{R}_k)] - \sum_{k=1}^K [H(\mathbf{R}_k | \mathbf{S}_k)].
 \end{aligned} \tag{8}$$

The difference between Equations (6) and (8) is

$$I(\mathcal{S}; \mathcal{R}) - \sum_{k=1}^K I(\mathbf{S}_k; \mathbf{R}_k) = H(\mathcal{R}) - \sum_{k=1}^K H(\mathbf{R}_k). \tag{9}$$

According to the definition of entropy, we have

$$H(\mathcal{R}) = H(\mathbf{R}_1 \mathbf{R}_2 \dots \mathbf{R}_K) \leq \sum_{k=1}^K H(\mathbf{R}_k). \tag{10}$$

As a result,

$$I(\mathcal{S}; \mathcal{R}) \leq \sum_{k=1}^K I(\mathbf{S}_k; \mathbf{R}_k). \tag{11}$$

For 3D NAND flash memory, the channel of each layer is a discrete memoryless channel. However, these channels are correlated due to the presence of the cell-to-cell interference. As mentioned in [11,35], the cell-to-cell interference can be mitigated by pre-distortion/post-compensation technique. Therefore, we assume that the channel of each layer is independent of each other. Hence, the equality sign of Equation (11) holds [38], yields

$$I(\mathcal{S}; \mathcal{R}) = \sum_{k=1}^K I(\mathbf{S}_k; \mathbf{R}_k). \tag{12}$$

The optimal read-voltage thresholds can be derived by MMI in Equation (12). To jointly optimize the read-voltage thresholds for all layers, the optimization problem can be formulated as

$$\begin{aligned}
 \mathcal{P} : \quad & \max \quad \sum_{k=1}^K I(\mathbf{S}_k; \mathbf{R}_k) \\
 \text{s.t.} \quad & 0 < d_1 < d_2 < \dots < d_J < \infty.
 \end{aligned} \tag{13}$$

Similarly, to optimize the read-voltage thresholds for each layer, this optimization problem can be formulated as

$$\begin{aligned} \mathcal{P}: \quad & \max I(\mathbf{S}_k; \mathbf{R}_k) \\ \text{s.t.} \quad & 0 < d_{1,k} < d_{2,k} < \cdots < d_{J,k} < \infty, \\ & k = 1, 2, \dots, K. \end{aligned} \quad (14)$$

3.2. MMI Quantization Design

DP is an effective method to solve the optimization of multi-stage decision-making process. Specifically, the problems in Equations (13) and (14) can be decomposed into several smaller local problems, and then be solved in turn according to the recurrence relationship of the local problems to reach the global optimum [39]. In this subsection, we propose a MMI-DP-based algorithm in Algorithm 1 to optimize the read-voltage thresholds for 3D MLC flash memory channel.

First, the flash memory channel is uniformly quantized into N intervals and the boundaries are $\{a_0, a_1, \dots, a_N\}$, where $J \ll N$ (e.g., $N = 1000$). We set $a_0 = -\infty$, $a_1 = V_{s0} - 5 \times \sigma_{s0}$, $a_{N-1} = V_{s3} + 5 \times \sigma_{s3}$, and $a_N = \infty$. Then, other boundaries can be obtained by $a_n = a_1 + (n-1)(a_{N-1} - a_1)/(N-2)$. Second, the MI in Equation (7) can also be written as $I(\mathbf{S}_k; \mathbf{R}_k) = H(\mathbf{S}_k) - H(\mathbf{S}_k | \mathbf{R}_k)$, where $H(\mathbf{S}_k)$ is a constant. Therefore, the maximization of $I(\mathbf{S}_k; \mathbf{R}_k)$ is equivalent to minimizing $H(\mathbf{S}_k | \mathbf{R}_k)$. Hence, the optimization problems Equations (13) and (14) can be reformulated as

$$\begin{aligned} \mathcal{P}: \quad & \min \sum_{k=k_0}^{k_1} H(\mathbf{S}_k | \mathbf{R}_k) \\ \text{s.t.} \quad & 0 < d_1 < d_2 < \cdots < d_J < \infty, \end{aligned} \quad (15)$$

where k_0 and k_1 denote the layer index of 3D flash memory, respectively. To jointly optimized thresholds for all layers, we set $k_0 = 1$, $k_1 = K$. To optimize thresholds for each layer, we set $k_0 = k_i$, $k_1 = k_i$, where k_i denotes the target layer index. The optimal read-voltage thresholds can be obtained by:

$$\mathcal{D}^* = \{d_1^*, \dots, d_J^*\} = \arg \min_{\{d_1, \dots, d_J\} \subset \{a_0, \dots, a_N\}} \sum_{k=k_0}^{k_1} H(\mathbf{S}_k | \mathbf{R}_k). \quad (16)$$

The conditional entropy of k -th layer $H(\mathbf{S}_k | \mathbf{R}_k)$ can be calculated by

$$H(\mathbf{S}_k | \mathbf{R}_k) = \sum_{j=0}^J \Delta(d_j, d_{j+1}), \quad (17)$$

where

$$\Delta(d_j, d_{j+1}) = \sum_{i=0}^3 P(s_i) P(r_{j,k} | s_i) \log \frac{\sum_{i=0}^3 P(r_{j,k} | s_i) P(s_i)}{P(r_{j,k} | s_i) P(s_i)}. \quad (18)$$

The MMI quantizer is a sequential deterministic quantizer [39]. We use $C(N, J+1)$ to denote the cost function of quantizing $\{a_0, a_1, \dots, a_N\}$ into $J+1$ levels. Let $C^*(N, J+1)$ denote the cost function of the optimal solution in Equation (15). For a sequential deterministic quantizer, it can be decomposed into several smaller quantizers. Then, we have

$$\begin{aligned} C^*(N, J+1) &= \sum_{k=k_0}^{k_1} \sum_{j=0}^J \Delta(d_j^*, d_{j+1}^*) \\ &= C^*(\lambda_J^*, J) + \sum_{k=k_0}^{k_1} \Delta(a_{\lambda_J^*}, a_N) \\ &= \min_{J \leq \lambda_J \leq N} C^*(\lambda_J, J) + \sum_{k=k_0}^{k_1} \Delta(a_{\lambda_J}, a_N), \end{aligned} \quad (19)$$

where $\{\lambda_1^*, \lambda_2^*, \dots, \lambda_J^*\} \subset \{1, 2, \dots, N-1\}$ and $\{a_{\lambda_1^*}, a_{\lambda_2^*}, \dots, a_{\lambda_J^*}\}$ are the optimal solutions of each smaller quantizer. From Equation (19), the optimal solution \mathcal{D}^* can be obtained by solving sub-problems in a recursive manner. The complexity of solving these problems by Algorithm 1 is $\mathcal{O}((N-J)^2J)$ [39].

Algorithm 1 MMI-DP algorithm for searching optimal read-voltage thresholds in 3D flash memory.

Input: $J, N, K, a_1, a_{N-1}, k_0, k_1$.

Output: $\mathcal{D}^* = \{d_1^*, \dots, d_J^*\}$.

```

1: for  $n = 2; n < N; n++$  do
2:    $a_n = a_1 + (n-1)(a_{N-1} - a_1)/(N-2)$ ;
3: end for
4:  $a_0 = -\infty, a_N = \infty$ ;
5: for  $n = 1; n \leq N; n++$  do
6:    $C(n, 1) = \sum_{k=k_0}^{k_1} \Delta(a_0, a_n)$ ;
7: end for
8: for  $q = 2; q \leq J+1; q++$  do
9:   for  $m = q; m \leq N-J+q-1; m++$  do
10:     $C(m, q) = \infty$ ;
11:    for  $t = q-1; t < m; t++$  do
12:      if  $C(m, q) > C(t, q-1) + \sum_{k=k_0}^{k_1} \Delta(a_t, a_m)$  then
13:         $C(m, q) = C(t, q-1) + \sum_{k=k_0}^{k_1} \Delta(a_t, a_m)$ ;
14:         $\lambda(m, q) = t$ ;
15:      end if
16:    end for
17:  end for
18: end for
19:  $t = N$ ;
20: for  $i = J+1; i > 1; i--$  do
21:    $t = \lambda(t, i), d_{i-1}^* = a_t$ ;
22: end for
23: Return  $\mathcal{D}^* = \{d_1^*, \dots, d_J^*\}$ .
```

To evaluate the performance of our proposed joint quantization design, the performance of the thresholds optimized for each layer by MMI is served as the benchmark. Note that this quantization design method requires a number of read operations which will result in a large read latency.

To examine the performance of Algorithm 1, the MI of 3D flash memory channel with different quantization schemes are presented in Figure 3. Note that the MMI-DP optimized for the first layer is to apply read-voltage thresholds optimized for the first layer to all layers. As illustrated in Figure 3, all MMI-DP based methods can approach the MI of 3D flash memory channel (with quantization level $J = \infty$) by using only nine read-voltage thresholds. Moreover, the MI of our proposed joint optimization algorithm is higher than that optimized for the first layer, and close to the optimum (i.e., optimized for each layer). The impact of layer-to-layer variation on the voltage distribution of s_0 results in a higher occurrence of errors in the upper layers [9]. The proposed joint optimization algorithm is capable of designing voltage thresholds to effectively mitigate these errors.

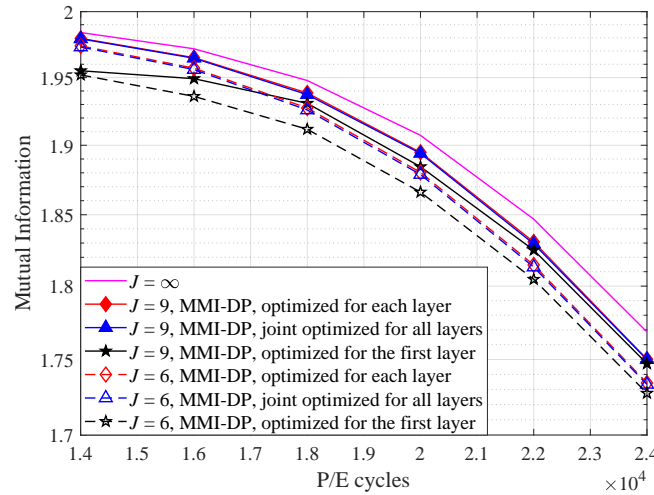


Figure 3. The mutual information of quantized 3D flash memory channel with the MMI quantizer at $t = 10^4$ s.

3.3. Read Thresholds Design for Hard Decision Decoding

In practical applications, to reduce the read latency and power consumption, the HDD of ECCs needs to be performed once before soft-decision decoding. Therefore, it is of great significance to design read-voltage thresholds for HDD of ECCs. Although the proposed DP algorithm can also design read-voltage thresholds for HDD, its computational complexity is high. To solve this problem, we proposed a MID-based quantization scheme to further reduce the complexity.

As described in Section 3. A, by MMI in Equation (12), we can optimize the read-voltage thresholds for flash memory. Since Equation (13) is locally concave, we can optimize the thresholds for HDD by calculating the derivative of the MI, given by

$$\frac{dI(\mathcal{S}; \mathcal{R})}{dh_j} = 0, \quad (20)$$

where $j = 1, 2, 3$, and h_j denotes the read-voltage threshold for HDD of ECCs.

$$\begin{aligned} I(\mathbf{S}_k; \mathbf{R}_k) &= \sum_{j=1}^J \sum_{i=1}^i \frac{1}{2} \int_{h_j}^{h_{j+1}} p_{s_{i,k}}(v) dv \log \frac{\int_{h_j}^{h_{j+1}} p_{s_{i,k}}(v) dv}{\sum_{i=i-1}^i \frac{1}{2} \int_{h_j}^{h_{j+1}} p_{s_{i,k}}(v) dv} \\ &= \frac{1}{2} \int_{-\infty}^{h_j} p_{s_{i-1,k}}(v) dv \log \frac{\int_{-\infty}^{h_j} p_{s_{i-1,k}}(v) dv}{\frac{1}{2} \left(\int_{-\infty}^{h_j} p_{s_{i-1,k}}(v) + p_{s_{i,k}}(v) dv \right)} + \\ &\quad \frac{1}{2} \int_{h_j}^{\infty} p_{s_{i-1,k}}(v) dv \log \frac{\int_{h_j}^{\infty} p_{s_{i-1,k}}(v) dv}{\frac{1}{2} \left(\int_{h_j}^{\infty} p_{s_{i-1,k}}(v) + p_{s_{i,k}}(v) dv \right)} + \\ &\quad \frac{1}{2} \int_{-\infty}^{h_j} p_{s_{i,k}}(v) dv \log \frac{\int_{-\infty}^{h_j} p_{s_{i,k}}(v) dv}{\frac{1}{2} \left(\int_{-\infty}^{h_j} p_{s_{i-1,k}}(v) + p_{s_{i,k}}(v) dv \right)} + \\ &\quad \frac{1}{2} \int_{h_j}^{\infty} p_{s_{i,k}}(v) dv \log \frac{\int_{h_j}^{\infty} p_{s_{i,k}}(v) dv}{\frac{1}{2} \left(\int_{h_j}^{\infty} p_{s_{i-1,k}}(v) + p_{s_{i,k}}(v) dv \right)}, \\ &j = 1, 2, 3, \quad i = 1, 2, 3. \end{aligned} \quad (21)$$

However, it is difficult to derive an analytical solution for Equation (20). We propose a MID-based method to optimize read-voltage thresholds by dividing the 2-bits quantization into three 1-bit quantization problems. As shown in Figure 4, each 1-bit quantization mainly depends on the adjacent state of flash memory, i.e., the red and blue region. With this simplification, the flash memory channel can be regarded as three binary asymmetric channels and the optimization problem can be solved by finding the root of the following equation:

$$\frac{d \sum_{k=k_0}^{k_1} I(\mathbf{S}_k; \mathbf{R}_k)}{dh_j} = 0, \quad j = 1, 2, 3. \quad (22)$$

According to Equation (7), the MI between two adjacent states in k -th layer can be calculated by

$$I(\mathbf{S}_k; \mathbf{R}_k) = \sum_{j=j-1}^j \sum_{i=i-1}^i P(s_i) P(r_{j,k} | s_i) \log \frac{P(r_{j,k} | s_i)}{\sum_{i=i-1}^i P(r_{j,k} | s_i) P(s_i)}, \quad i = 1, 2, 3, \quad j = 1, 2, 3, \quad (23)$$

where $P(s_i) = \frac{1}{2}$.

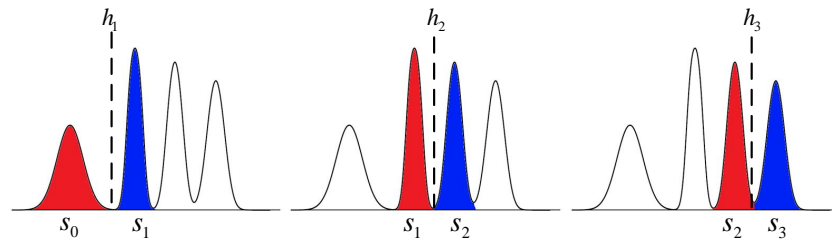


Figure 4. Simplified model of MLC flash memory.

By expanding Equation (23) to Equation (21) and substitute it into Equation (22), we can calculate the solution of Equation (22). Let h_j^* denote the solution of Equation (22). To determine h_j^* , we first compute the derivative of Equation (21) with respect to h_j . The detailed derivative is given in Appendix A. By using the proposed MID-based quantization, we can obtain the solution of Equation (22) which is a suboptimal solution of Equation (20). To solve Equation (22), the root can be found by employing binary search method. Hence, the complexity of the MID-based quantization is $\mathcal{O}(\log n)$, where n is the number of samples, and $n = 1024$ has been found to be sufficient. Therefore, the computational complexity of the MID-based quantization is significantly lower than the MMI-DP algorithm which is $\mathcal{O}((N - J)^2 J)$.

To evaluate the performance of thresholds designed by MID, the symbol error probability (SEP) of 3D MLC flash memory with full channel knowledge is further derived. For given $\{h_1, h_2, h_3\}$, the SEP of an uncoded 3D flash memory channel can be calculated by

$$P_e = \sum_{k=k_0}^{k_1} \frac{P_e(s_0, k) + P_e(s_1, k) + P_e(s_2, k) + P_e(s_3, k)}{4(k_1 - k_0 + 1)}, \quad (24)$$

where

$$P_e(s_0, k) = \int_{h_1}^{\infty} p_{s_0, k}(v) dv, P_e(s_1, k) = \int_{-\infty}^{h_1} p_{s_1, k}(v) dv + \int_{h_2}^{\infty} p_{s_1, k}(v) dv, \\ P_e(s_2, k) = \int_{-\infty}^{h_2} p_{s_2, k}(v) dv + \int_{h_3}^{\infty} p_{s_2, k}(v) dv, P_e(s_3, k) = \int_{-\infty}^{h_3} p_{s_3, k}(v) dv.$$

To evaluate the performance of our proposed MID scheme, the thresholds optimized by minimizing SEP (MSEP) is served as a benchmark. To find the thresholds that minimize the SEP, a cross iterative searching algorithm is adopted by utilizing genetic and iterative searching algorithm [24]. Through plenty of iterations, the cross iterative searching algorithm can approach the global optimum for given number of quantization levels. Therefore, it can serve as the benchmark of our proposed low-complexity MID-based quantization design.

Figure 5 compares the MI of 3D flash memory channel with the MSEP, MMI-DP, and MID quantizers. The MI of the unquantized channel is also included as a reference. First, as shown in Figure 5, the MI of MMI-DP quantizer and MSEP quantizer are almost the same, which indicates the DP algorithm has reached the global optimum. Second, the MI of the MID quantizer close approaches to that of the MMI-DP quantizer, which demonstrates the effectiveness of the proposed MID quantizer. Finally, the MI of joint optimization scheme for 3D flash memory approaches that of optimizing for each layer in 3D flash memory, which verifies that the joint optimization scheme can achieve near-optimal performance for 3D flash memory.

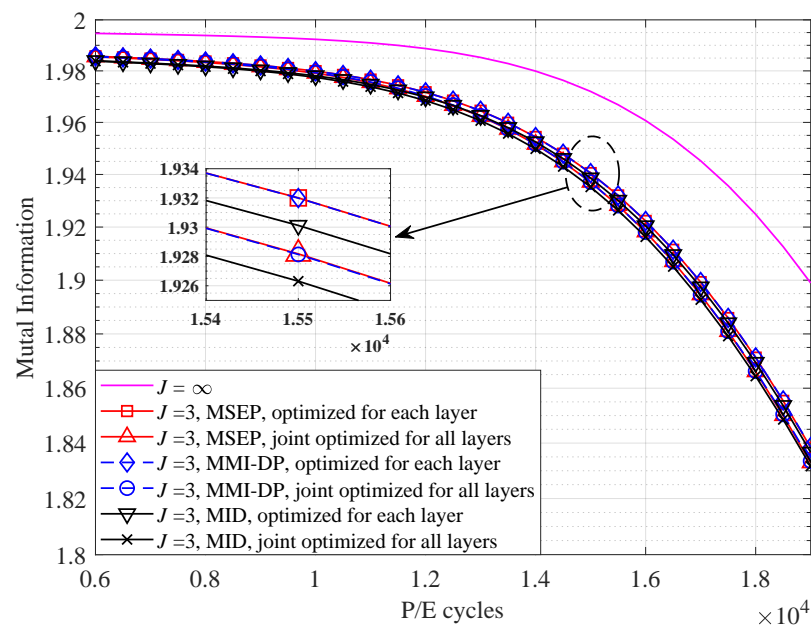


Figure 5. The MI of the MSEP, MMI-DP, and MID quantizers with 3 read-voltage thresholds over different P/E cycles at $t = 5 \times 10^6$ s.

4. Numerical and Simulation Results

In the simulations, the total number of layer in 3D flash memory is set to 30. We first investigate the uncoded SEP performance with different quantization schemes over different P/E cycles at $t = 5 \times 10^6$ s. The quantization scheme that directly minimizes the SEP is served as a benchmark. With read-voltage thresholds optimized by these quantizers, the SEP of 3D flash memory channel can be calculated by Equation (24). As shown in Figure 6, the SEP of these quantization schemes is consistent with their MI in Figure 5. It is observed that the SEP of the MID quantization is close to that of the MMI-DP quantization. In addition, the quantization scheme of joint optimized for all layers shows superior performance when compared with the first layer optimization quantization scheme.

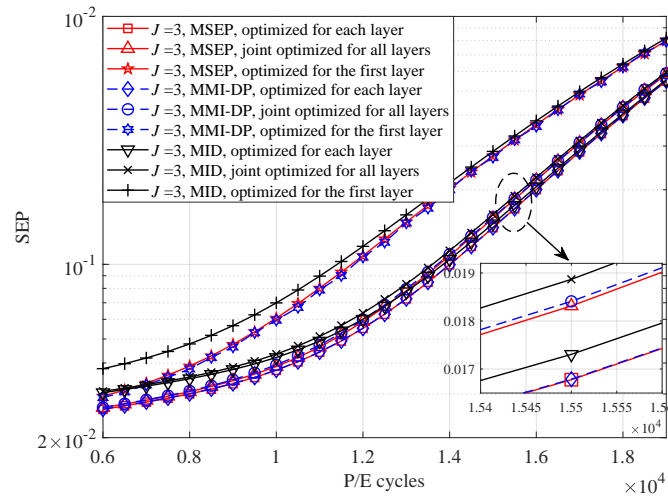


Figure 6. The SEP of the MSEP, MMI-DP, and MID quantizers with 3 read-voltage thresholds over different P/E cycles at $t = 5 \times 10^6$ s.

Next, the LDPC-coded frame error rate (FER) performance of different quantization scheme is examined. The decoding algorithm of LDPC codes is sum-product algorithm with maximum 25 iterations ($I_{\max} = 25$). In our simulations, a 4K-code is employed and constructed by progressive-edge-growth algorithm [21]. The code length for 4K-code is 4544 bits with code rate 0.9. The degree distribution of this code is given as

$$\epsilon(x) = 0.0682x + 0.1822x^2 + 0.1329x^3 + 0.6167x^4,$$

$$\theta(x) = 0.22x^{38} + 0.78x^{39},$$

where $\epsilon(x)$ and $\theta(x)$ are the variable-node and check-node degree distribution optimized by density-evolution, respectively.

First, the LDPC-coded performance of different quantizers with 3-level quantization is presented in Figure 7. The FER of the MSEP quantizer is also included as the benchmark. Similarly, it can be seen that, the FER performance of joint optimization for all layers by MID scheme is superior to that of optimization for the first layer and close to the optimal performance. For example, at FER = 10^{-4} , the proposed quantizer improves the endurance of 3D flash memory by 2000 P/E cycles compared with the quantizer that optimized for the first layer.

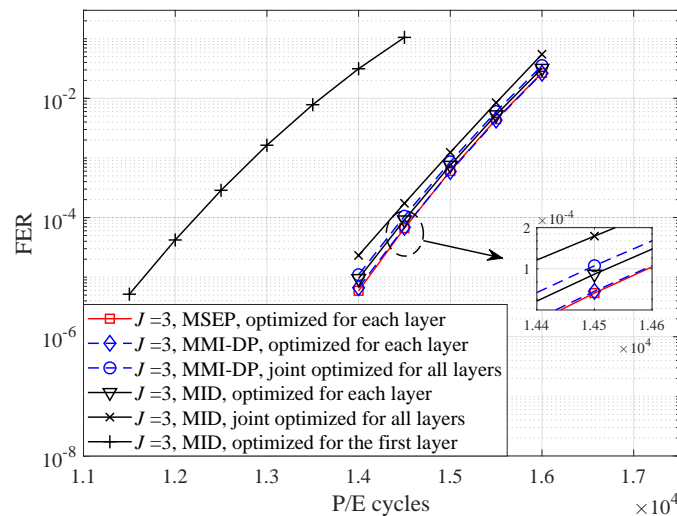


Figure 7. FER performance of LDPC 4K-code over different P/E cycles with 3 read-voltage thresholds, $I_{\max} = 25$ and $t = 5 \times 10^6$ s.

To enable soft-decision decoding, more read-voltage thresholds are essential. Figure 8 illustrates the FER performance of $4K$ -code with six read-voltage thresholds. In addition, the performance of the uniform quantizer is also included as a reference. Note that the uniform quantization has 20 read-voltage thresholds and our designed non-uniform quantizers has 6 read-voltage thresholds. It is observed that the FER performance of joint optimized thresholds is much better than that of optimized for the first layer and close to the optimum (i.e., the thresholds optimized for each layer). The FER performance is consistent with the MI in Figure 3, which all demonstrate the superiority of our proposed algorithm. For example, at $\text{FER} = 10^{-4}$, the proposed algorithm improves the endurance of 3D flash memory by 3100 P/E cycles compared with that optimized for the first layer. Compared with uniform quantization, the proposed algorithm only needs 6 read-voltage thresholds to surpass the performance of 20 uniform thresholds. This demonstrates that non-uniform read-voltage design reduces the number of read operations while still maintaining desirable error-correction performance.

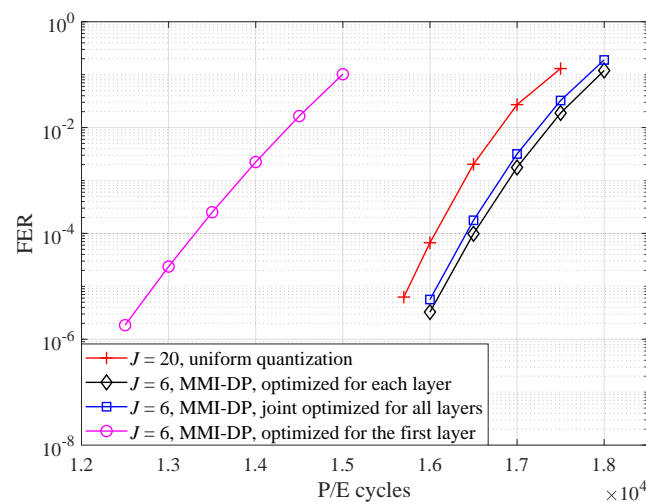


Figure 8. FER performance of LDPC $4K$ -code over different P/E cycles with $I_{\max} = 25$ and $t = 5 \times 10^6$ s.

5. Conclusions

In this paper, we have derived the channel model for 3D MLC flash memory based on the experimental data. Next, the 3D MLC flash memory with K layers is regarded as a joint channel and the channel capacity has been derived. By maximizing the MI of 3D flash memory channel, we have further proposed a MMI-DP algorithm to optimize read-voltage thresholds. In addition, to reduce the complexity of the MMI-DP algorithm, we have simplified the 3D MLC flash memory channel model and proposed a MID-based quantization scheme to obtain read-voltage thresholds for ECCs with HDD. Simulation results have shown that the FER performance of our proposed joint optimization algorithm can almost achieve the performance that is optimized for each layer with much less read-voltage thresholds, such that the read latency can be significantly reduced.

Author Contributions: Conceptualization, C.W., Z.M. and J.L.; methodology, C.W., Z.M. and X.H.; software, C.W.; validation, C.W., Z.M. and X.H.; formal analysis, C.W.; investigation, C.W.; resources, C.W.; data curation, C.W.; writing—original draft preparation, C.W. and Z.M.; writing—review and editing, C.W., Z.M., J.L. and X.H.; visualization, C.W.; supervision, J.L., L.K. and F.S.; project administration, Z.M. and J.L.; funding acquisition, Z.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 62201258, by the open research fund of National Mobile Communications Research Laboratory, Southeast University (No. 2023D12).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivation of Equation (23)

For MLC flash memory, the HDD of ECCs requires three read-voltage thresholds. Hence, Equation (13) can be divided into three parts. As shown in Figure 4, for each part, the voltages are quantized into two regions, i.e., the red and green region. First, for h_1^* between s_0 and s_1 , it can be obtained by calculating the root of Equation (22).

To simplify the derivative, we decompose the elements of Equation (21) into several parts. We define

$$\begin{aligned} Q_0 &= \int_{-\infty}^{h_1} p_{s_{0,k}}(v)dv; 1 - Q_0 = \int_{h_1}^{\infty} p_{s_{0,k}}(v)dv; \\ Q_1 &= \int_{-\infty}^{h_1} p_{s_{1,k}}(v)dv; 1 - Q_1 = \int_{h_1}^{\infty} p_{s_{1,k}}(v)dv. \end{aligned} \quad (A1)$$

As $p_{s_{0,k}}$ follows Gaussian distribution, Q_0 can be expressed as

$$Q_0 = 1 - Q\left(\frac{h_1 - \mu_{s_{0,k}}}{\sigma_{s_{0,k}}}\right), \quad (A2)$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{t^2}{2}\right) dt. \quad (A3)$$

As $p_{s_{1,k}}$ includes Gaussian error function, Q_1 can be expressed as:

$$\begin{aligned} Q_1 &= \frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{2V_p} \operatorname{erf}\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{\sqrt{2}\sigma_{s_{1,k}}}\right) + 0.5 - \\ &\quad \frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{2V_p} \operatorname{erf}\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{\sqrt{2}\sigma_{s_{1,k}}}\right) + \\ &\quad \frac{\sigma_{s_{1,k}}}{\sqrt{2\pi}V_p} \left(e^{-\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{\sqrt{2}\sigma_{s_{1,k}}}\right)^2} - e^{-\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{\sqrt{2}\sigma_{s_{1,k}}}\right)^2} \right). \end{aligned} \quad (A4)$$

Substituting Equation (A1) into Equation (21), we have Equation (A5). The derivative of each term in Equation (A5) is given as Equations (A6)–(A9).

$$\begin{aligned} I(\mathbf{S}_k; \mathbf{R}_k) &= \frac{1}{2} Q_0 \left(\log Q_0 - \log \left(\frac{1}{2} Q_0 + \frac{1}{2} Q_1 \right) \right) + \frac{1}{2} Q_1 \left(\log Q_1 - \log \left(\frac{1}{2} Q_0 + \frac{1}{2} Q_1 \right) \right) + \\ &\quad \frac{1}{2} (1 - Q_0) \left(\log(1 - Q_0) - \log \left(\frac{1}{2} (1 - Q_0) + \frac{1}{2} (1 - Q_1) \right) \right) + \\ &\quad \frac{1}{2} (1 - Q_1) \left(\log(1 - Q_1) - \log \left(\frac{1}{2} (1 - Q_0) + \frac{1}{2} (1 - Q_1) \right) \right) \\ &= \frac{1}{2} \sum_{i=0}^1 (Q_i (\log Q_i - \log(\frac{Q_0 + Q_1}{2})) + (1 - Q_i) (\log(1 - Q_i) - \log(\frac{2 - Q_0 - Q_1}{2}))). \end{aligned} \quad (A5)$$

$$\frac{dQ_0}{dh_1} = \frac{d1 - Q\left(\frac{h_1 - \mu_{s_{0,k}}}{\sigma_{s_{0,k}}}\right)}{dh_1} = \frac{1}{\sqrt{2\pi}\sigma_{s_{0,k}}} e^{-\frac{(h_1 - \mu_{s_{0,k}})^2}{2\sigma_{s_{0,k}}^2}}, \quad (A6)$$

$$\frac{d \log Q_0}{dh_1} = \frac{1}{\sqrt{2\pi}\sigma_{s_{0,k}}} e^{-\frac{(h_1 - \mu_{s_{0,k}})^2}{2\sigma_{s_{0,k}}^2}} \frac{1}{1 - Q\left(\frac{h_1 - \mu_{s_{0,k}}}{\sigma_{s_{0,k}}}\right)}, \quad (A7)$$

$$\begin{aligned} \frac{dQ_1}{dh_1} = & \frac{\operatorname{erf}\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{\sqrt{2}\sigma_{s_{1,k}}}\right)}{\sqrt{2}\sigma_{s_{1,k}} V_p} - \frac{\sqrt{2}(h_1 + \mu_{s_{1,k}} - V_{s_1})}{\sqrt{\pi}\sigma_{s_{1,k}}} e^{-\frac{(h_1 + \mu_{s_{1,k}} - V_{s_1})^2}{2\sigma_{s_{1,k}}^2}} - \\ & \frac{\operatorname{erf}\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{\sqrt{2}\sigma_{s_{1,k}}}\right)}{\sqrt{2}\sigma_{s_{1,k}} V_p} + \frac{\sqrt{2}(h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p)}{\sqrt{\pi}\sigma_{s_{1,k}}} e^{-\frac{(h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p)^2}{2\sigma_{s_{1,k}}^2}} + \\ & \frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{\sqrt{\pi}\sigma_{s_{1,k}}^2 V_p} e^{-\frac{(h_1 + \mu_{s_{1,k}} - V_{s_1})^2}{2\sigma_{s_{1,k}}^2}} - \frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{\sqrt{\pi}\sigma_{s_{1,k}}^2 V_p} e^{-\frac{(h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p)^2}{2\sigma_{s_{1,k}}^2}}, \end{aligned} \quad (A8)$$

$$\begin{aligned} \frac{d \log Q_1}{dh_1} &= \frac{dQ_1}{dh_1} / Q_1 \\ &= \frac{dQ_1}{dh_1} / \left[\frac{\frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{2V_p} \operatorname{erf}\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{\sqrt{2}\sigma_{s_{1,k}}}\right) - \frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{2V_p} \operatorname{erf}\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{\sqrt{2}\sigma_{s_{1,k}}}\right)}{\frac{\sigma_{s_{1,k}}}{\sqrt{2\pi}V_p} \left(e^{-\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1}}{\sqrt{2}\sigma_{s_{1,k}}}\right)^2} - e^{-\left(\frac{h_1 + \mu_{s_{1,k}} - V_{s_1} - V_p}{\sqrt{2}\sigma_{s_{1,k}}}\right)^2} \right) + 0.5} \right]. \end{aligned} \quad (A9)$$

With these derivatives, we can easily compute the derivative of the MI. By solving $d \sum_{k=0}^{K-1} I(\mathbf{S}_k; \mathbf{R}_k) / dh_1 = 0$, we can obtain the threshold h_1^* . Similarly, we can obtain h_2^* and h_3^* .

References

- Kim, K. Future memory technology: Challenges and opportunities. In Proceedings of the International Symposium on VLSI Technology, Systems and Applications, Hsinchu, Taiwan, 21–23 April 2008; pp. 5–9.
- Dong, G.; Xie, N.; Zhang, T. On the use of soft-decision error-correction codes in NAND flash memory. *IEEE Trans. Circuits Syst. I Reg. Pap.* **2011**, *58*, 429–439. [\[CrossRef\]](#)
- Li, Q.; Jiang, Q.; Haratsch, E.F. Noise modeling and capacity analysis for NAND flash memories. In Proceedings of the IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2262–2266.
- Luo, Y.; Ghose, S.; Cai, Y.; Haratsch, E.F.; Mutlu, O. HeatWatch: Improving 3D NAND Flash Memory Device Reliability by Exploiting Self-Recovery and Temperature Awareness. In Proceedings of the IEEE International Symposium on High Performance Computer Architecture, Vienna, Austria, 24–28 February 2018; pp. 504–517.
- Wang, Y.; Tan, J.; Mao, R.; Li, T. Temperature-aware persistent data management for LSM-tree on 3-D NAND flash memory. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2020**, *39*, 4611–4622. [\[CrossRef\]](#)
- Wang, Y.; Huang, J.; Chen, J.; Mao, R. PVSensing: A Process-Variation-Aware Space Allocation Strategy for 3D NAND Flash Memory. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2021**, *42*, 1302–1315. [\[CrossRef\]](#)
- Wang, Y.; Dong, L.; Mao, R. P-Alloc: Process-variation tolerant reliability management for 3D charge-trapping flash memory. *ACM Trans. Embed. Comput. Syst.* **2017**, *16*, 1–19. [\[CrossRef\]](#)
- Michelsoni, R. *3D Flash Memories*; Springer: Dordrecht, The Netherlands, 2016.
- Luo, Y.; Ghose, S.; Cai, Y.; Haratsch, E.F.; Mutlu, O. Improving 3D NAND flash memory lifetime by tolerating early retention loss and process variation. *Proc. ACM Meas. Anal. Comput. Syst.* **2018**, *2*, 1–48. [\[CrossRef\]](#)
- Xu, Q.; Gong, P.; Chen, T.; Michael, J.; Li, S. Modelling and characterization of NAND flash memory channels. *Measurement* **2015**, *70*, 225–231. [\[CrossRef\]](#)
- Wang, K.; Du, G.; Lun, Z.; Chen, W.; Liu, X. Modeling of program Vth distribution for 3-D TLC NAND flash memory. *Sci. China Inf. Sci.* **2019**, *62*, 1–10. [\[CrossRef\]](#)
- Shim, Y.; Kim, M.; Chun, M.; Park, J.; Kim, Y.; Kim, J. Exploiting process similarity of 3D flash memory for high performance SSDs. In Proceedings of the Annual IEEE/ACM International Symposium on Microarchitecture, New York, NY, USA, 12–16 October 2019; pp. 211–223.
- Papandreou, N.; Pozidis, H.; Parnell, T.; Ioannou, N.; Pletka, R.; Tomic, S.; Breen, P.; Tressler, G.; Fry, A.; Fisher, T. Characterization and analysis of bit errors in 3D TLC NAND flash memory. In Proceedings of the IEEE International Reliability Physics Symposium, Monterey, CA, USA, 31 March–4 April 2019; pp. 1–6.

14. Choi, B.; Jang, S.H.; Yoon, J.; Lee, J.; Jeon, M.; Lee, Y.; Han, J.; Lee, J.; Kim, D.M.; Kim, D.H.; et al. Comprehensive evaluation of early retention (fast charge loss within a few seconds) characteristics in tube-type 3-D NAND flash memory. In Proceedings of the IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 14–16 June 2016; pp. 1–2.
15. Gallager, R. Low-density parity-check codes. *IRE Trans. Inf. Theory* **1962**, *8*, 21–28. [\[CrossRef\]](#)
16. Du, Y.; Zhou, Y.; Zhang, M.; Liu, W.; Xiong, S. Adapting layer RBERs variations of 3D flash memories via multi-granularity progressive LDPC reading. In Proceedings of the Annual Design Automation Conference, Las Vegas, NV, USA, 2–6 June 2019; pp. 1–6.
17. Zhang, M.; Wu, F.; Yu, Q.; Liu, W.; Cui, L.; Zhao, Y.; Xie, C. BeLDPC: Bit errors aware adaptive rate LDPC codes for 3D TLC NAND flash memory. In Proceedings of the Design, Automation and Test in Europe Conference and Exhibition, Grenoble, France, 9–13 March 2020; pp. 302–305.
18. Wang, J.; Vakili, K.; Chen, T.; Courtade, T.; Dong, G.; Zhang, T. Enhanced precision through multiple reads for LDPC decoding in flash memories. *IEEE J. Sel. Areas. Commun.* **2014**, *32*, 880–891. [\[CrossRef\]](#)
19. Ma, R.; Wu, F.; Zhang, M.; Lu, Z.; Wan, J.; Xie, C. RBER-aware lifetime prediction scheme for 3D-TLC NAND flash memory. *IEEE Access* **2019**, *7*, 44696–44708. [\[CrossRef\]](#)
20. Yu, D.; Hsieh, J. Differential Evolution Algorithm with Asymmetric Coding for Solving the Reliability Problem of 3D-TLC CT Flash-Memory Storage Systems. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2021**, *41*, 2863–2876. [\[CrossRef\]](#)
21. Aslam, C.; Guan, Y.; Cai, K. Read and write voltage signal optimization for multi-level-cell (MLC) NAND flash memory. *IEEE Trans. Comm.* **2016**, *64*, 1613–1623. [\[CrossRef\]](#)
22. Yu, X.; He, J.; Li, Q.; Zhang, B.; Wang, X.; Yang, L.; Huo, Z. LIAD: A Method for Extending the Effective Time of 3D TLC NAND Flash Hard Decision. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2022**, *42*, 1705–1717. [\[CrossRef\]](#)
23. Peleato, B.; Agarwal, R.; Cioffi, J.; Qin, M.; Siegel, P. Adaptive read thresholds for NAND flash. *IEEE Trans. Commun.* **2015**, *63*, 3069–3081. [\[CrossRef\]](#)
24. Wang, C.; Wei, K.; Kong, L.; Shi, L.; Mei, Z.; Li, J.; Cai, K. DNN-aided read-voltage threshold optimization for MLC flash memory with finite block length. *IET Commun.* **2022**, *16*, 120–130. [\[CrossRef\]](#)
25. Mei, Z.; Cai, K.; Shi, L. Information theoretic bounds based channel quantization design for emerging memories. In Proceedings of the IEEE Information Theory Workshop, Guangzhou, China, 25–29 November 2018; pp. 1–5.
26. Li, Q.; Ye, M.; Cui, Y.; Shi, L.; Li, X.; Kuo, T. W.; Xue, C.J. Shaving retries with sentinels for fast read over high-density 3D flash. In Proceedings of the Annual IEEE/ACM International Symposium on Microarchitecture, Athens, Greece, 9–13 March 2020; pp. 483–495.
27. Li, Q.; Shi, L.; Cui, Y.; Xue, C. Exploiting asymmetric errors for LDPC decoding optimization on 3D NAND flash memory. *IEEE Trans. Comput.* **2020**, *69*, 475–488. [\[CrossRef\]](#)
28. Zhang, M.; Wu, F.; Yu, Q.; Liu, W.; Wang, Y.; Xie, C. Exploiting error characteristic to optimize read voltage for 3-D NAND flash memory. *IEEE Trans. Electron Devices* **2020**, *67*, 5490–5496. [\[CrossRef\]](#)
29. Park, S.; Moon, J. Characterization of Inter-Cell Interference in 3D NAND Flash Memory. *IEEE Trans. Circuits Syst. I Reg. Pap.* **2021**, *68*, 1183–1192. [\[CrossRef\]](#)
30. Liu, W.; Wu, F.; Zhou, J.; Zhang, M.; Yang, C.; Lu, Z.; Wang, Y.; Xie, C. Modeling of threshold voltage distribution in 3d nand flash memory. In Proceedings of the Design, Automation and Test in Europe Conference and Exhibition, Grenoble, France, 1–5 February 2021; pp. 1729–1732.
31. Bez, R.; Camerlenghi, E.; Modelli, A.; Visconti, A. Introduction to flash memory. *Proc. IEEE* **2003**, *91*, 489–502. [\[CrossRef\]](#)
32. Cai, Y.; Ghose, S.; Haratsch, E.; Luo, Y.; Mutlu, O. Error characterization, mitigation, and recovery in flash-memory-based solid-state drives. *Proc. IEEE* **2017**, *105*, 1666–1704. [\[CrossRef\]](#)
33. Suh, K.; Suh, B.; Lim, Y.; Kim, J.; Choi, Y.; Koh, Y.; Lee, S.; Kwon, S.; Choi, B.; Yum, J.; et al. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme. *IEEE J. Solid-State Circuits* **1995**, *30*, 1149–1156.
34. Cui, J.; Zeng, Z.; Huang, J.; Yuan, W.; Yang, L. Improving 3D NAND SSD read performance by parallelizing read-retry. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2022**, *42*, 768–780. [\[CrossRef\]](#)
35. Dong, G.; Li, S.; Zhang, T. Using data postcompensation and predistortion to tolerate cell-to-cell interference in MLC NAND flash memory. *IEEE Trans. Circuits Syst. I Reg. Pap.* **2010**, *57*, 2718–2728. [\[CrossRef\]](#)
36. Aslam, C.; Guan, Y.; Cai, K. Decision-directed retention-failure recovery with channel update for MLC NAND flash memory. *IEEE Trans. Circuits Syst. I Reg. Pap.* **2018**, *65*, 353–365. [\[CrossRef\]](#)
37. Dong, G.; Xie, N.; Zhang, T. Enabling NAND flash memory use soft-decision error correction codes at minimal read latency overhead. *IEEE Trans. Circuits Syst. I Reg. Pap.* **2013**, *60*, 2412–2421. [\[CrossRef\]](#)
38. Ash, R.B. *Information Theory*; Courier Corporation: Chicago, IL, USA, 2012.
39. He, X.; Cai, K.; Song, W.; Mei, Z. Dynamic programming for sequential deterministic quantization of discrete memoryless channels. *IEEE Trans. Commun.* **2021**, *69*, 3638–3651. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.