Noise Modeling and Capacity Analysis for NAND Flash Memories

Qing Li*, Anxiao (Andrew) Jiang*, and Erich F. Haratsch†
† Flash Components Division, LSI Corporation, San Jose, CA, 95131
* Computer Sci. and Eng. Dept.,Texas A & M University, College Station, TX, 77843
*{qingli, ajiang}@cse.tamu.edu

***Abstract*—Flash memories have become a significant storage technology. However, they have various types of error mechanisms, which are drastically different from traditional communication channels. Understanding the error models is necessary for developing better coding schemes in the complex practical settings. This paper endeavors to survey the noise and disturbs in NAND flash memories, and construct channel models for them. The capacity of flash memory under these models is analyzed, particularly regarding capacity degradation with flash operations, the trade-off of sub-thresholds for soft cell-level information, and the importance of dynamic thresholds.**

I. Introduction

Flash memories have become a significant storage technology. Despite their wide applications, flash memories are far from ideal. Flash memories have quite a few noise or disturb mechanisms, including retention errors, inter-cell interference, random noise, programming errors, read and write disturbs, and stuck cells [1]. These mechanisms have quite different characteristics from traditional communication channels.

Understanding the channel models for noise/disturbs in flash memories is important for designing better coding schemes in the complex practical settings. However, information-theoretical work on channel modeling for flash memories has been limited. This paper is an endeavor to survey the noise/disturbs for NAND flash memories, build their corresponding channel models, and explore their special features. Due to space limitation, we omit many details. More details can be found in the full paper [5].

II. Fundamental concepts on flash memories

In this section, we briefly survey the fundamental concepts on NAND flash memories, which are necessary for understanding the channel models of noise and disturbs.

A flash memory cell is a MOS transistor with a floating-gate layer. Its structure is illustrated in Fig. 1 (a). We represent it with the simplified symbol in Fig. 1 (b).

A flash cell stores data by storing charge in its floating-gate layer. And the mount of charge affects its threshold voltage, which is the minimum required voltage added to CG to open the gate. When electrons are stored, the more electrons are trapped in the floating-gate layer, the higher the threshold voltage is. We call the analog value of a cell's threshold voltage its *analog level*. In practice, a cell's analog levels are quantized into discrete values to represent one or more bits. We denote the q discrete levels of a cell by level 0, 1, $\cdots$, $q-1$. When $q = 2$, 4, the flash memory cells are called SLC (Single-Level Cell) and MLC (Multi-Level Cell), respectively. A cell's discrete level is read by comparing it to several reference levels.

TABLE I
Basic notations

Notation	Meaning
q	Number of discrete levels of a cell
W	Number of WLs in a block
$\mathcal{W}_i$	The i-th WL in a block
B	Number of BLs in a block
$\mathcal{B}_j$	The j-th BL in a block
$c_{i,j}$	The cell in the i-th page (corresponding to the i-th WL $\mathcal{W}_i$) and the j-th column (corresponding to the j-th BL $\mathcal{B}_j$)
$V_{i,j}(0)$	The analog level of cell $c_{i,j}$ right after it is programmed
$V_{i,j}(t)$	The analog level of cell $c_{i,j}$ after time t has elapsed since it was programmed
$V_{i,j}$	A simplified notation of $V_{i,j}(t)$
V_i	The average analog level of the i-th discrete level

There are three basic operations on a flash cell: *read*, *write/program* and *erase*, and their typical configurations of voltages are shown in Fig. 1 (c).

The cells in a flash memory are organized as (often tens of thousands or more) blocks, where every block is a two dimensional array. We illustrate a block in Fig. 2 (a). Each row of a block is called a page, which is the unit of read and write operations. A block usually has 32, 64 or more pages, and a page stores thousands of bits. An erase operation is applied to a whole block.

We illustrate the typical voltage configurations for read and write in Fig. 2 (b), (c), respectively.

III. Channel modeling for errors

In this section, we survey the various noise and disturb mechanisms, and present channel models for them. Throughout the paper, a number of notations will be used; for convenience, we summarize them in Table I.

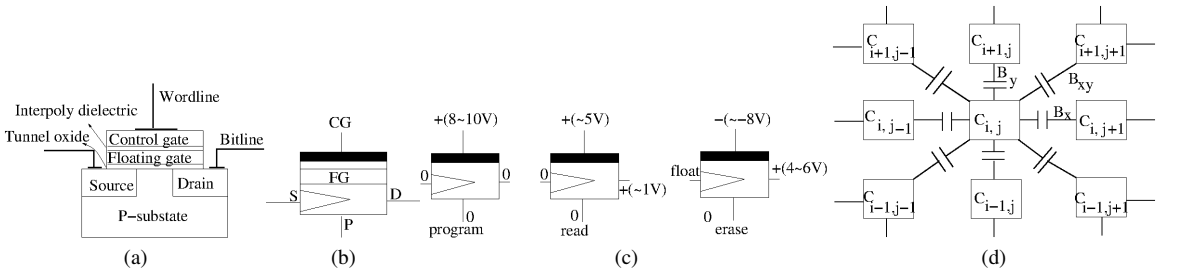


Fig. 1. (a) The structure of NAND flash cell. (b) A symbol for NAND flash cell, where "CG", "FG", "S", "D" and "P" stand for control gate, floating gate, source, drain and P-substrate, respectively. (c) The estimate of voltage biases on S, CG, P and D during flash operations. (d) Illustration of cell-to-cell interference, where B_x, B_y and B_{xy} refer to coupling parameters between neighboring cells in the row direction, the column direction and the diagonal direction, respectively.

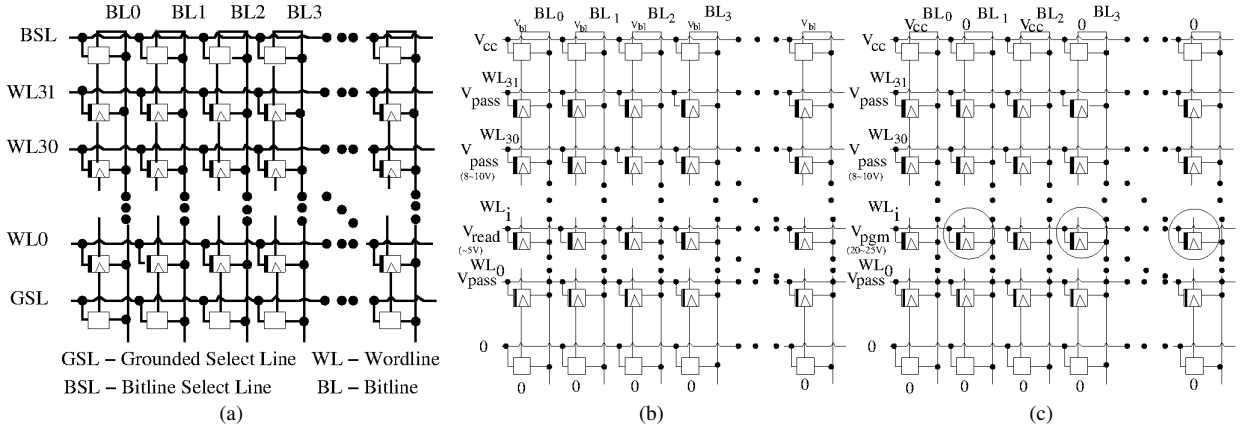


Fig. 2. (a) Structure of a NAND flash memory block, where every vertical wire BL is called a bitline, and every horizontal wire WL is called a wordline. (b) A typical voltage configuration for the read operation. Here the i-th page shown in the figure is being read. (c) A typical voltage configuration for the write operation. Here the i-th page is being programmed. (In particular, the programmed cells are shown in circles.)

A. Inaccurate programming

For cells of q levels, for $k = 0, 1, ..., q-1$, when we program a cell to the k-th level (for $k = 0$ it is the erasure state), let V_k denote the target analog level. For a cell $c_{i,j}$ programmed to the k-th level, let $V_{i,j}(0)$ denote its actual programmed analog level. We call $Z_k = V_{i,j}(0) - V_k$ the programming noise. For simplicity, we assume $Z_k \sim \mathcal{N}(0, \sigma_k)$ has a Gaussian distribution.

B. Retention error

After cells are programmed, the charge stored in them leaks away gradually. When cells experience more program/erase (P/E) cycles, their quality degrades, and charge leakage becomes more serious.

It is reported in [3] that the number of leaked electrons depends on the leaking time t and the initial number of electrons $n(0)$. The number of electrons at time t, $n(t)$, can be modeled as $n(t) = n(0)e^{-\upsilon t}$, where υ is a constant parameter. This parameter υ can vary for cells. So for cell $c_{i,j}$, we use $\upsilon_{i,j}$ to denote its υ.

Further, we use an additive noise Z_{re} to account for the possible deviation from the above model. Based on the linear relationship between the analog level and the number of electrons in the cell's FG, we model $V_{i,j}(t)$ – the analog level of cell $c_{i,j}$ after time t has elapsed since it was programmed – as $V_{i,j}(t) = V_{i,j}(0)e^{-\upsilon_{i,j}t} + Z_{re}$. For simplicity, we assume $Z_{re} \sim \mathcal{N}(0, \sigma_{re})$ has a Gaussian distribution.

C. Cell-to-cell interference

Due to the parasitic capacitance-coupling effect between neighboring cells, the analog level we can read from a cell depends on the cell's own level and those of its neighboring cells. For this reason, we differentiate the concept of *intrinsic analog level* from the *extrinsic analog level* of a cell. By intrinsic (respectively, extrinsic) analog level, we refer to the cell's analog level when there is no (respectively, there is) interference from neighboring cells. For cell $c_{i,j}$, we use $\hat{V}_{i,j}$ to denote its intrinsic analog level, and use $V_{i,j}$ to denote its extrinsic analog level.

A model for cell-to-cell interference is proposed in [2], as shown in Fig. 1 (d). We model the effect of cell-to-cell interference as

$$\begin{aligned} V_{i,j} &= \hat{V}_{i,j} + B_x(\hat{V}_{i,j-1} + \hat{V}_{i,j+1}) + B_y(\hat{V}_{i-1,j} \\ &+ \hat{V}_{i+1,j}) + B_{xy}(\hat{V}_{i-1,j+1} + \hat{V}_{i-1,j-1} \\ &+ \hat{V}_{i+1,j+1} + \hat{V}_{i+1,j-1}) + Z_{inter}, \end{aligned} \quad (1)$$

where the noise Z_{inter} accounts for the possible deviation from the above linear model. For simplicity, we assume $Z_{inter} \sim \mathcal{N}(0, \sigma_{inter})$.

Note that cell-to-cell interference can be compensated by adaptive programming.

D. Read disturb

When the k-th page is read (for $k \in \{0, 1, \cdots, W-1\}$), the other pages are softly programmed due to the voltage V_{pass} added on their control gates. For a disturbed cell $c_{i,j}$ (for $i \in \{0, 1, \cdots, W-1\} - \{k\}$ and $j \in \{0, 1, \cdots, B-1\}$), we denote its analog level before read disturb by $V_{i,j}$, and denote that after read disturb by $V'_{i,j}$. We model read disturb as

$$V'_{i,j} = V_{i,j} + \gamma^{rd}_{i,j} + Z_{rd}, \quad (2)$$

where $\gamma^{rd}_{i,j}$ is a parameter that depends on the time interval for the read operation, the strength of the electrical field between cell $c_{i,j}$'s control gate and P-substrate, and the cell's capacitance; and the noise Z_{rd} accounts for the possible deviation from the above simple linear model. For simplicity, we assume $Z_{rd} \sim \mathcal{N}(0, \sigma_{rd})$ has a Gaussian distribution.

E. Program disturb

When the i-th page is programmed (for $i \in \{0, 1, \cdots, W-1\}$), let $\mathcal{S} \subseteq \{0, 1, \cdots, B-1\}$ denote the indices of those programmed cells in that page. The unprogrammed cells in that page, which have indices in $\{0, 1, \cdots, B-1\} - \mathcal{S}$, will be softly programmed, which is called program disturb. For a disturbed cell $c_{i,j}$ (for $j \in \{0, 1, \cdots, B-1\} - \mathcal{S}$), we model program disturb similarly to read disturb as: $V'_{i,j} = V_{i,j} + \gamma^{prod}_{i,j} + Z_{prod}$, where $\gamma^{prod}_{i,j}$ has a similar meaning as in function (2) (but it may have a different value due to changed parameters such as the time interval for programming). Here the noise Z_{prod} accounts for the possible deviation from the above simple linear model. For simplicity, we assume $Z_{prod} \sim \mathcal{N}(0, \sigma_{prod})$ has a Gaussian distribution.

F. Pass disturb

When the k-th page is programmed, the other pages are softly programmed due to the voltage V_{pass} added on their control gates. The process is similar to read disturb, and we model it as $V'_{i,j} = V_{i,j} + \gamma^{pasd}_{i,j} + Z_{pasd}$, where $\gamma^{pasd}_{i,j}$ has a similar meaning as in function (2) (but with possible different values, as for program disturb). And as before, Z_{pasd} accounts for the additive noise term, and for simplicity we assume $Z_{pasd} \sim \mathcal{N}(0, \sigma_{pasd})$.

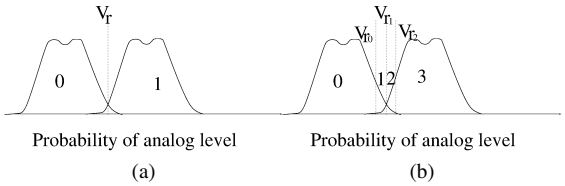


Fig. 3. (a) SLC with one reference level; (b) SLC with three reference levels (i.e., three sub-thresholds)

IV. Capacity Analysis for Flash Memories

In this section, we focus on flash memory capacity's special features: how the storage capacity degrades with flash operations, the trade-off between instantaneous capacity and read disturbs when sub-thresholds are used, and the importance of dynamic thresholds.

A. Basic model for write and read operations

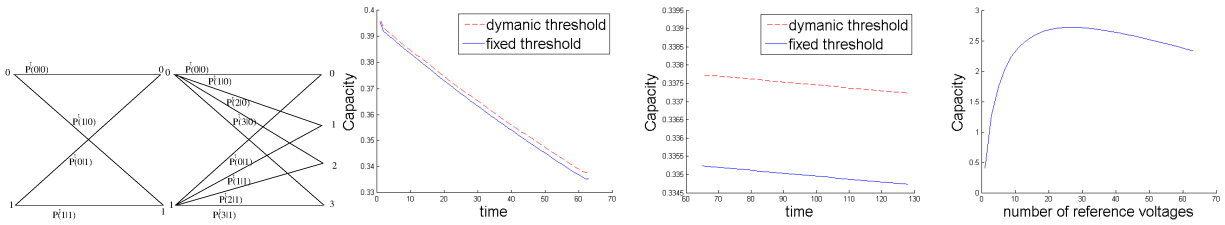
In this section, we consider the following simplistic write/read model for an SLC block of W pages. We first program the W pages sequentially from $\mathcal{W}_0$ to $\mathcal{W}_{W-1}$, and we assume every cell has an equal likelihood of being programmed to 0 or 1. We then have $n-1$ rounds of reading; in each round, we read the pages $\mathcal{W}_0, \mathcal{W}_1, \cdots, \mathcal{W}_{W-1}$ sequentially. Although the noise in cells is correlated (e.g., via inter-cell interference), when we compute capacity, we treat them as having independent noise. (This is, of course, a restrictive model for capacity.) Furthermore, when we analyze capacity, we assume B (the number of cells in a page) approaches infinity.

The above simple model for SLC can be extended to more complex cases. However, the basic observations derived here still hold for more general cases.

B. Capacity degradation with flash operations

In this subsection, we analyze how the analog-level distribution of cells changes with more and more write/read operations (under the model introduced earlier), and compute the corresponding storage capacity. (Note that P/E cycles degrade cells' quality, which is another source of capacity degradation, but we do not consider it here.)

We assume that the W writes and $(n-1)W$ reads in our model happen at time $0, 1, \cdots, W-1, W, W+1, \cdots, (n-1)W-1$, respectively. So for the i-th page $\mathcal{W}_i$, it was programmed at time i and was read at time $i+W, i+2W, \cdots, i+(n-1)W$. For a cell $c_{i,j}$, which is intended to be programmed to $V_k \in \{V_0, V_1\}$, let $V_{i,j}(k,t)$ be the analog level of $c_{i,j}$ after the t-th operation. We first present the recursive formula for $V_{i,j}(0,t)$



(a) (b) (c) (d) (e)

Fig. 4. (a) Noise channel model with single reference voltage. (b) Noise channel model with three reference voltages. (c), (d) Comparison between the fixed-reference-voltage scheme and the dynamic-reference-voltage scheme during sequential write operations ((c)) and sequential read operations ((d)). Here the x-axis is the discrete time when write or read operations happen, the y-axis is the storage capacity, the solid curve corresponds to $\mathcal{C}_2(t)$ $t = 0, 1, \cdots, 127$ for fixed reference voltage, and the dashed curve corresponds to $\mathcal{C}_2^d(t)$ for dynamic reference voltage. (e) The trade-off between the number of sub-thresholds (reference voltages) and capacity.

below. (For simplicity, we assume $0 < i < W-1$ to avoid boundary cases.) $V_{i,j}(0,t) =$

$$\begin{cases} V_0 + Z_0 & t < i, \\ V_{i,j}(k,t-1) + Z_{prod} & t = i, \\ V_{i,j}(k,t-1) + B_y V_{i+1,j}(k_1,t) & \\ +B_{xy}(V_{i+1,j-1}(k_2,t)+ & \\ V_{i+1,j+1}(k_3,t)) + Z_{inter} & \\ +\gamma_{i,j}^{pasd} + Z_{pasd} & t = i+1, \\ V_{i,j}(k,t-1) + \gamma_{i,j}^{pasd} + Z_{pasd} & t \in [i+2, W-1], \\ V_{i,j}(k,t-1) & t = W + mi, \\ V_{i,j}(k,t-1) + \gamma_{i,j}^{rd} + Z_{rd} & \text{otherwise.} \end{cases} \tag{3}$$

In order to obtain the probability distribution for $V_{i,j}(k,t)$, we make the following assumptions for simplicity: for cell-to-cell interference, B_{xy} is negligible compared to B_y; given any i and j, $\gamma_{i,j}^{pasd}$ and $\gamma_{i,j}^{rd}$ are constant over time, so they have no effect on $V_{i,j}(k,t)$'s probability distribution; $V_{i,j}(0,t)$ is independent of j, and we write it as $V_i(k,t)$ sometimes. Thus, $V_{i,j}(0,t) \sim \mathcal{N}(V_0(1+B_y), \sigma_i(0,t))$ or $V_{i,j}(0,t) \sim \mathcal{N}(V_0 + B_y V_1, \sigma_i(0,t))$ with equal probability, where $\sigma_i^2(0,t) =$

$$\begin{cases} \sigma_0^2 & t < i, \\ \sigma_i^2(0,t-1) + \sigma_{prod}^2 & t = i, \\ \sigma_i^2(0,t-1) + (B_y \sigma_{i+1}(k,t))^2 & \\ +\sigma_{pasd}^2 + \sigma_{inter}^2 & t = i+1,\ k \in \{0,1\}, \\ \sigma_i^2(0,t-1) + \sigma_{pasd}^2 & t \in [i+2, W-1], \\ \sigma_i^2(0,t-1) & t = W + mi, \\ \sigma_i^2(0,t-1) + \sigma_{rd}^2 & \text{otherwise.} \end{cases} \tag{4}$$

Suppose the reference voltage for reading (that separates the two levels) is V_r. Also suppose $c_{i,j}$ represents 1 if $V_{i,j}(k,t) > V_r$ and 0 otherwise (see Fig. 3 (a)). $P^t(Y|X)$ $(Y, X \in \{0,1\})$ is the probability that $c_{i,j}$ (which is intended to be programmed to $V_X \in \{V_0, V_1\}$) represents data Y after t operations. Thus,

$$P^t(1|0) = \frac{1}{2}\left(Q\left(\frac{V_r - V_0(1+B_y)}{\sigma_i(0,t)}\right) + Q\left(\frac{V_r - (V_0 + B_y V_1)}{\sigma_i(0,t)}\right)\right), \tag{5}$$

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

A similar process and assumptions as above can be applied to $V_{i,j}(1,t)$, and due to space limitation we skip it here. Naturally, $V_{i,j}(k,t)$ (for $t = 0, 1, \cdots$) form a Markov chain.

Let $\mathcal{X} = \mathcal{Y} = \{0,1\}$, and our channel model is $\mathbb{P} = (\mathcal{X}, \mathcal{Y}, P^t(Y|X))$. An example is presented in Fig. 4 (a).

Thus, the capacity of $\mathcal{W}_i$ after the t-th operation is $\mathcal{C}_i(t) = 1 - \frac{1}{2} \sum_{X,Y \in \{0,1\}} P^t(Y|X) \log_2 \frac{P^t(Y|X)}{\sum_X P^t(Y|X)}$. Due to data-processing inequality, we conclude that $\mathcal{C}_i(t+1) \le \mathcal{C}_i(t)$ for $t \in \mathbb{N}$.

Let V_r be 1.4 and the remaining parameters be fixed values in Table II. We numerically calculate $\mathcal{C}_2(t)$ for $t = 0, 1, \cdots, 127$ as shown by the solid line of Fig. 4 (c) and (d). We can clearly see that the storage capacity $\mathcal{C}_2(t)$ decreases with more and more operations.

TABLE II
PARAMETERS USED IN COMPUTING CAPACITY

σ_0^2	σ_1^2	σ_{inter}^2	σ_{pasd}^2	σ_{rd}^2	σ_{prod}^2
2	1	9×10^{-3}	5×10^{-3}	10^{-4}	8×10^{-3}
$(B_y \sigma_{i+1}(0,t))^2$	$(B_y \sigma_{i+1}(1,t))^2$	B_y	V_0	V_1	W
10^{-3}	10^{-3}	0.01	0	2.5	64

C. The impact of sub-threshold for flash capacity

Sub-thresholds [4], [6] are the scheme that there are multiple reference voltages between adjacent discrete levels, e.g., Fig. 3 (b). Sub-thresholds can obtain more soft information on cell levels, and improve coding performance (e.g., for LDPC codes). However, sub-thresholds lead to more reads and read disturbs. Although having sub-thresholds can increase the precision of reading at the moment, the additional noise caused by read disturbs also distorts cell levels and is accumulated for future reading. Therefore, there is an optimal way to

set sub-thresholds to maximize capacity over the flash memory's lifetime (which is not necessarily the more sub-thresholds the better).

In this subsection, we explore the impact of sub-thresholds for flash capacity, focusing on the trade-off between read precision and read disturbs. (How to set the positions of sub-thresholds is beyond the scope of this paper.) Consider SLC, for $l = 1, 3, 5, \cdots$, let $\mathcal{V}(l) = \{V_{r_0}, V_{r_1}, \cdots, V_{r_{l-1}}\}$ denote the set of l sub-thresholds we use for separating discrete level 0 from level 1. Let V_r be the single sub-threshold when $l = 1$. Let L be the maximum number of sub-thresholds used. Let $\delta = \frac{V_1 - V_0}{2L}$. We set $\mathcal{V}(l-1)$ as $\{V_{r_k} = V_r - (\lfloor \frac{l}{2} \rfloor - k)\delta | k = 0, 1, \cdots, l-2\}$ in this paper.

An SLC with l sub-thresholds can be modeled by a 2-input $(l+1)$-output channel, where the $(l+1)$ outputs– $0, 1, \cdots, l$– corresponds to $l+1$ regions separated by the l sub-thresholds. Let $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, \cdots, l\}$, and $P^t(Y|X)$ ($X \in \mathcal{X}$, $Y \in \mathcal{Y}$) be the probability that $c_{i,j}$ (which is intended to be programmed to $V_X \in \{V_0, V_1\}$) is read as $Y \in \mathcal{Y}$ after t operations. $P^t(Y|X)$ can be obtained in a similar way as before. (With the same setting and the similar analysis of the previous subsection, we obtain that $V_{i,j}(k,t) = V_{i,j}(k,t-1) + \gamma_{i,j}^{rd} + l \times Z_{rd}$ when it is suffered from read disturbs. The remaining cases of $V_{i,j}(k,t)$ are the same as those of the previous subsection.) Our proposed channel model is $\mathbb{P}^m = (\mathcal{X}, \mathcal{Y}, P^t(Y|X))$. Fig. 4 (b) presents an illustration of the channel model with three sub-thresholds.

Let the capacity of the i-th page $\mathcal{W}_i$ (with l sub-thresholds) after t write/read operations be $\mathcal{C}_i(l,t) = I(X;Y)$. With parameters listed in Table II except $\sigma_0^2 = \sigma_1^2 = 1$, we present $\mathcal{C}_2(l, 500)$ for different l in Fig. 4 (e). (The capacity for other values of i and t has similar shapes.) As shown in Fig. 4(e), there is a trade-off between the number of sub-thresholds and storage capacity. When there are too many sub-thresholds, the impact of read disturbs becomes dominant, and the corresponding capacity decreases.

D. Dynamically adjust reference threshold voltages

It can be seen from the error models that flash disturbs are highly correlated (both in time and space), and the noise has a tendency to be non-symmetric (e.g., disturbs tend to increase cell levels). Therefore, it is important to set reference voltages adaptively over time to reduce errors and maximize capacity. Such a scheme is called *dynamic threshold*, and has been studied before [7], [8]. In this subsection, we study how dynamic thresholds can help improve storage capacity based on our flash models.

Let $V_r(t)$ be the reference voltage we adaptively choose for the t-th operation. Let $ER_k(t)$ denote the error probability of quantizing $V_i(k,t)$ (for simplicity, we assume $t \geq i$). Therefore $ER_k(t) =$

$$\begin{cases} \frac{1}{2}(Q(\frac{V_r(t) - V_0(1+B_y)}{\sigma_i(0,t)}) + Q(\frac{V_r - (V_0 + B_y V_1)}{\sigma_i(0,t)})) & k = 0, \\ 1 - \frac{1}{2}(Q(\frac{V_r(t) - V_1(1+B_y)}{\sigma_i(1,t)}) + Q(\frac{V_r - (V_1 + B_y V_0)}{\sigma_i(1,t)})) & k = 1. \end{cases} \quad (6)$$

Assume that k is uniformly distributed over $\{0, 1\}$. Thus the total quantizing error probability for the t-th operation is $TER(t) = \frac{1}{2} +$

$$\frac{1}{4}(Q(\frac{V_r(t) - V_0(1+B_y)}{\sigma_i(0,t)}) + Q(\frac{V_r(t) - (V_0 + B_y V_1)}{\sigma_i(0,t)})) - \frac{1}{4}(Q(\frac{V_r(t) - V_1(1+B_y)}{\sigma_i(1,t)}) + Q(\frac{V_r(t) - (V_1 + B_y V_0)}{\sigma_i(1,t)})).$$

The objective of dynamic reference voltage is to choose $V_r(t)$ such that $TER(t)$ is minimized, therefore $V_r(t)$ should satisfy

$$\frac{\partial(Q(\frac{V_r(t) - V_0(1+B_y)}{\sigma_i(0,t)}) + Q(\frac{V_r(t) - (V_0 + B_y V_1)}{\sigma_i(0,t)}))}{\partial V_r(t)} = \frac{\partial(Q(\frac{V_r(t) - V_1(1+B_y)}{\sigma_i(1,t)}) + Q(\frac{V_r(t) - (V_1 + B_y V_0)}{\sigma_i(1,t)}))}{\partial V_r(t)}. \quad (7)$$

Similarly, we can obtain the probability distribution of $V_{i,j}(k,t)$, and $P^t(Y|X)$ for $X \in \mathcal{X} = \{0, 1\}$ and $Y \in \mathcal{Y} = \{0, 1\}$. The channel model of $\mathcal{W}_i$ for dynamic reference voltages is denoted by $\mathbb{P}^d = (\mathcal{X}, \mathcal{Y}, P^t(Y|X))$, and its capacity is $\mathcal{C}_i^d(t) = I(X;Y)$. With parameters of Table II, we numerically compute $\mathcal{C}_2^d(t)$, which is shown by the dashed curve in Fig. 4 (c) and (d). We can see that after dynamically adjusting the reference voltages, the channel (with quantization) becomes less noisy and the storage capacity increases correspondingly.

V. Acknowledgment

This work was partially supported by the University Research Program of LSI Corporation.

References

[1] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Error patterns in MLC NAND flash memory: measurement, characterization, and analysis," in proceedings of the Conference on Design, Automation and Test in Europe, DATE'12.

[2] Y. Cai, O. Mutlu, E. F. Haratsch, and K. Mai, "Program Interference in MLC NAND Flash Memory: Characterization, Modeling, and Application," ICCD'13.

[3] G. Crisenza, C. Clementi, G. Ghidini, and M. Tosi, "Floating Gate Memroies Reliability," *in Proc. ESREF*, vol. 8, pp. 177–187, 1992.

[4] G. Dong, N. Xie, and T. Zhang, "On the Use of Soft-Decision Error-Correction Codes in NAND Flash Memory," *IEEE Trans. on Circuits and Systems*, vol. 58, no. 2, pp. 429–439, 2011.

[5] Q. Li, A. Jiang, and E. F. Haratsch, "Noise Modeling and Capacity Analysis for NAND Flash Memories," available at faculty.cs.tamu.edu/ajiang/Publications/2014/noiseModelFull.pdf.

[6] J. Wang, T. Courtade, H. Shankar, and R. D. Wesel, "Soft Information for LDPC Decoding in Flash: Mutual-Information Optimized Quantization," in proceedings of Globecom 2011.

[7] W. Xu and T. Zhang, "A time-aware fault tolerance scheme to improve reliability of multilevel phase-change memory in the presence of significant resistance drift," *IEEE Trans. VLSI Systems*, vol. 19, no. 8, pp. 1357–1367, 2011.

[8] H. Zhou, A. A. Jiang, and J. Bruck, "Error-Correcting Schemes with Dynamic Thresholds in Nonvolatile Memories," ISIT, 2011.