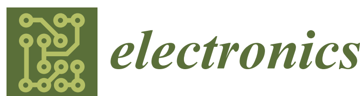

Article

Read Reference Calibration and Tracking for Non-Volatile Flash Memories

Johann-Philipp Thiers [1], Daniel Nicolas Bailon [1], Jürgen Freudenberger [1,*] and Jianjie Lu [2]

[1] Institute for System Dynamics (ISD), HTWG Konstanz, University of Applied Sciences, 78462 Konstanz, Germany; jthiers@htwg-konstanz.de (J.-P.T.); dnicolas@htwg-konstanz.de (D.N.B.)
[2] Hyperstone GmbH, 78467 Konstanz, Germany; jlu@hyperstone.com
* Correspondence: juergen.freudenberger@htwg-konstanz.de; Tel.: +49-7531-206-647

Abstract: The performance and reliability of nonvolatile NAND flash memories deteriorate as the number of program/erase cycles grows. The reliability also suffers from cell-to-cell interference, long data retention time, and read disturb. These processes effect the read threshold voltages. The aging of the cells causes voltage shifts which lead to high bit error rates (BER) with fixed predefined read thresholds. This work proposes two methods that aim on minimizing the BER by adjusting the read thresholds. Both methods utilize the number of errors detected in the codeword of an error correction code. It is demonstrated that the observed number of errors is a good measure for the voltage shifts and is utilized for the initial calibration of the read thresholds. The second approach is a gradual channel estimation method that utilizes the asymmetrical error probabilities for the one-to-zero and zero-to-one errors that are caused by threshold calibration errors. Both methods are investigated utilizing the mutual information between the optimal read voltage and the measured error values. Numerical results obtained from flash measurements show that these methods reduce the BER of NAND flash memories significantly.

Keywords: nonvolatile NAND flash; program/erase cycles; data retention time; channel estimation; threshold calibration

Citation: Thiers, J.-P.; Nicolas Bailon, D.; Freudenberger, J.; Lu, J. Read Reference Calibration and Tracking for Non-Volatile Flash Memories. *Electronics* **2021**, *10*, 2306. https://doi.org/10.3390/electronics10182306

Academic Editors: Gino Giusi and Changhwan Shin

Received: 29 July 2021
Accepted: 17 September 2021
Published: 19 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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

Nonvolatile flash memories are often applied for storage systems that require high data reliability, e.g., industrial robots, scientific and medical instruments. Flash memories are resistant to mechanical shock and provide fast read access times. The information is stored in floating gate transistors that hold the charge in the absence of power [1–3]. The cells can be programmed and erased. To distinguish the charge levels and read the stored information, several reference voltages are applied to the control gate. However, this reading operation can cause additional disturbances which lead to errors in stored information [4].

Due to the error characteristics of the flash memory, it is necessary to use error correction coding (ECC) to guarantee reliability and data integrity [1,4–6]. Traditionally, Bose–Chaudhuri–Hocquenghem (BCH) codes with so-called hard-input decoding were used for error correction [7,8]. The term hard-input refers to a reading procedure where only a single bit of information is read from each cell. However, the performance of error correction can be improved with soft-input decoding. Soft-input decoding algorithms utilize reliability information about the state of the cells, where the channel output is quantized using a small number of bits [9,10]. Low-density parity-check (LDPC) codes [11–15], generalized concatenated codes (GCC) [16–18], and polar codes [19,20] were proposed for flash memories with soft-input decoding.

The error probability of the flash cells depends on the flash memory technology (single-level cell (SLC), multilevel cell (MLC), triple-level cell (TLC), or quad-level cell (QLC)), the storage density, and the number of programming and erasing cycles. A lot of research

exists on the error characteristics considers the flash memory endurance [1,21–23], i.e., the number of program/erase (P/E) cycles. Over time, charge losses occur, which cause a change of the threshold voltage distribution [24]. In many applications, data retention time is more important than flash endurance. Likewise, read disturb, inter-cell interference, and temperature related charge losses occur [25–28]. Such charge losses result in shifts of the threshold voltages. These voltage shifts increase the bit error probability if the reference voltages are not adapted to the actual life cycle conditions.

Compared with hard-input read operation, the threshold voltage sensing operation to obtain the soft information causes a larger latency and a higher energy consumption [29]. Hence, the soft-input decoding should only be used for blocks where the hard-input decoding fails. Algorithms that adapt the read references can reduce the latency and energy consumption by avoiding soft-reads. Moreover, many soft-input decoding approaches assume reading procedures that maximize the mutual information (MI) between the input and output of flash memory. This requires an accurate adaptation of the read voltages and additional estimation of log-likelihood ratios (LLR) for the quantization intervals [15,22].

Several publications proposed threshold adaptation concepts [25,30–35]. These approaches adjust the read references to minimize bit error rates. For instance, the flash controller may apply a read-retry mechanism to adapt the reference voltages. Firstly, the controller reads the data with the default read reference. If the ECC decoder corrects the errors, there is no need to adjust the default value. In the case of decoding failures, the flash controller can perform repeated reads with certain threshold voltages until all errors are corrected by the decoder. However, such read-retry mechanisms can be very time-consuming.

To obtain statistical information about the cell state, a sensing directed estimation method was proposed in [31]. This method approximates the threshold voltage distribution as a Gaussian mixture. It estimates the mean and standard deviation values based on measured distributions. This scheme does not utilize any known pilot bits but employs extra memory sensing operations similar to the read-retry mechanisms. Such voltage sensing operations incur a larger energy consumption and a latency penalty. Similarly, a parameter estimation approach and a threshold adaptation policy is derived in [33]. This adaptation method is also based on the assumption of Gaussian voltage distributions. However, the cell threshold voltage distributions of MLC and TLC are actually highly asymmetric with exponential tails [23,36,37] and cannot be estimated with Gaussian distribution. In [25,30], coding methods based on Berger codes are suggested that exploit the asymmetric error probabilities of the channel. Berger codes are used for error detection. They can detect any number of one-to-zero errors, as long as no zero-to-one errors occurred in the same codeword. Likewise, Berger codes can detect any number of zero-to-one errors, as long as no one-to-zero bit-flip errors occur in the same code word. However, Berger codes cannot correct any errors.

In this work, we propose two methods for reference calibration that aim on minimizing the BER. Similar to the methods proposed in [32,34], we exploit the error correction code to obtain information about the cell state. Based on the number of observed errors in the read codeword, we predict the shift of the reference voltages. It is demonstrated that the observed number of errors is a good feature for the voltage shifts and is utilized for the initial calibration of the read thresholds. However, the total number of errors is only a good measure if there is a relatively large difference between the threshold for the first readout and the optimal read threshold. On the other hand, the total number of errors is not suited to estimate small changes of the read voltages, which occur for different pages from the same block. To cope with this issue, we propose a second approach which is a gradual channel estimation method. Like the approach based on Berger codes, we exploit the asymmetric error probabilities by counting the one-to-zero and zero-to-one errors independently. Based on the asymmetry, we apply a tracking method to adjust the read references from page to page.

We use information-theoretic measures to investigate the relation between the optimal read references and the statistical information obtained by counting errors. Both methods are investigated utilizing the mutual information between the optimal read voltage and the measured error values. Mutual information is also used to determine the parameters for the adaption algorithms. Numerical results based on flash measurements demonstrate that both methods can significantly reduce the BER of NAND flash memories and achieve nearly optimal performance.

This publication is organized as follows. In Section 2, we review some basic properties of flash memories and introduce a model for the threshold distributions. In Section 3, we present measurement results that motivate the proposed adaption methods and describe the principal approach. Moreover, we introduce the information-theoretical measures that are used to analyze and optimize the threshold calibration. The calibration of the read reference voltages is detailed in Section 4, and the gradual estimation method in Section 5. Finally, we conclude our work in Section 6.

2. Models for Voltage Distributions and Measurement Data

In this section, we discuss some basic properties of flash memories and introduce the notation. Furthermore, we consider a model to estimate the threshold voltage distributions. This curve fitting approach is used to determine the optimal read voltages.

Flash memory consists of transistors which have an insulated floating gate below the control gate. The information is stored as electric charges in the floating gate. The charge levels affect the threshold voltage of the transistor. The cell can be charged by applying a high program voltage to the control gate and erased by applying a high erase voltage to the substrate. The information can be read out by applying a reference threshold to differentiate erased and charged cells. The cells are organized in arrays of blocks and pages, where the control gates of the cells inside a page are connected using the so-called word line. Multiple pages constitute a block by connecting the drain of one cell to the source of the next cell. These connections form the bit lines.

2.1. Threshold Voltage Distributions

NAND flash technologies are differentiated by the number of bits stored per cell. While SLC flash stores only one bit per cell, TLC flash stores three bits per cell. Hence, a TLC flash requires eight different charge levels ($L_0, \dots, L_7$) and seven reference voltages ($r_1, \dots, r_7$) to read all three bits. The three bits of each cell are typically mapped to three different pages called MSB-, CSB-, and LSB-page, where MSB, CSB, and LSB stand for most significant bit, center significant bit, and least significant bit, respectively. Figure 1 illustrates the conditional distributions $f(V_{th}|L)$ of the threshold voltages V_{th} of the flash cells given the charge levels. The figure also provides a possible bit-labeling for the charge levels, where the colors indicate the corresponding page, i.e., MSB in red, CSB in blue, and LSB in green. Most flash memories support page-wise read operations. For instance, to read the MSB-page with the bit-labeling in Figure 1, we need to read at the reference voltages r_3 and r_7.

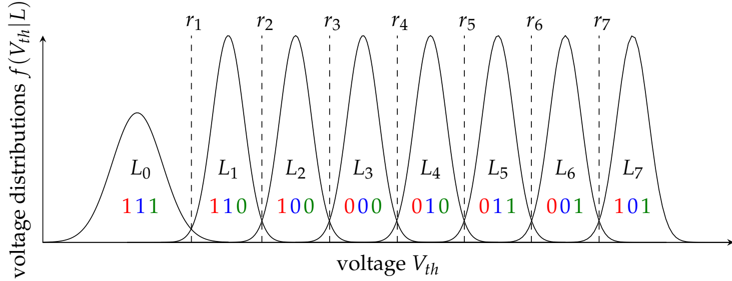

Figure 1. Model of voltage distribution of a TLC flash memory with labeled reference threshold voltages (dashed lines) and bit labeling.

The distributions $f(V_{th}|L)$ are not static but change over the lifetime of the flash. Two of the most important factors for these changes are the number of program/erase (P/E) cycles and the data retention time. The P/E cycles lead to a wear-out of the floating gate, while the data retention time leads to a charge loss. The distributions also depend on effects such as the program and read temperature and the charges of adjacent cells. Moreover, they can vary over pages and blocks.

The threshold distributions can be measured by successively applying all possible threshold voltages and counting the number of cells that are activated for the given voltage and charge level. For those measurements the data retention time is simulated by baking the chip for a specific time [38,39]. Figure 2 shows measured histograms for three different life-cycle states. The measurements were performed using fresh cells, cells with 1500 P/E-cycles and 13 h of baking, which simulates about two months of data retention, and cells with 3000 P/E-cycles and 80 h of baking, which corresponds to the end of life (EOL) defined by the manufacturer. Note that only the charged states L_1 to L_7 are shown. The voltages are represented by the index numbers of the possible voltage steps.

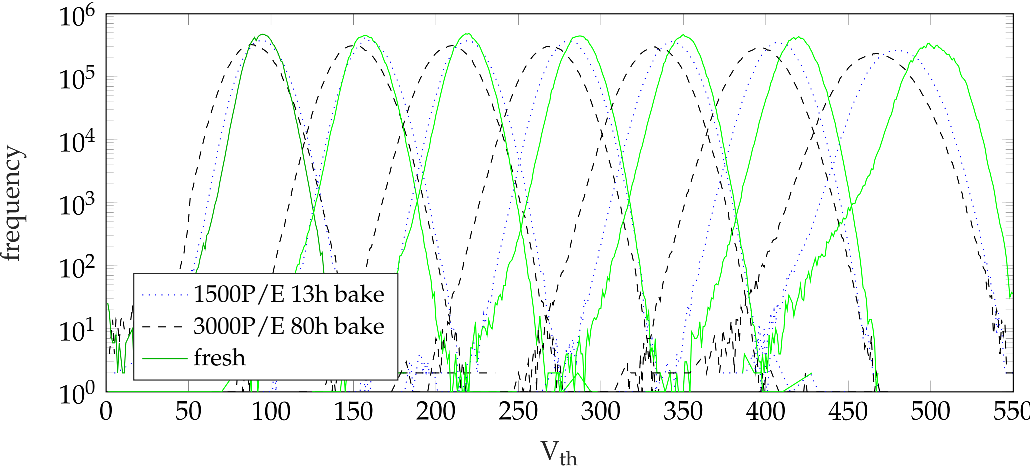

Figure 2. Measurements of TLC flash memory in different life-cycle states.

This figure shows the charge loss due to data retention time as well as the widening of the distributions due to wear-out. Due to the logarithmic scale of the y-axis, it is apparent that the distributions are not symmetric and the voltages are not normally distributed. While the threshold voltage distributions are often modeled as Gaussian distributions, the measured distributions show an exponential tail toward the lower voltages. To cope with this exponential tail, the exponentially-modified Gaussian distribution was proposed [23] which is the sum of two independent normal and exponential random variables. In [40], a Normal–Laplace mixture model was proposed. Similarly, the model in [41] copes with program errors in MLC-Flash memory. In our measurements, all three bit-pages are programmed simultaneously; hence, no write errors happen. Moreover, neither the exponentially-modified Gaussian distribution nor the Normal–Laplace mixture model is able to fit our measurements because these models do not allow for an inflection point, as observed in the measurement data.

2.2. Parameter Estimation

To approximate the measured distributions, we use a parameter estimation approach presented in [37]. The distributions $f(x \mid L_i)$ are modeled by a piecewise-defined function

$$f(x \mid L_i) = \frac{1}{n_i} \cdot \begin{cases} c_i \cdot e^{\lambda_i (x - x_i)} & x < x_i \\ \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x-\mu_i)^2}{2\sigma_i^2}} & x \geq x_i \end{cases} \tag{1}$$

with

$$c_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x_i-\mu_i)^2}{2\sigma_i^2}}. \tag{2}$$

With the cumulative distribution function $\phi(x)$ of the standard normal distribution

$$n_i = 1 + \frac{c_i}{\lambda_i} - \phi\left(\frac{x_i - \mu_i}{\sigma_i}\right) \tag{3}$$

ensures the normalization

$$\int_{-\infty}^{\infty} f(x \mid L_i)dx = 1. \tag{4}$$

The proposed model requires four parameters, i.e., $\sigma_i, \mu_i, \lambda_i$, and x_i. We model the exponential tail to the left by an exponential distribution and the rest by a normal distribution. This leads to a precise estimation of the measured distributions, which is shown by the fitted curve in Figure 3. This figure shows measurements of the voltage distribution $f(V_{th}|L_7)$ of a fresh memory together with a fitted distribution.

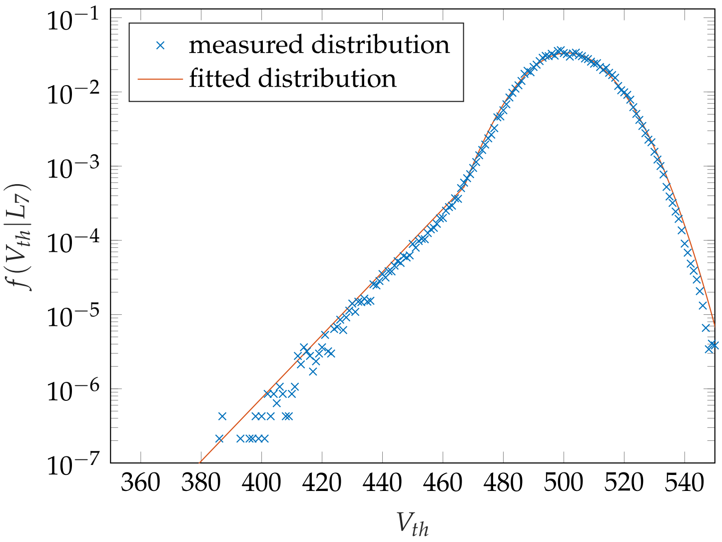

Figure 3. Fitted voltage distribution $f(V_{th}|L_7)$ of fresh flash memory.

The magnitude of the skewness of the measured distributions is very small. Consequently, the parameters σ_i^2 and μ_i can be computed from the sample mean and the sample variance, as the effects of the exponential tail on the mean and variance can be neglected. For the exponential distribution, a least squares regression line is fitted to the logarithm of the measured distribution. The distributions of the lower states also have an exponential tail towards the higher voltages, which is not modeled as it is negligible compared with the tail to lower voltages.

We use the curve fitting to determine the optimal threshold voltage $V_{opt}(r_i)$ for each read reference such that the bit error probability is minimized, i.e., the optimal threshold voltage is defined as

$$V_{opt}(r_i) = \underset{V}{\operatorname{argmin}} \int_{V}^{\infty} f(x \mid L_{i-1})\, dx + \int_{-\infty}^{V} f(x \mid L_i)\, dx. \tag{5}$$

2.3. Measurement Data

Note that the measured distributions are also conditioned on the life-cycle state, i.e., the number of P/E-cycles and data retention time, amongst many others. Since the distributions may vary over pages and blocks, the page and block numbers are also

conditions of the distributions. For simplicity, we omit those conditions in the notation. To obtain the unconditioned distributions, we would need measurements of all possible life-cycle states together with their a priori probabilities, which depend on the usage.

The measurements used in this work are 185 data sets each consisting of one block with 256 pages. The measurements are from eight different samples of the same flash memory model. The blocks were chosen uniformly over all available blocks on the chip. For the life-cycle states, the P/E cycles were measured up to 3000 in steps of 200 cycles. The data retention time simulated by baking was measured at 0, 13, 27, 42, 55, and 83 h of baking, where 83 h simulates one year. The value 3000 P/E with 83 h baking corresponds to the EOL state defined by the manufacturer. Additionally, measurements of read disturb as well as cross temperature effects were taken into account. For the read disturb, many read cycles were performed before the measurements. For the cross temperature effects, the pages were either programmed at −35 °C and read at a 85 °C or the other way around, while the data sets without cross temperature effects were measured at 25 °C. Most data sets combine several effects.

3. Problem Statement and Principal Approach

This section presents measurement results that motivate the suggested adaptation methods. We describe the principal approach. Moreover, we introduce the information-theoretical measures that are used to analyze and optimize the threshold calibration.

The voltage distributions of a flash memory change over the lifetime. Hence, a constant read reference voltage cannot minimize the bit error rate and the number of errors can surpass the error correction capability of the ECC decoder. To prevent this, the read reference voltage has to be adapted to the current life-cycle state of the flash to minimize the bit error rate. In this section, we discuss and analyze the statistical features that are used for the voltage adaptation.

Similarly to the methods proposed in [32,34], we exploit the error correction code to obtain information about the cell state. Based on the error correction code, we can determine the number of errors and the asymmetry of the error distribution, i.e., the probabilities of 1-to-0 errors and 0-to-1 errors are not equal [23,25,30,42]. In a nutshell, the number of errors provides information about the distance to the optimal threshold and the asymmetry provides information about the direction, i.e, whether to increase or decrease the threshold voltage. We use information-theoretic measures to investigate the relation between the optimal read references and the statistical information obtained by counting errors.

3.1. Asymmetric Error Probabilities

In [23,25,30,42], asymmetric error probabilities of the flash channel are reported, i.e., the probabilities of 1-to-0 errors and 0-to-1 errors are not equal. In our measurements, we found basically two causes for this asymmetry. Firstly, for some charge levels, the asymmetric shape of the threshold voltage distributions causes a certain asymmetry for the bit errors. Secondly, an offset of the reference voltage leads to asymmetric probabilities of 1-to-0 errors and 0-to-1 errors. Later on, we will exploit the asymmetric error probabilities due to the offsets for the adaptation of the read references.

The asymmetric error probabilities are indicated in Figure 4 which shows the number of errors for 1-to-0 errors and 0-to-1 errors. This figure considers reference r_3 at an EOL scenario. The error probabilities are balanced for the optimal threshold voltage which is indicated by the value $\Delta V_{th} = 0$, whereas the asymmetric error counts for other voltages can be exploited for the threshold adaptation.

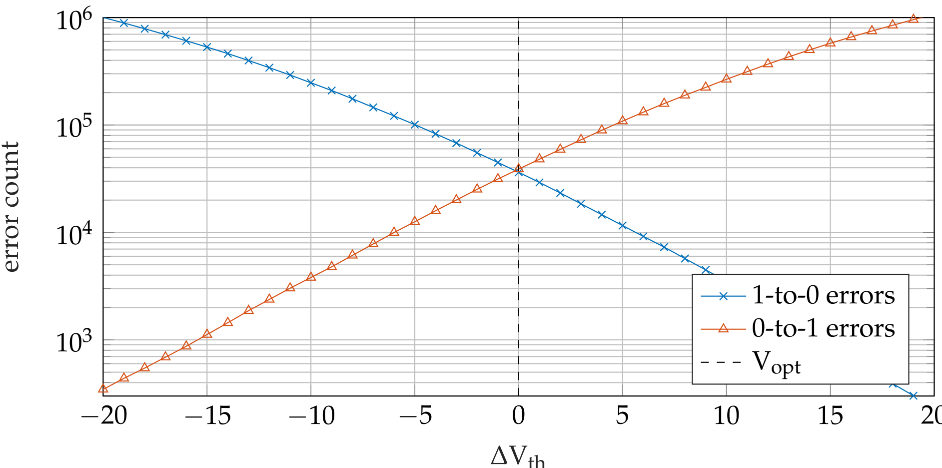

Figure 4. Error count for 1-to-0 errors and 0-to-1 errors versus the threshold voltage for r_3 at EOL.

As shown in Figure 2, the higher distributions are highly asymmetric for early-life cases. Figure 5 shows the error count for 1-to-0 errors, 0-to-1 errors, and the overall error count for reference r_7 for a fresh memory block. The error probabilities are asymmetric at the optimal threshold voltage indicated by the value $\Delta V_{th} = 0$, which needs to be considered for voltage adaptation.

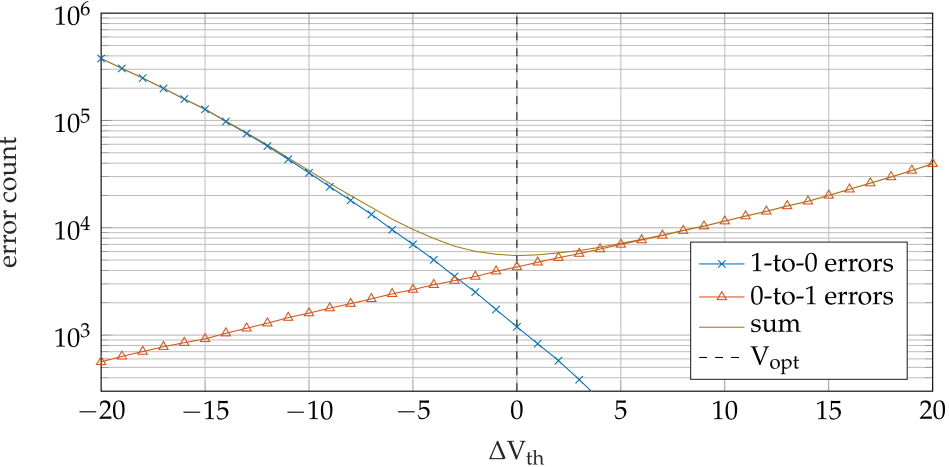

Figure 5. Error count for 1-to-0 errors and 0-to-1 errors versus threshold voltage for r_7 of fresh flash memory.

3.2. Number of Errors

Typically, different error correction codes are applied for the user-data and for meta data. Along with the user-data, a flash controller stores meta information about the life-cycle state, e.g., the number of P/E cycles. Typically, the meta data is protected by a significantly better error correction code than the user-data. Hence, it is possible to correct the errors in the codeword of the meta ECC even at a relatively large offset in the reference voltage. On the other hand, the number of bits for the meta data is small. Consequently, the number of errors in the meta ECC is not always a statistically relevant sample size.

Figure 6 illustrates the number of errors in the meta-data versus the read threshold at different life-cycle states. We consider a code with length $n = 508$ with a correction capability of up to $t = 21$ errors. The horizontal line in Figure 6 indicates the maximum number of correctable errors. The number of errors for a given voltage provides information about the flash aging. For instance, consider the value $V_{th} = 120$. For the EOL condition, we observed 11 errors on average, whereas the expected number of errors for 1500 P/E cycles

is only about two. Hence, the observed number of errors enables a prediction of the offset to the optimal read voltage. However, the provided information depends on V_{th}. In Figure 6, the slope of the error curves is very low near the optimum. When threshold voltages close to the optimal voltage V_{opt} are applied, only little information is gained by the number of errors. On the other hand, if the offset between V_{th} and V_{opt} is too large then the read data is likely to be undecodable. We use the mutual information between the optimal reference voltage V_{opt} and the observed number of errors to measure the information that is gained by counting errors.

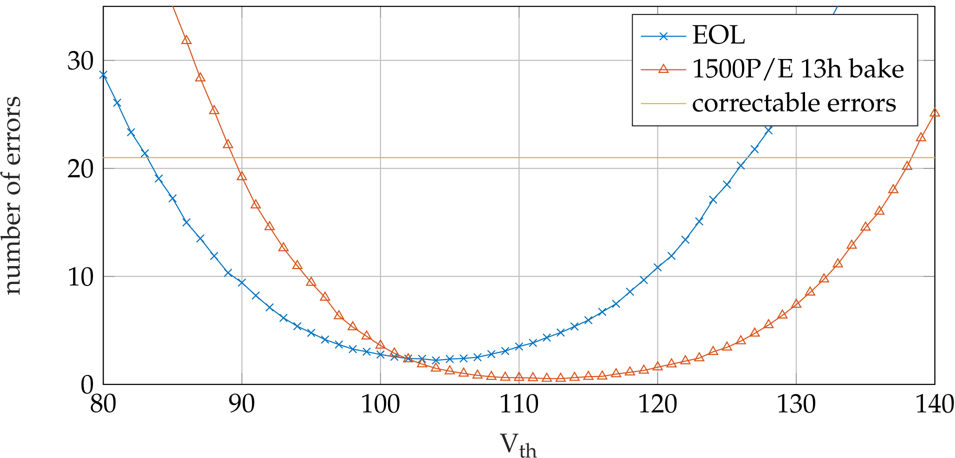

Figure 6. Number of errors in meta-data over threshold voltage.

3.3. Mutual Information

Next, we investigate the information that is obtained by counting the number of errors in the read codewords. Later, we predict the shift of the reference voltages based on the number of observed errors in the read codeword. In addition to the total number of errors, we exploit the asymmetry between 0-errors and 1-errors.

We use information theoretic measures to investigate the relation between the optimal read references and the statistical information obtained by counting errors. In particular, we consider the mutual information. The mutual information $I(X;Y)$ of two random variables is a measure for the information in bits gained about one variable when knowing the other variable. Let X and Y be two random variables with $X \in \{x_1, x_2, ..., x_N\}$ and $Y \in \{y_1, y_2, ..., y_M\}$ with probability distributions $f_X(x_i)$ and $f_Y(y_j)$ and joint probability distribution $f_{XY}(x_i, y_j)$. The mutual information between X and Y is defined as [43]

$$I(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} f_{XY}(x_i, y_j) \cdot log_2 \left(\frac{f_{X|Y}(x_i|y_j)}{f_X(x_i)} \right). \tag{6}$$

The mutual information is non-negative and upper bounded by the minimum of the entropies of the two variables [43]. For practical applications, it can have advantages using other measures, e.g., the residual error rate of a given decoder. But those measures are very specific to a given implementation.

Figure 7 shows the mutual information $I(E; V_{opt})$ between the optimal reference voltage V_{opt} and the observed number of errors E. The mutual information is determined over different life-cycle states. As a second observation, we consider a binary random variable A, which indicates the direction of the asymmetry between 0-errors and 1-errors. $A = 1$ indicates more 1-errors, and $A = 0$ indicates more 0-errors. The mutual information $I\big(A; V_{opt}\big)$ is also depicted in Figure 7. Note that this figure shows the curves for reference r_3.

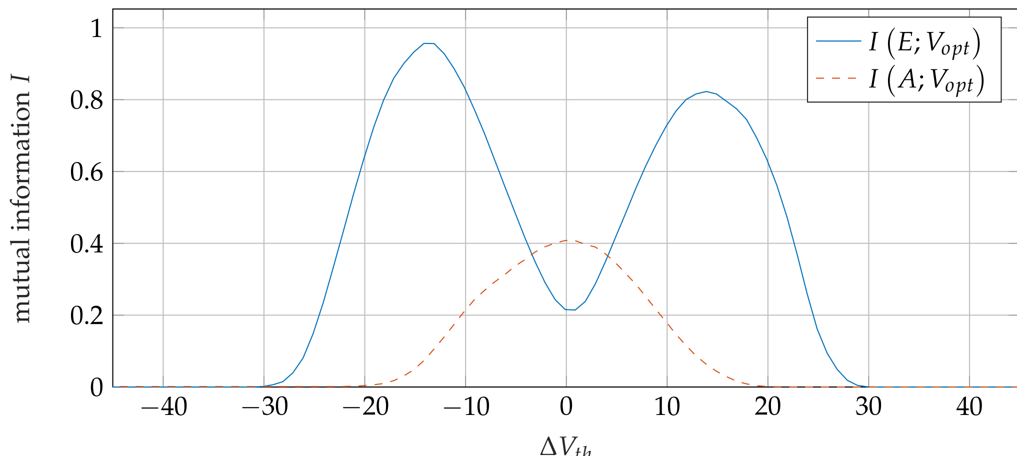

Figure 7. Mutual information between number of errors or binary variable A and optimal threshold voltage for reference r_3 over read voltage.

When using the number of errors in the meta-data, we get the maximum information about the optimum reference when reading at about -14, i.e., 14 voltage steps lower than the mean optimum reference. Near the optimum, the mutual information is fairly low, and this is due to the fact that the entropy of the number of errors is low in that region. Reading far from the optimum ($|\Delta V_{th}| \geq 30$), we obtain no information about V_{opt} because the read data is not correctable. In the case of a decoding failure, the exact number of errors remains unknown. Near the optimum $\Delta V_{th} = 0$, we can see that the binary variable A that indicates the asymmetry provides more information about V_{opt} than the number of errors E.

The asymmetry for higher charge values shown in Figure 5 also leads to an asymmetry in the mutual information. This can be seen in Figure 8, where the mutual information versus the read voltage is plotted for reference r_7. As we can see, the maximum of the mutual information between the asymmetry A and the optimum threshold voltage is not at $\Delta V_{th} = 0$. Similarly, the local minimum for $I(E; V_{opt})$ does not occur at $\Delta V_{th} = 0$.

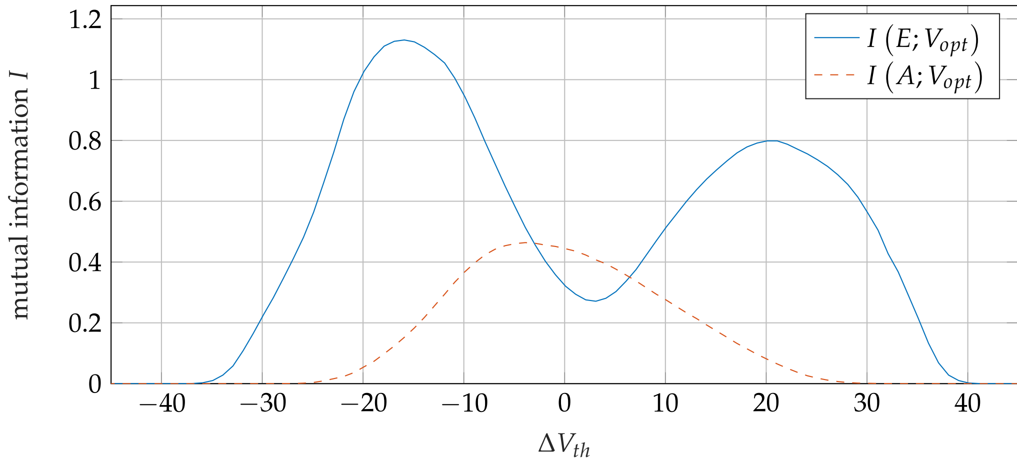

Figure 8. Mutual information between number of errors or binary variable A and optimal threshold voltage for reference r_7 over read voltage.

3.4. Principal Approach

In the following, we consider two adaptation methods depending on the a priori knowledge about the life-cycle state of the block. Firstly, a calibration approach aims on finding a good reference threshold with a small number of read attempts when we have no knowledge about the life-cycle state. Secondly, a tracking approach, based on a gradual

channel estimation method, adjusts the read threshold while iterating over successive pages, i.e., we calibrate the first page of a block and sequentially read other pages from the same block. The second approach applies only small changes to the read reference voltage because adjacent pages in a block typically have similar distributions. On the other hand, a tracking algorithm should be as time efficient as possible, whereas the calibration may require multiple reads and voltage adjustment steps, which is quite time intensive. We have some knowledge about the life-cycle state, and we aim to use this information to update the reference without any additional reads.

The proposed methods are also motivated by the above observations about the mutual information. We use the number of errors in the meta-data for the calibration because it provides information about V_{opt} over a relatively wide range of read voltages V_{th}. On the other hand, we propose a tracking method based on the asymmetry of 0- and 1-errors since the asymmetry has a higher mutual information for voltages near to the optimum, where the total number of error provides little information.

4. Calibration of the Read Reference Voltages

In this section, we discuss a data-driven calibration method to predict the optimal threshold voltage without any previous knowledge of the flash life-cycle state. We exploit the observed number of errors in the codeword for the meta-data. We assume that the error correction for meta-data is stronger than the code used for the user-data. Hence, the voltage range that allows for successful decoding is larger.

For the calibration, we determine a voltage V_{cal} for the first reading over all available life-cycle states. This voltage is obtained by maximizing the mutual information between the number of errors and the optimum threshold voltage over all data sets. Consider Figure 7 for reference r_3, the best voltage V_{cal} is at $V_{th} = -14$. The analysis of the mutual information has to be done for each reference.

Firstly, the meta-data are read at the threshold voltage V_{cal} for reference r_3 and decoded. In the case of successful decoding, the number of errors is known and can be used to adapt the threshold voltage for reading the user-data. This adaptation is done using a simple look-up in a table, which specifies a voltage offset for a given number of observed errors. The calculation for this table is described later on. In the case of a decoding failure, we apply a read-retry mechanism. We adapt the threshold voltage and read the meta-data again. For this second adaptation step, the same analysis of mutual information is performed but using only data sets of the pages, where the meta-data were not correctable at V_{cal}. Again, we choose the threshold voltage which maximizes the mutual information between the number of errors and the optimum threshold voltage. In our measurements, a maximum of two read trials was sufficient for the calibration. In general, the maximum number of read-retries will depend on the error correction capability of the code.

The number of errors in the meta-data is used to adapt the threshold voltage, which is then used to read the user-data. To obtain the best threshold voltage given a specific number of errors, we propose a look-up table filled with a data-driven approach. To determine the look-up table for the adaptation, we cluster the data sets of the measurements based on the number of errors in the meta-data when reading at the calibration voltage V_{cal}. The clustering approach is similar to the well-known k-means clustering [44]. The number of clusters is determined by the error correction capability t of the code. We have $t + 1$ clusters, one for every possible number of errors. In contrast to k-means clustering, we require no iterative updates because the partition of the data set is obtained based on the number of errors observed in each page. Similar to k-means clustering, we determine the representatives (cluster centroid) of the clusters by minimizing the Euclidean distance to all samples in the cluster.

Figure 9 shows the clusters for 0, 10, and 21 errors in the meta-data of the MSB page. The MSB page is defined by references r_3 and r_7. In the figures the optimum threshold voltage for the reference r_7 is plotted versus the optimum threshold for reference r_3. Each point depicts a possible pair of optimal voltages that was observed for a given number

of errors in the meta-data. The small crosses in the plots for the three clusters, indicate the representatives of the clusters. As can be seen in Figure 9, the clusters are not well separated. This results from the small number of bits in the meta-data ($n = 504$), which leads to a high variance in the number of errors.

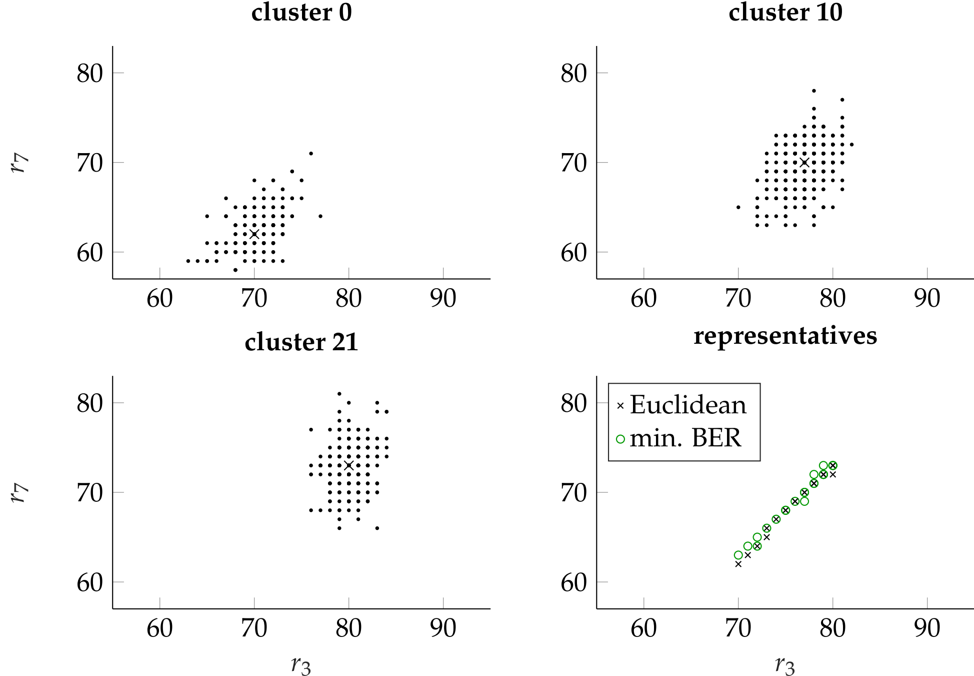

Figure 9. Exemplary clusters for filling calibration table.

The lower right-hand figure shows the representatives of all clusters obtained by minimizing the Euclidean distance or by minimizing the mean bit error rate (BER) in the cluster, respectively. We observe that both ways to determine the representatives achieve similar results. Since the V_{cal} is typically lower than the V_{opt} (cf. Figure 7), a higher error count leads to larger predicted threshold voltage of the representative.

Figure 10 shows the results of this calibration procedure for three different life-cycle states. The x-axis shows the page number and the y-axis the bit error rate (BER) when reading at the default threshold voltage, the optimum threshold voltage for each page, and the calibrated threshold voltage. Each data set consists of 256 pages; the page numbers from 1 to 256 correspond to 1500 P/E cycles and 13 h baking; the pages from 257 to 512 consider 1500 P/E cycles and 80 h baking; the last 256 pages correspond to 3000 P/E cycles and 80 h of baking (EOL). The default threshold voltage is constant and only suitable for fresh flash memories, which is not shown in this figure because typically no calibration is required. The optimum threshold voltage is calculated for each page individually using the measurements.

While the calibration typically does not result in the optimum threshold voltage for each page, the BER is very close to the minimum possible error rate. Hence, this method provides a sufficiently precise calibration. On the other hand, it requires up to two times reading the meta-data and up to three times changing the threshold voltage, which leads to a relatively high overall latency.

Next, we analyze the impact of the calibration on the error correction performance. For this discussion, we consider a regular LDPC code of rate $R = 0.9$ and with a dimension $k = 16,384$, i.e., the code is designed for 2kByte user data. For decoding, we consider the decoder proposed in [45]. With soft-input decoding using three soft-bits (five read voltages for each reference), the decoder achieves a frame error rate (FER) of less than 10^{-9} for bit error rates below 0.0088. With hard-input decoding, the FER value of 10^{-9} is achieved at a BER value of 0.0038.

For the EOL case, the minimum BER without calibration is about 0.03. This error rate is above the error correction capability of the decoder even with soft-input decoding. With calibration, on the other hand, the maximum observed BER values among all pages in the EOL data is about 0.008. Hence, a frame error rate (FER) of less than 10^{-9} is achieved with calibration even in the worst-case condition. Thus, for end-of-life conditions, calibration is required to reduce the probability of decoding failures.

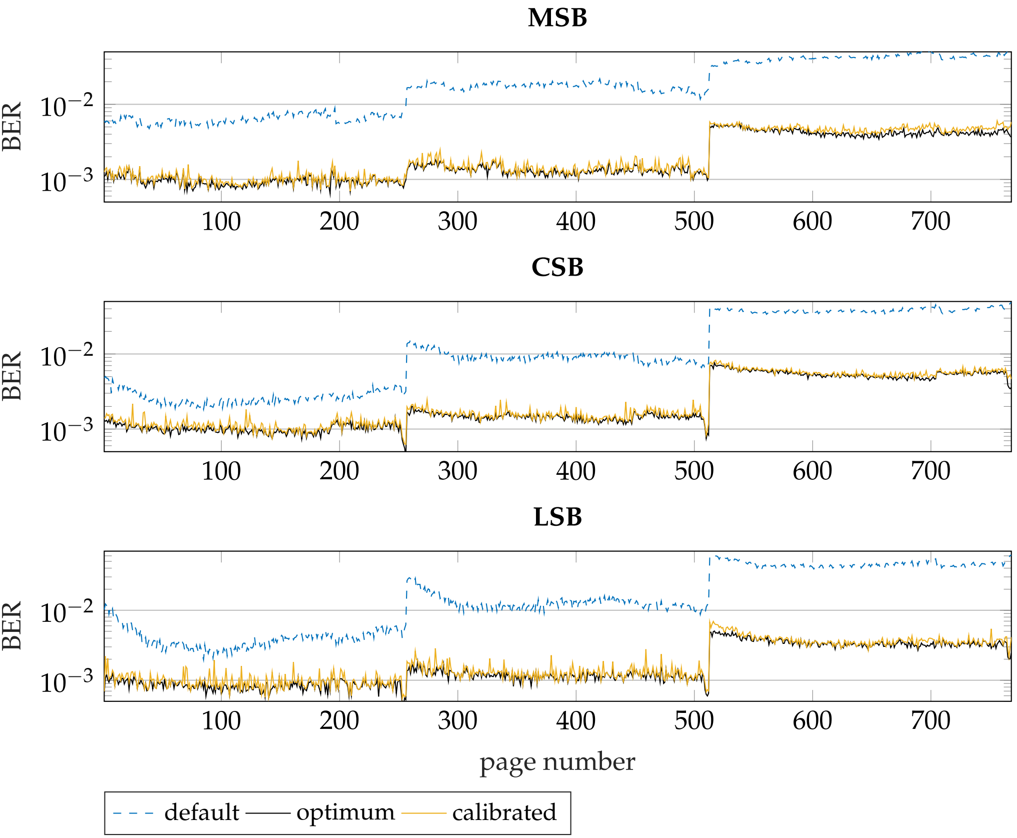

Figure 10. Residual BER of proposed calibration compared to minimum BER.

However, reading soft information is extremely time-intensive. Consequently, it is interesting whether soft-decoding is required or hard-input decoding is sufficient. For a BER of 0.006, the hard-input decoding results in a FER of about 0.01, which is unacceptable. Hence, those cases definitively require soft-reading. Consider for instance the first life-cycle state shown in Figure 10. Note that for about 22% of the pages, the BER is higher than 0.006 without calibration. With calibration, on the other hand, the highest BER in the first life-cycle state is about 0.002, which leads to a FER lower than 10^{-9} with hard-input decoding. Hence, for pages with mid-life conditions, the calibration reduces the number of pages, which require soft-reading. This reduces the read latency and the power consumption for the read operation and the decoding.

In the next section, we discuss a tracking approach which reduces the latency for the threshold adaption of the remaining pages of a block.

5. Tracking the Read Threshold

As described before, each page of a flash block has a specific readout threshold voltage V_{opt} at which the lowest number of errors is observed. Unfortunately, these voltages may vary from page to page of the same block even when all pages have the same life-cycle condition, as shown in Figure 11.

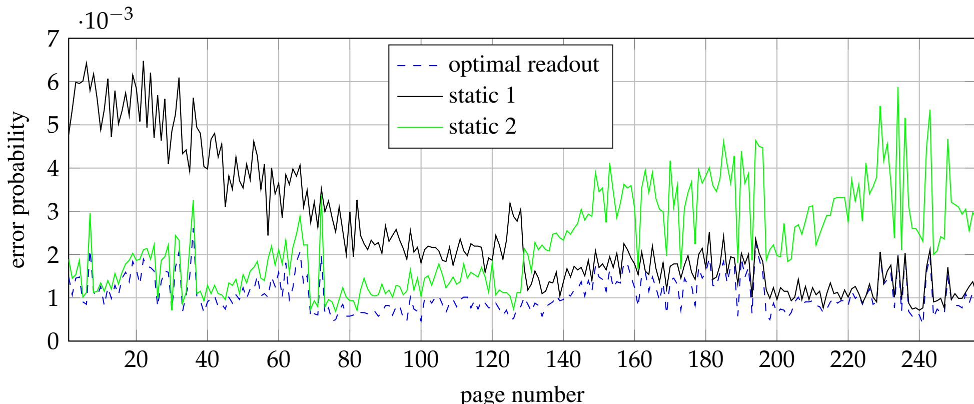

Figure 11. Error probability of a TLC block on r_7 (55 h bake, 3000 P/E) for optimal readout and 2 worst case static readouts.

This figure illustrates the error probability for each page of a single TLC block at reference r_7 with 3000 P/E-cycles and 55 h of baking. Note that the error rate at reference r_7 is not equal to the BER of the MSB page because the MSB page uses references r_3 and r_7 (cf. Figure 1). In Figure 11, the curve labeled by *optimal readout* indicates the smallest possible error probability for each page using the individual V_{opt} values. The two other curves correspond to two static threshold voltages. For the curve *static 1*, the threshold voltage was calibrated to the end of the block and for curve 2, the threshold voltage was calibrated to the beginning of the block. We observe that the static threshold voltages lead to significantly higher bit error rates.

The previous example shows that constant reference voltages for a complete black can cause a considerable performance degradation. To cope with this problem, we propose a tracking procedure for the reference voltages for consecutive pages. We assume that the calibration algorithm determined a near-optimal threshold voltage for the first page of the block.

As shown in Figure 4, the optimal threshold voltage corresponds to a balanced error probability for 1-to-0 errors and 0-to-1 errors. The binary variable A indicates the direction of the asymmetry between 0-errors and 1-errors. $A = 1$ indicates more 1-to-0 errors and $A = 0$ indicates more 0-to-1 errors. Hence, A also provides the direction for the voltage change toward the equilibrium of the error probabilities.

A simple approach for the tracking adapts the threshold voltage for the next page by a ± 1 voltage step depending on A. This concept is illustrated in Figure 12. This figure depicts the adaptation of the threshold voltages over one block for reference r_3. The blue line shows the values of V_{opt} for each page. The red curve presents the results with tracking, where ± 1 voltage steps are applied. We can observe that the adaptation follows the optimal voltage values.

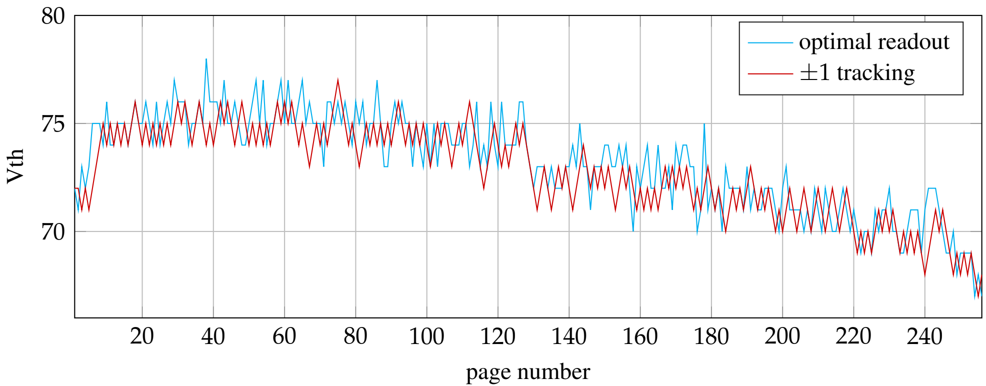

Figure 12. Threshold voltage adaptation of a TLC-Block on reference r_3 (55 h bake, 3000 P/E).

The equilibrium of the error probabilities does not always coincide with V_{opt}. This is indicated in Figures 5 and 8. For instance, the equilibrium of the error probabilities corresponds to the value $\Delta V_{th} = -3$ in Figure 5. In such a case, the tracking towards the equilibrium results in a bias, i.e., the estimated voltages systematically overestimate or underestimate V_{opt}. This effect is displayed on the top part of Figure 13. This figure shows the adaptation of the threshold voltages over one block for reference r_7. The dashed black lines correspond to the static voltages from Figure 11, and the blue line shows the values of V_{opt} for each page. The red curve depicts the results with tracking, where ± 1 voltage steps are applied. We can observe that the adaptation follows the optimal voltage values, but diverges slightly from the optimal curve, where V_{opt} is mostly underestimated. On average, the bias corresponds to the shift $\Delta V_{th} = -3$ of the equilibrium.

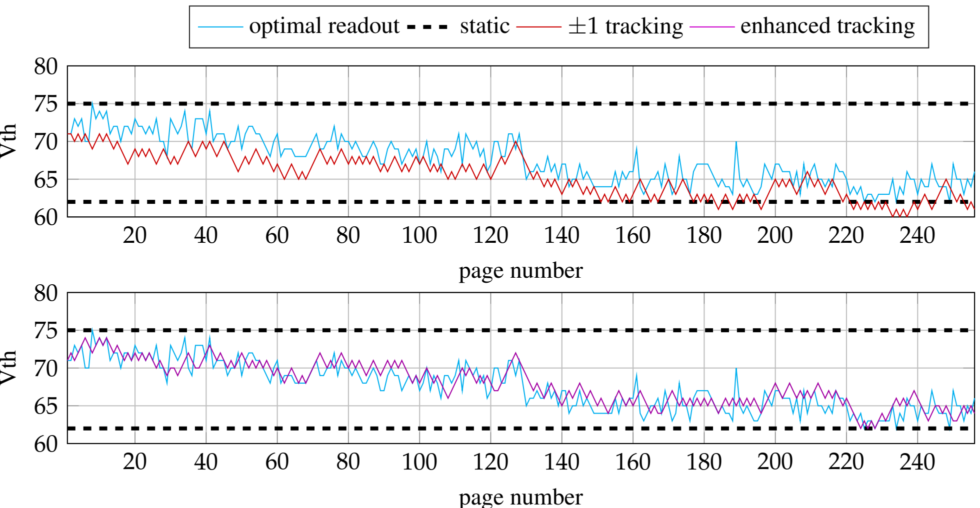

Figure 13. Threshold voltage adaptation of a TLC-Block on reference r_7 (55 h bake, 3000 P/E).

In such a case, an enhanced tracking algorithm can be applied that corrects this bias. Instead of changing the value of A at the equilibrium of the error probabilities, we calculate the rate of error number corresponding to the optimal voltage V_{opt} or the shift $\Delta V_{th} = 0$, respectively. For instance, in Figure 5 the number of 1-to-0 errors is about 4-times the number of 0-to-1 errors. In this case, we can choose the value if A such that $A = 1$ indicates four times more 1-to-0 errors then 0-to-1 errors. To average the ratio over different life-cycle states, we choose the ratio that corresponds to the value $\Delta V_{th} = 0$ for the mutual information $I(A; V_{opt})$. The lower part of Figure 13 shows the adjusted readout voltage with this approach. Now the adaptation is closer to the optimal voltage V_{opt}.

Figures 12 and 13 show results for one block in a single life-cycle state. Figure 14 demonstrates the tracking accuracy averaged over different life-cycle states. In this box plot, the median of the error probability at r_7 is represented as the central marker and a box represents the interquartile range, i.e., the interval for 50% of all measurements when ordered from lowest-to-highest error probability. The whiskers represent the most extreme data points without outliers.

The box plots consider the simple ± 1 voltage adaptation, the enhanced adaptation, and the worst-case static readout voltages. For every block in the data set, the worst-case readout voltage was selected within the range of optimal readout voltages. Hence, these values correspond to an optimal readout voltage of one page within a block, which was selected as static readout voltage for the entire block. The results in Figure 14 show that a static readout voltage can significantly increase the error probability compared with the optimal voltage.

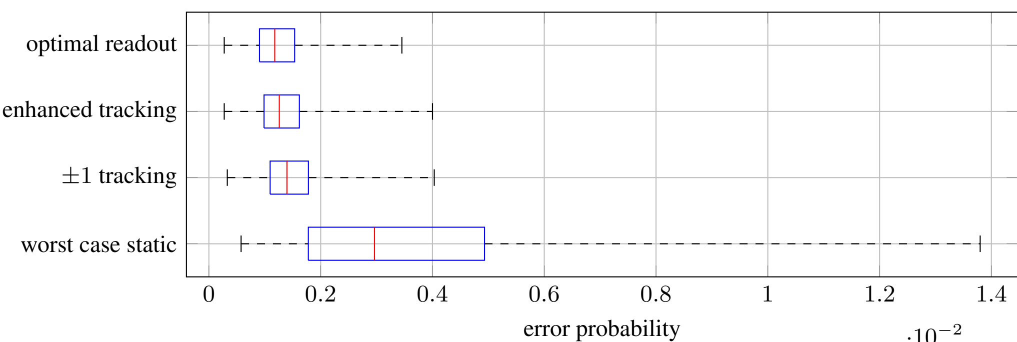

Figure 14. Error statistic on reference r_7 of all pages of 27 Blocks with 55 h bake 3000 P/E.

Again, we consider the LDPC code used in Section 4 for the discussion of the ECC performance. We discuss results for the FER using hard-input decoding. For a BER of 0.006, the hard-input decoding results in a FER of about 0.01. Therefore, we assume that a soft-read operation is required for all BER values greater or equal 0.006. For the worst-case static readout voltage, a significant amount of pages exceed the BER of 0.006. When using the tracking approach the maximum BER is about 0.004, which results in a FER of less than 10^{-9} with hard-input decoding. Consequently, the tracking improves the latency and the power consumption by reducing the profitability of soft-read and soft-input decoding procedures.

The ± 1 adaptation and the enhanced adaptation obtain a similar performance which is close to the optimal read voltage. Consequently, the bias compensation is not critical with respect to the achievable error rate. However, the bias compensation may be required for soft-input decoding [15,22]. Many soft-input decoding algorithms assume quantization intervals that maximize the mutual information between the input and output of flash memory, which requires an accurate voltage adaptation.

6. Conclusions

The read threshold voltages of nonvolatile NAND flash memories depend on various factors such as the number of program/erase cycles, data retention times, and read disturb. Based on measured flash data, fixed predefined read thresholds lead to high bit error rates.

In this work, we proposed two methods for adapting the read thresholds based on the number of errors observed in the codeword of an error correction code. The total number of errors in the codeword as well as the numbers of 1-to-0 and 0-to-1 errors are used as statistical features. These features were investigated utilizing the mutual information between the optimal read voltage and the measured error values. The total number of errors provides information about the distance to the optimal threshold and the asymmetry of the errors provides information about the direction. The total number of errors enables a prediction of the voltage shift for the initial calibration of the read thresholds. The second approach is a gradual channel estimation to cope with threshold calibration errors and voltage shifts over different pages of the same block. This approach is based on the asymmetric error distribution. Numerical results based on flash measurements demonstrate that both methods can significantly reduce the BER of NAND flash memories.

In future work, the predictions of the voltage offsets for larger distances between pages should be considered. The proposed tracking method assumes sequential reads of pages. This simple adaption method is only suitable for pages nearby. Random access to pages from the same block would still require frequent calibration steps in end-of-life scenarios. Moreover, the impact of calibration errors on soft-input decoding algorithms requires more investigations [15,22,29]. Such soft-input decoding approaches typically assume quantization intervals that maximize the mutual information between the input and output of flash memory, which requires an accurate voltage adaptation.

Author Contributions: The research for this article was exclusively undertaken by J.-P.T., D.N.B., J.F. and J.L.; conceptualization and investigation, J.-P.T., D.N.B. and J.F.; software and validation, J.-P.T. and D.N.B.; measurements, J.L.; writing—review and editing, J.-P.T., D.N.B., J.F. and J.L.; writing—original draft preparation, J.-P.T., D.N.B. and J.F.; supervision, project administration, and funding acquisition J.F. All authors have read and agreed to the published version of the manuscript.

Funding: The German Federal Ministry of Research and Education (BMBF) supported the research for this article (16ES1045) as part of the PENTA project 17013 XSR-FMC.

Acknowledgments: The authors would like to thank Hyperstone GmbH, Konstanz, for supporting the research for this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Spinelli, A.S.; Compagnoni, C.M.; Lacaita, A.L. Reliability of NAND Flash Memories: Planar Cells and Emerging Issues in 3D Devices. *Computers* **2017**, *6*, 16. [CrossRef]
2. Khosla, R.; Rolseth, E.G.; Kumar, P.; Vadakupudhupalayam, S.S.; Sharma, S.K.; Schulze, J. Charge Trapping Analysis of Metal/Al_2O_3/SiO_2/Si, Gate Stack for Emerging Embedded Memories. *IEEE Trans. Device Mater. Reliab.* **2017**, *17*, 80–89. [CrossRef]
3. Soni, M.; Soni, A.; Sharma, S.K. NrGO Floating Gate/SiO_XN_Y Tunneling Layer Stack for Nonvolatile Flash Memory Applications. *IEEE Trans. Device Mater. Reliab.* **2020**, *20*, 570–575. [CrossRef]
4. Zambelli, C.; Micheloni, R.; Olivo, P. Reliability challenges in 3D NAND Flash memories. In Proceedings of the IEEE 11th International Memory Workshop (IMW), Monterey, CA, USA, 12–15 May 2019; pp. 1–4. [CrossRef]
5. Micheloni, R.; Marelli, A.; Ravasio, R. *Error Correction Codes for Non-Volatile Memories*; Springer: Berlin/Heidelberg, Germany, 2008.
6. Dolecek, L.; Cassuto, Y. Channel Coding for Nonvolatile Memory Technologies: Theoretical Advances and Practical Considerations. *Proc. IEEE* **2017**, *105*, 1705–1724. [CrossRef]
7. Zhang, X.; Parhi, K.K. High-speed architectures for parallel long BCH encoders. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **2005**, *13*, 872–877. [CrossRef]
8. Freudenberger, J.; Spinner, J. A configurable Bose-Chaudhuri-Hocquenghem codec architecture for flash controller applications. *J. Circuits Syst. Comput.* **2014**, *23*, 1–15. [CrossRef]
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 Regul. Pap.* **2011**, *58*, 429–439. [CrossRef]
10. Cho, S.; Kim, D.; Choi, J.; Ha, J. Block-Wise Concatenated BCH Codes for NAND Flash Memories. *IEEE Trans. Commun.* **2014**, *62*, 1164–1177. [CrossRef]
11. Zhao, K.; Zhao, W.; Sun, H.; Zhang, X.; Zheng, N.; Zhang, T. LDPC-in-SSD: Making Advanced Error Correction Codes Work Effectively in Solid State Drives. In Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST 13), San Jose, CA, USA, 12–15 February 2013; pp. 243–256.
12. Wang, J.; Vakilinia, K.; Chen, T.Y.; Courtade, T.; Dong, G.; Zhang, T.; Shankar, H.; Wesel, R. Enhanced Precision Through Multiple Reads for LDPC Decoding in Flash Memories. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 880–891. [CrossRef]
13. Lin, W.; Yen, S.W.; Hsu, Y.C.; Lin, Y.H.; Liang, L.C.; Wang, T.C.; Shih, P.Y.; Lai, K.H.; Cheng, K.Y.; Chang, C.Y. A low power and ultra high reliability LDPC error correction engine with Digital Signal Processing for embedded NAND Flash Controller in 40 nm CMOS. In Proceedings of the Symposium on VLSI Circuits Digest of Technical Papers, Honolulu, HI, USA, 10–13 June 2014; pp. 1–2.
14. Haymaker, K.; Kelley, C.A. Structured Bit-Interleaved LDPC Codes for MLC Flash Memory. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 870–879. [CrossRef]
15. Sandell, M.; Ismail, A. Machine Learning for LLR Estimation in Flash Memory With LDPC Codes. *IEEE Trans. Circuits Syst. II Express Briefs* **2021**, *68*, 792–796. [CrossRef]
16. Spinner, J.; Freudenberger, J.; Shavgulidze, S. A Soft Input Decoding Algorithm for Generalized Concatenated Codes. *IEEE Trans. Commun.* **2016**, *64*, 3585–3595. [CrossRef]
17. Spinner, J.; Rohweder, D.; Freudenberger, J. Soft input decoder for high-rate generalised concatenated codes. *IET Circuits Devices Syst.* **2018**, *12*, 432–438. [CrossRef]
18. Rajab, M.; Freudenberger, J.; Shavgulidze, S. Soft-input Bit-flipping Decoding of Generalized Concatenated Codes for Application in Non-volatile Flash Memories. *IET Commun.* **2018**, *133*, 460–467.
19. Bian, J.; Zhao, S.; Kong, L. Rate-adaptive Polar Codes Design for MLC NAND Flash Memory. In Proceedings of the IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; pp. 11–16. [CrossRef]

20. Kong, L.; Liu, Y.; Liu, H.; Zhao, S. Protograph QC-LDPC and Rate-Adaptive Polar Codes Design for MLC NAND Flash Memories. *IEEE Access* **2019**, *7*, 37131–37140. [CrossRef]
21. Taranalli, V.; Uchikawa, H.; Siegel, P.H. Channel Models for Multi-Level Cell Flash Memories Based on Empirical Error Analysis. *IEEE Trans. Commun.* **2016**, *64*, 3169–3181. [CrossRef]
22. Freudenberger, J.; Rajab, M.; Shavgulidze, S. Estimation of channel state information for non-volatile flash memories. In Proceedings of the IEEE 7th International Conference on Consumer Electronics (ICCE), Berlin, Germany, 3–6 September 2017.
23. Freudenberger, J.; Rajab, M.; Shavgulidze, S. A Source and Channel Coding Approach for Improving Flash Memory Endurance. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **2018**, *26*, 981–990. [CrossRef]
24. Cai, Y.; Luo, Y.; Haratsch, E.F.; Mai, K.; Mutlu, O. Data retention in MLC NAND flash memory: Characterization, optimization, and recovery. In Proceedings of the IEEE 21st International Symposium on High Performance Computer Architecture (HPCA), Burlingame, CA, USA, 7–11 February 2015; pp. 551–563. [CrossRef]
25. Sala, F.; Gabrys, R.; Dolecek, L. Dynamic Threshold Schemes for Multi-Level Non-Volatile Memories. *IEEE Trans. Commun.* **2013**, *61*, 2624–2634. [CrossRef]
26. Cao, C.; Fair, I. Mitigation of Inter-Cell Interference in Flash Memory With Capacity-Approaching Variable-Length Constrained Sequence Codes. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 2366–2377. [CrossRef]
27. Yassine, H.; Coon, J.P.; Simmons, D.E. Index Programming for Flash Memory. *IEEE Trans. Commun.* **2017**, *65*, 1886–1898. [CrossRef]
28. Park, S.K.; Moon, J. Characterization of Inter-Cell Interference in 3D NAND Flash Memory. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2021**, *68*, 1183–1192. [CrossRef]
29. Qi, S.; Feng, D.; Liu, J. Optimal voltage signal sensing of NAND flash memmory for LDPC code. In Proceedings of the 2014 IEEE Workshop on Signal Processing Systems (SiPS), Belfast, UK, 20–22 October 2014; pp. 1–6. [CrossRef]
30. Zhou, H.; Jiang, A.; Bruck, J. Error-correcting schemes with dynamic thresholds in nonvolatile memories. In Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings, St. Petersburg, Russia, 31 July–5 August 2011; pp. 2143–2147.
31. Lee, D.; Sung, W. Estimation of NAND Flash Memory Threshold Voltage Distribution for Optimum Soft-Decision Error Correction. *IEEE Trans. Signal Process.* **2013**, *61*, 440–449. [CrossRef]
32. Lee, D.H.; Sung, W. Decision Directed Estimation of Threshold Voltage Distribution in NAND Flash Memory. *IEEE Trans. Signal Process.* **2014**, *62*, 919–927. [CrossRef]
33. Peleato, B.; Agarwal, R.; Cioffi, J.M.; Qin, M.; Siegel, P.H. Adaptive read thresholds for NAND flash. *IEEE Trans. Commun.* **2015**, *63*, 3069–3081. [CrossRef]
34. Rajab, M.; Thiers, J.; Freudenberger, J. Read Threshold Calibration for Non-Volatile Flash Memories. In Proceedings of the IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, 8–11 September 2019; pp. 119–123.
35. Yang, L.; Wang, Q.; Li, Q.; Yu, X.; He, J.; Huo, Z. Gradual Channel Estimation Method for TLC NAND Flash Memory. *IEEE Embed. Syst. Lett.* **2021**, 1. [CrossRef]
36. Cai, Y.; Haratsch, E.F.; Mutlu, O.; Mai, K. Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling. In Proceedings of the 2013 Design, Automation Test in Europe Conference Exhibition (DATE), Grenoble, France, 18–22 March 2013; pp. 1285–1290. [CrossRef]
37. Thiers, J.P.; Bailon, D.N.; Freudenberger, J. Bit-Labeling and Page Capacities of TLC Non-Volatile Flash Memories. In Proceedings of the IEEE 10th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, 9–11 November 2020; pp. 1–6. [CrossRef]
38. Miccoli, C.; Paolucci, G.M.; Compagnoni, C.M.; Spinelli, A.S.; Goda, A. Cycling pattern and read/bake conditions dependence of random telegraph noise in decananometer NAND flash arrays. In Proceedings of the IEEE International Reliability Physics Symposium, Monterey, CA, USA, 19–23 April 2015; pp. MY.9.1–MY.9.6. [CrossRef]
39. Samadder, T.; Kumar, S.; Thakor, K.; Mahapatra, S. A Theoretical Framework for Trap Generation and Passivation in NAND Flash Tunnel Oxide During Distributed Cycling and Retention Bake. In Proceedings of the IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 21–25 March 2021; pp. 1–6. [CrossRef]
40. Parnell, T.; Papandreou, N.; Mittelholzer, T.; Pozidis, H. Modelling of the threshold voltage distributions of sub-20nm NAND flash memory. In Proceedings of the IEEE Global Communications Conference, Austin, TX, USA, 8–12 December 2014; pp. 2351–2356.
41. Luo, Y.; Ghose, S.; Cai, Y.; Haratsch, E.F.; Mutlu, O. Enabling Accurate and Practical Online Flash Channel Modeling for Modern MLC NAND Flash Memory. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 2294–2311. [CrossRef]
42. Yaakobi, E.; Ma, J.; Grupp, L.; Siegel, P.; Swanson, S.; Wolf, J. Error characterization and coding schemes for flash memories. In Proceedings of the IEEE GLOBECOM Workshops, Miami, FL, USA, 6–10 December 2010; pp. 1856–1860.
43. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1991.
44. MacKay, D. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
45. Vasić, B.; Declercq, D.; Reynwar, B. Iterative Decoders Robust to Threshold Voltage Uncertainty. In Proceedings of the 9th Annual Non-Volatile Memories Workshop, San Diego, CA, USA, 11–13 March 2018.