Performance of Flash Memories with Different Binary Labelings: A Multi-User Perspective

Pengfei Huang*, **Paul H. Siegel***, and **Eitan Yaakobi**‡
*Electrical and Computer Engineering Dept., University of California, San Diego, La Jolla, CA 92093 U.S.A
‡Computer Science Dept., Technion – Israel Institute of Technology, Haifa 32000, Israel
{*pehuang,psiegel*}@*ucsd.edu*, *yaakobi@cs.technion.ac.il*

***Abstract*—In this work, we study the performance of different decoding schemes for multilevel flash memories where each page in every block is encoded independently. We focus on the multi-level cell (MLC) flash memory, which is modeled as a two-user multiple access channel suffering from asymmetric noise. The uniform rate regions and sum rates of Treating Interference as Noise (TIN) decoding and Successive Cancelation (SC) decoding are investigated for a Program/Erase (P/E) cycling model and a data retention model. We examine the effect of different binary labelings of the cell levels, as well as the impact of further quantization of the memory output (i.e., additional read thresholds). Finally, we extend our analysis to the three-level cell (TLC) flash memory.**

I. Introduction

NAND flash memory is a promising non-volatile data storage medium, and has been widely used in customer electronics as well as enterprise data centers. It has many advantages over traditional magnetic recording, e.g., higher read throughput and less power consumption. The basic storage unit in a NAND flash memory is a floating-gate transistor referred to as a cell. The voltage levels of a cell can be adjusted by a program operation and are used to represent the stored data. The cells typically have 2, 4, and 8 voltage levels (1, 2, and 3 bits/cell respectively) and are referred to as single-level cell (SLC), multi-level cell (MLC), and three-level cell (TLC) respectively. Cells are grouped into pages, which are grouped into blocks.

The two bits belonging to a multi-level cell (MLC) are separately mapped to two pages. The most significant bit (MSB) is mapped to the lower page while the least significant bit (LSB) is mapped to the upper page. We represent the four voltage levels in MLC flash memory as A_0, A_1, A_2, and A_3 in increasing order of voltage levels. The corresponding 2-bit patterns written to the lower page (MSB) and upper page (LSB) are '11', '10', '00', and '01', which are called *Gray* labeling. Similarly, the three bits belonging to a three-level cell (TLC) are separately mapped to three pages. We refer to the first bit as the most significant bit (MSB), the second bit as the center significant bit (CSB), and the third bit as the least significant bit (LSB). We represent the eight voltage levels in TLC flash memory as B_0, B_1, B_2, B_3, B_4, B_5, B_6, and B_7 in increasing order of voltage levels. The corresponding 3-bit patterns for the MSB, CSB, and LSB also use Gray labeling and are '111', '110', '100', '101', '001', '000', '010', and '011', respectively.

The channel characterization of flash memory is important for understanding the fundamental density limits as well as designing effective error-correcting codes (ECC). Many experiments have shown that the noise in flash memories is asymmetric [2], [8]. In [6], a mixed normal-Laplace model was proposed and validated to capture this asymmetry feature well. The capacity of MLC flash memory was analyzed in [9] by modeling MLC flash memory as a 4-ary input point-to-point channel with additive white Gaussian noise. Similarly, the capacity of TLC flash memory was recently studied in [7] by considering TLC flash memory as an 8-ary input point-to-point channel with asymmetric mixed normal-Laplace noise.

However, in current MLC (or TLC) flash memories, 2 (or 3) bits in a cell are mapped to 2 (or 3) pages, which are actually encoded independently. Hence, previous works [7], [9] based on a point-to-point channel model only give an upper bound on the sum rate of all pages. In this paper, we take a different perspective, and model the flash memory as a multi-user system, where each user corresponds to a page and is encoded independently. To the best of our knowledge, this is the first time that flash memories have been studied in this framework.

Our goal is to study the fundamental performance limits of flash memories with different decoding schemes. Here, we consider both low-complexity Treating Interference as Noise (TIN) decoding and relatively high-complexity Successive Cancelation (SC) decoding, and derive the conditions when the sum rate of TIN decoding equals that of SC decoding. Then, achievable rate regions and sum rates of both decoding schemes are determined for different channel models, represented by channel transition matrices from cell voltage levels to quantized readback outputs. The effect of different binary labelings of the cell levels is also studied, and the optimal labeling for each decoding scheme and channel model is identified. It is shown that TIN and SC decodings both outperform the Default Setting (DS) decoding, a model of current flash memory technology, which uses Gray labeling of cell levels, along with separate quantization and decoding of each page.

The remainder of the paper is organized as follows. In Section II, we model the MLC flash memory as a multi-user system. In Section III, we introduce three different decoding schemes. In Section IV, we study the decoding performance of MLC flash memory for different channel models. Different labelings and multiple reads are discussed. In Section V, we further investigate TLC flash memories. We conclude the paper in Section VI. Throughout the paper, we follow the notations in [4]. Due to space constraints, we omit some proofs, which can be found online in the longer version of this paper [5].

II. System Model for Flash Memories

We model a flash memory as a k-user multiple access channel with k independent inputs $X_1, \cdots, X_k$, and one output Y ($k=2$ for MLC flash, and $k=3$ for TLC flash).

Specifically, the readback signal $\tilde{Y} \in \mathbb{R}$ in a flash memory

978-1-5090-1806-2/16/$31.00 ©2016 IEEE

is expressed as

$$\tilde{Y} = \sigma(X_1, \cdots, X_k) + Z, \quad (1)$$

where $X_1, \cdots, X_k \in \{0,1\}$ represent data from k independent pages, $Z \in \mathbb{R}$ stands for the asymmetric noise, and σ maps an input $(x_1, \cdots, x_k)$ to a voltage level v. More specifically, σ is a bijective mapping from the set $\mathcal{T}$ which consists of all k-length binary strings to the set $\mathcal{V}$ which consists of 2^k voltage level values. For $k = 2$ (MLC flash), $\mathcal{T}_{MLC} = \{11, 10, 01, 00\}$ and $\mathcal{V}_{MLC} = \{A_0, A_1, A_2, A_3\}$. By a slight abuse of notation, we write the mapping σ as a vector $\sigma = (w_0, w_1, w_2, w_3)$ (where the w_i, $i = 0,1,2,3$, represent the full set of possible 2-tuples) to represent the mapping $\sigma(w_i) = A_i$ for $i = 0,1,2,3$. For example, $\sigma = (11,10,00,01)$ means $\sigma(11) = A_0$, $\sigma(10) = A_1$, $\sigma(00) = A_2$, and $\sigma(01) = A_3$. Similarly, for $k = 3$ (TLC flash), $\mathcal{T}_{TLC} = \{111, 110, 101, 100, 011, 010, 001, 000\}$ and $\mathcal{V}_{TLC} = \{B_0, B_1, \ldots, B_7\}$. We write $\sigma = (w_0, w_1, \ldots, w_7)$ (where the w_i, $i = 0,1,\ldots,7$, represent the full set of possible 3-tuples) to represent the mapping $\sigma(w_i) = B_i$ for $i = 0,1,\ldots,7$.

During the readback process, a quantizer Q is used to quantize $\tilde{Y}$ to obtain an output Y, i.e., $Y = Q(\tilde{Y})$, where the function $Q(\cdot)$ is a mapping from $\mathbb{R}$ to a finite alphabet set $\mathcal{Y} = \{s_0, s_1, \ldots, s_{q-1}\}$ of cardinality q.

In this paper, we refer to a *mapping* σ as a *labeling*.

III. Decoding Schemes for MLC Flash Memories

In this section, we investigate three decoding schemes for MLC flash memories.

Given a labeling σ and a quantizer Q, the MLC flash memory channel can be modeled as a 2-user discrete memoryless multiple access channel $\mathcal{W}_{MLC}$: $(\mathcal{X} \times \mathcal{X}, p(y|x_1,x_2), \mathcal{Y})$, where $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{s_0, s_1, \ldots, s_{q-1}\}$, and $p(y|x_1,x_2)$ is the transition probability for any $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$. Define $P(Y = y|X_1 = x_1, X_2 = x_2) \overset{\text{def}}{=} p_{BD(x_1,x_2),y}$, where $BD(\cdot)$ is a function that converts a binary string into its decimal value, e.g., $P(Y = s_0|X_1 = 1, X_2 = 0) = p_{2,s_0}$.

The capacity region of this multiple access channel is fully characterized [3], [4]. However, in this paper, to make analysis simple yet representative, we are only interested in the *uniform* rate region for different decoding schemes. For a multiple access channel, the uniform rate region is the achievable region corresponding to the case that the input distributions are uniform. For other input distributions, the analysis is similar.

For a channel $\mathcal{W}_{MLC}$, the **Treating Interference as Noise (TIN)** decoding scheme decodes X_1 and X_2 independently based on Y [3], [4]. Its uniform rate region $\mathscr{R}^{TIN}$ for lower page X_1 and upper page X_2 is the set of all pairs (R_1, R_2) such that 1) $0 \leqslant R_1 \leqslant I(X_1;Y)$ and 2) $0 \leqslant R_2 \leqslant I(X_2;Y)$. In $\mathscr{R}^{TIN}$, the sum rate[1] $r_s^{TIN} = \max\{R_1 + R_2 : (R_1, R_2) \in \mathscr{R}^{TIN}\} = I(X_1;Y) + I(X_2;Y)$.

For a channel $\mathcal{W}_{MLC}$, the **Successive Cancelation (SC)** decoding scheme decodes X_1 and X_2 in some order based on Y [3], [4]. Its uniform rate region $\mathscr{R}^{SC}$ for lower page X_1 and upper page X_2 is the set of all pairs (R_1, R_2) such that 1) $R_1 \leqslant I(X_1;Y|X_2)$, 2) $R_2 \leqslant I(X_2;Y|X_1)$, and 3) $R_1 + R_2 \leqslant I(X_1,X_2;Y)$. In $\mathscr{R}^{SC}$, the sum rate $r_s^{SC} = \max\{R_1 + R_2 : (R_1, R_2) \in \mathscr{R}^{SC}\} = I(X_1,X_2;Y)$.

[1]In this paper, for the sake of brevity, we use the term "sum rate" to represent the maximum sum rate in the corresponding rate region.

Remark 1 For TIN decoding, X_1 and X_2 are decoded independently and can be implemented in parallel. However, for SC decoding, X_1 and X_2 are decoded in a certain order. In general, TIN decoding is preferred for its low decoding complexity, but the uniform rate region $\mathscr{R}^{TIN} \subseteq \mathscr{R}^{SC}$ and the sum rate $r_s^{TIN} \leqslant r_s^{SC}$. □

The following theorem gives the condition when the sum rates of TIN decoding and SC decoding are the same.

Theorem 1. *For a channel $\mathcal{W}_{MLC}$, the sum rates satisfy $r_s^{TIN} \leqslant r_s^{SC}$ with equality if and only if $p_{3,s_j}p_{0,s_j} = p_{2,s_j}p_{1,s_j}$ for all $j = 0,1,\ldots,q-1$. If $r_s^{TIN} = r_s^{SC}$, then $\mathscr{R}^{TIN} = \mathscr{R}^{SC}$ and the rate region is a rectangle.*

Proof: We bound $r_s^{SC} - r_s^{TIN}$ as

$$\begin{aligned} r_s^{SC} - r_s^{TIN} &= I(X_1,X_2;Y) - I(X_1;Y) - I(X_2;Y) \\ &= I(X_2;Y|X_1) - I(X_2;Y) \\ &= H(X_2|X_1) - H(X_2|X_1,Y) - \Big(H(X_2) - H(X_2|Y)\Big) \\ &\overset{(a)}{=} I(X_1;X_2|Y) = \sum_{j=0}^{q-1}\Big(\sum_{i=0}^{3}\frac{p_{i,s_j}}{4}\Big) I(X_1;X_2|Y=s_j) \geqslant 0, \end{aligned} \quad (2)$$

where in step (a) we use $H(X_2|X_1) = H(X_2)$ which follows from the fact that X_1 and X_2 are independent. Noting that $I(X_1;X_2|Y) \geqslant 0$, with equality if and only if X_1 and X_2 are conditionally independent given $Y = s_j$, i.e., $P(X_1,X_2|Y=s_j) = P(X_1|Y=s_j)P(X_2|Y=s_j)$, we conclude that $r_s^{TIN} = r_s^{SC}$ if and only if $p_{3,s_j}p_{0,s_j} = p_{2,s_j}p_{1,s_j}$ for all $j = 0,1,\ldots,q-1$.

Finally, assuming $r_s^{TIN} = r_s^{SC}$, i.e., $I(X_1;Y) + I(X_2;Y) = I(X_1,X_2;Y)$, since $I(X_1,X_2;Y) = I(X_1;Y) + I(X_2;Y|X_1) = I(X_2;Y) + I(X_1;Y|X_2)$, we have $I(X_1;Y) = I(X_1;Y|X_2)$ and $I(X_2;Y) = I(X_2;Y|X_1)$, which means $\mathscr{R}^{TIN}$=$\mathscr{R}^{SC}$. ■

The third decoding scheme we consider is modeled upon current MLC flash memory technology. The Gray labeling $\sigma = (11,10,00,01)$ is used to map binary inputs (X_1, X_2) to cell levels. The lower page X_1 and upper page X_2 are decoded independently according to different quantizations and a total of three reads are employed: 1) to decode X_1, $\tilde{Y}$ is quantized by one read between voltage levels A_1 and A_2, and its corresponding output is Y_1; 2) to decode X_2, $\tilde{Y}$ is quantized by two reads between voltage levels A_0 and A_1, and between A_2 and A_3, respectively, and its corresponding output is Y_2. We call this **Default Setting (DS)** decoding, and it is used as our *baseline* decoding scheme. Its uniform rate region $\mathscr{R}^{DS}$ for lower page X_1 and upper page X_2 is the set of all pairs (R_1, R_2) such that 1) $0 \leqslant R_1 \leqslant I(X_1;Y_1)$ and 2) $0 \leqslant R_2 \leqslant I(X_2;Y_2)$. In $\mathscr{R}^{DS}$, the sum rate $r_s^{DS} = \max\{R_1 + R_2 : (R_1, R_2) \in \mathscr{R}^{DS}\} = I(X_1;Y_1) + I(X_2;Y_2)$.

IV. Performance of MLC Flash Memory with Different Decoding Schemes and Labelings

In this section, we study the uniform rate region and sum rate of MLC flash memories with different decoding schemes and labelings for different channel models which are characterized by channel transition matrices from voltage levels to quantized outputs. Both a *Program/Erase (P/E) cycling* model and a *data retention* model are considered. In the following, we will use the function $f(x) = x\log_2 x$.

TABLE I

TRANSITION MATRIX $p^E_{MLC}(y|v)$ AT EARLY STAGE OF P/E CYCLING

V	Inputs: (X_1, X_2)			Output: Y			
Levels	Gray	NO	EO	s_0	s_1	s_2	s_3
A_0	(11)	(11)	(11)	a_1	$1-a_1$	0	0
A_1	(10)	(10)	(00)	0	b_1	$1-b_1$	0
A_2	(00)	(01)	(01)	0	0	c_1	$1-c_1$
A_3	(01)	(00)	(10)	0	0	0	1

TABLE II

UNIFORM RATE REGIONS AND SUM RATES OF DS, TIN, AND SC DECODINGS AT EARLY STAGE OF P/E CYCLING FOR MLC FLASH

Gray	DS	$\mathscr{R}_G^{DS}$	$0 \leqslant R_1 \leqslant \lambda_1$, $0 \leqslant R_2 \leqslant \lambda_2$	$r_{s(G)}^{DS} = \lambda_1 + \lambda_2$
	TIN	$\mathscr{R}_G^{TIN}$	$0 \leqslant R_1 \leqslant \lambda_3$, $0 \leqslant R_2 \leqslant \lambda_4$	$r_{s(G)}^{TIN} = \lambda_3 + \lambda_4$
	SC	$\mathscr{R}_G^{SC}$	$0 \leqslant R_1 \leqslant \lambda_3$, $0 \leqslant R_2 \leqslant \lambda_4$	$r_{s(G)}^{SC} = \lambda_3 + \lambda_4$
NO	TIN	$\mathscr{R}_{NO}^{TIN}$	$0 \leqslant R_1 \leqslant \lambda_3$, $0 \leqslant R_2 \leqslant \lambda_5$	$r_{s(NO)}^{TIN} = \lambda_3 + \lambda_5$
	SC	$\mathscr{R}_{NO}^{SC}$	$0 \leqslant R_1 \leqslant 1$, $0 \leqslant R_2 \leqslant \lambda_4$, $R_1 + R_2 \leqslant 1 + \lambda_5$	$r_{s(NO)}^{SC} = 1 + \lambda_5$
EO	TIN	$\mathscr{R}_{EO}^{TIN}$	$0 \leqslant R_1 \leqslant \lambda_4$, $0 \leqslant R_2 \leqslant \lambda_5$	$r_{s(EO)}^{TIN} = \lambda_4 + \lambda_5$
	SC	$\mathscr{R}_{EO}^{SC}$	$0 \leqslant R_1 \leqslant 1$, $0 \leqslant R_2 \leqslant \lambda_3$, $R_1 + R_2 \leqslant 1 + \lambda_5$	$r_{s(EO)}^{SC} = 1 + \lambda_5$

Definition 2 *For MLC flash memory, the mapping* $\sigma_G = (11, 10, 00, 01)$ *is called Gray labeling,* $\sigma_{NO} = (11, 10, 01, 00)$ *is called Natural Order (NO) labeling, and* $\sigma_{EO} = (11, 00, 01, 10)$ *is called Even Odd (EO) labeling.*

For each labeling, the mapping between inputs $(X_1, X_2) \in \mathcal{T}_{MLC}$ and voltage levels $V \in \mathcal{V}_{MLC}$ is shown in Table I.

A. Quantization with Three Reads

In this subsection, we fix our quantizer $Q(\cdot)$ with three reads, which are placed between every pair of adjacent voltage levels. Hence, the output alphabet $\mathcal{Y}_{MLC} = \{s_0, s_1, s_2, s_3\}$. For DS decoding, we assume that the output alphabet for lower page X_1 is $\mathcal{Y}^1_{MLC} = \{s_{0\cup 1}, s_{2\cup 3}\}$ of cardinality two, and the output alphabet for upper page X_2 is $\mathcal{Y}^2_{MLC} = \{s_0, s_{1\cup 2}, s_3\}$ of cardinality three [2]. DS decoding also requires a total of three reads.

We study the performance of MLC flash memories using the *P/E cycling* model, which has different channel characteristics for early and late stages of the memory lifetime.

1) Early Stage P/E Cycling Model: The channel transition matrix $p^E_{MLC}(y|v)$, for output $y \in \mathcal{Y}_{MLC}$ and voltage level $v \in \mathcal{V}_{MLC}$, reflects empirical results in [8] and is shown in Table I, where a_1, $1-a_1$, b_1, $1-b_1$, c_1, and $1-c_1$ represent non-zero transition probabilities.

As shown in Table I, note that the *transition probability* of inputs (X_1, X_2) to output Y is determined by 1) the *labeling* which maps inputs to voltage levels, and 2) the *channel transition matrix* from voltage levels to output.

Lemma 3. *For channel transition matrix* $p^E_{MLC}(y|v)$*, using Gray labeling, we have* $r_s^{TIN} = r_s^{SC}$ *and* $\mathscr{R}^{TIN} = \mathscr{R}^{SC}$*. Using either NO labeling or EO labeling, we have* $r_s^{TIN} < r_s^{SC}$*.*

Proof: With Gray labeling, in Table I, for the column $Y = s_0$, we have $p_{0,s_0} = 0$, $p_{1,s_0} = 0$, $p_{2,s_0} = 0$, and $p_{3,s_0} = a_1$. Thus, $p_{3,s_0} p_{0,s_0} = p_{2,s_0} p_{1,s_0}$. We can also verify $p_{3,s_i} p_{0,s_i} = p_{2,s_i} p_{1,s_i}$ for $i = 1, 2, 3$. Thus, from Theorem 1, we conclude $r_s^{TIN} = r_s^{SC}$ and $\mathscr{R}^{TIN} = \mathscr{R}^{SC}$. On the other hand, under NO labeling $p_{3,s_2} p_{0,s_2} \neq p_{2,s_2} p_{1,s_2}$, and under EO labeling $p_{3,s_3} p_{0,s_3} \neq p_{2,s_3} p_{1,s_3}$. Thus, from Theorem 1, for these two labelings, $r_s^{TIN} < r_s^{SC}$. ■

[2]We use notation $s_{u\cup v}$ to represent an output by merging two outputs s_u and s_v in $\mathcal{Y}_{MLC}$, i.e., $P(Y = s_{u\cup v}|X_1 = x_1, X_2 = x_2) = \sum_{i\in\{u,v\}} P(Y = s_i|X_1 = x_1, X_2 = x_2)$ for any $x_1, x_2 \in \{0, 1\}$. Strictly speaking, for the upper page decoding, current MLC flash memories use an output alphabet $\mathcal{Y}^2_{MLC} = \{s_{1\cup 2}, s_{0\cup 3}\}$. The resulting performance cannot exceed that obtained with the output alphabet $\{s_0, s_{1\cup 2}, s_3\}$ used in this paper.

TABLE III

TRANSITION MATRIX $p^L_{MLC}(y|v)$ AT LATE STAGE OF P/E CYCLING

V	Inputs: (X_1, X_2)			Output: Y			
Levels	Gray	NO	EO	s_0	s_1	s_2	s_3
A_0	(11)	(11)	(11)	$\hat{a}_1$	$\hat{a}_2$	0	$1-\hat{a}_1-\hat{a}_2$
A_1	(10)	(10)	(00)	0	$\hat{b}_1$	$1-\hat{b}_1$	0
A_2	(00)	(01)	(01)	0	0	$\hat{c}_1$	$1-\hat{c}_1$
A_3	(01)	(00)	(10)	0	0	0	1

TABLE IV

UNIFORM RATE REGIONS AND SUM RATES OF DS, TIN, AND SC DECODINGS AT LATE STAGE OF P/E CYCLING FOR MLC FLASH

Gray	DS	$\mathscr{R}_G^{DS}$	$0 \leqslant R_1 \leqslant \tau_1$, $0 \leqslant R_2 \leqslant \tau_2$	$r_{s(G)}^{DS} = \tau_1 + \tau_2$
	TIN	$\mathscr{R}_G^{TIN}$	$0 \leqslant R_1 \leqslant \tau_3$, $0 \leqslant R_2 \leqslant \tau_4$	$r_{s(G)}^{TIN} = \tau_3 + \tau_4$
	SC	$\mathscr{R}_G^{SC}$	$0 \leqslant R_1 \leqslant \tau_5$, $0 \leqslant R_2 \leqslant \tau_6$, $R_1 + R_2 \leqslant \tau_4 + \tau_5$	$r_{s(G)}^{SC} = \tau_4 + \tau_5$
NO	TIN	$\mathscr{R}_{NO}^{TIN}$	$0 \leqslant R_1 \leqslant \tau_3$, $0 \leqslant R_2 \leqslant \tau_7$	$r_{s(NO)}^{TIN} = \tau_3 + \tau_7$
	SC	$\mathscr{R}_{NO}^{SC}$	$0 \leqslant R_1 \leqslant \tau_8$, $0 \leqslant R_2 \leqslant \tau_6$, $R_1 + R_2 \leqslant \tau_7 + \tau_8$	$r_{s(NO)}^{SC} = \tau_7 + \tau_8$
EO	TIN	$\mathscr{R}_{EO}^{TIN}$	$0 \leqslant R_1 \leqslant \tau_4$, $0 \leqslant R_2 \leqslant \tau_7$	$r_{s(EO)}^{TIN} = \tau_4 + \tau_7$
	SC	$\mathscr{R}_{EO}^{SC}$	$0 \leqslant R_1 \leqslant \tau_8$, $0 \leqslant R_2 \leqslant \tau_5$, $R_1 + R_2 \leqslant \tau_7 + \tau_8$	$r_{s(EO)}^{SC} = \tau_7 + \tau_8$

Next, we calculate uniform rate regions and sum rates of the three decoding schemes under different labelings. The results are shown in Table II, where λ_i, $i = 1, 2, \ldots, 5$, are

$$
\begin{aligned}
\lambda_1 =& \frac{f(1-b_1) - f(3-b_1)}{4} + \frac{3}{2}, \\
\lambda_2 =& 1 + \frac{1}{4}\Big(f(1-a_1) + f(1+c_1) + f(1-c_1) - f(2-c_1) - f(2-a_1+c_1)\Big), \\
\lambda_3 =& 1 + \frac{1}{4}\Big(f(1-b_1) + f(c_1) - f(1-b_1+c_1)\Big), \\
\lambda_4 =& 1 + \frac{1}{4}\Big(f(1-a_1) + f(b_1) + f(1-c_1) - f(1-a_1+b_1) - f(2-c_1)\Big), \\
\lambda_5 =& 1 - \frac{1}{4}\Big(f(1-a_1+b_1) + f(2-c_1) + f(1-b_1+c_1)\Big) \\
&+ \frac{1}{4}\Big(f(1-a_1) + f(c_1) + f(1-c_1) + f(b_1) + f(1-b_1)\Big).
\end{aligned}
$$

2) Late Stage P/E Cycling Model: The channel transition matrix $p^L_{MLC}(y|v)$, for output $y \in \mathcal{Y}_{MLC}$ and voltage level $v \in \mathcal{V}_{MLC}$, reflects measurements in [8] and its structure is shown in Table III where $\hat{a}_1$, $\hat{a}_2$, $1-\hat{a}_1-\hat{a}_2$, $\hat{b}_1$, $1-\hat{b}_1$, $\hat{c}_1$, and $1-\hat{c}_1$ represent non-zero transition probabilities.

Lemma 4. *For channel transition matrix* $p^L_{MLC}(y|v)$*, we have* $r_s^{TIN} < r_s^{SC}$ *with Gray labeling, NO labeling or EO labeling.*

Proof: For any of the three labelings, consider the column $Y = s_3$ in Table III. Since three of p_{0,s_3}, p_{1,s_3}, p_{2,s_3}, and p_{3,s_3} are positive, it is impossible to make $p_{3,s_3} p_{0,s_3} = p_{2,s_3} p_{1,s_3} = 0$. Thus, from Theorem 1, we conclude $r_s^{TIN} < r_s^{SC}$. ■

Next, we calculate uniform rate regions and sum rates of the three decoding schemes under different labelings. The results are shown in Table IV, where τ_i, $i = 1, 2, \ldots, 8$, are

$$
\begin{aligned}
\tau_1 =& \frac{f(2-\hat{a}_1-\hat{a}_2-\hat{b}_1) - f(4-\hat{a}_1-\hat{a}_2-\hat{b}_1)}{4} + \frac{3}{2}, \\
\tau_2 =& 1 + \frac{1}{4}\Big(f(\hat{a}_2) + f(2-\hat{a}_1-\hat{a}_2) + f(1+\hat{c}_1) + f(1-\hat{c}_1)\Big) \\
&- \frac{1}{4}\Big(f(3-\hat{a}_1-\hat{a}_2-\hat{c}_1) + f(\hat{a}_2+\hat{c}_1+1)\Big), \\
\tau_3 =& 1 + \frac{1}{4}\Big(f(1-\hat{b}_1) + f(\hat{c}_1) + f(1-\hat{a}_1-\hat{a}_2) + f(2-\hat{c}_1)\Big) \\
&- \frac{1}{4}\Big(f(1-\hat{b}_1+\hat{c}_1) + f(3-\hat{a}_1-\hat{a}_2-\hat{c}_1)\Big), \\
\tau_4 =& 1 + \frac{1}{4}\Big(f(\hat{a}_2) + f(2-\hat{a}_1-\hat{a}_2) + f(\hat{b}_1) + f(1-\hat{c}_1)\Big) \\
&- \frac{1}{4}\Big(f(\hat{a}_2+\hat{b}_1) + f(3-\hat{a}_1-\hat{a}_2-\hat{c}_1)\Big), \\
\tau_5 =& 1 + \frac{1}{4}\Big(f(1-\hat{a}_1-\hat{a}_2) + f(1-\hat{b}_1) + f(\hat{c}_1) - f(2-\hat{a}_1-\hat{a}_2) - f(1-\hat{b}_1+\hat{c}_1)\Big), \\
\tau_6 =& 1 + \frac{1}{4}\Big(f(\hat{a}_2) + f(\hat{b}_1) + f(1-\hat{c}_1) - f(\hat{a}_2+\hat{b}_1) - f(2-\hat{c}_1)\Big), \\
\tau_7 =& 1 - \frac{1}{4}\Big(f(\hat{a}_2+\hat{b}_1) + f(1-\hat{b}_1+\hat{c}_1) + f(3-\hat{a}_1-\hat{a}_2-\hat{c}_1)\Big) \\
&+ \frac{1}{4}\Big(f(\hat{a}_2) + f(\hat{c}_1) + f(2-\hat{a}_1-\hat{a}_2-\hat{c}_1) + f(\hat{b}_1) + f(1-\hat{b}_1)\Big), \\
\tau_8 =& 1 + \frac{1}{4}\Big(f(1-\hat{c}_1) + f(1-\hat{a}_1-\hat{a}_2) - f(2-\hat{a}_1-\hat{a}_2-\hat{c}_1)\Big).
\end{aligned}
$$

TABLE V
TRANSITION MATRIX $p_{MLC}^{DR}(y|v)$ FOR DATA RETENTION MODEL

V	Inputs: (X_1, X_2)			Output: Y			
Levels	Gray	NO	EO	s_0	s_1	s_2	s_3
A_0	(11)	(11)	(11)	1	0	0	0
A_1	(10)	(10)	(00)	$1-\tilde{a}_1$	$\tilde{a}_1$	0	0
A_2	(00)	(01)	(01)	0	$1-\tilde{b}_1$	$\tilde{b}_1$	0
A_3	(01)	(00)	(10)	0	0	$1-\tilde{c}_1$	$\tilde{c}_1$

From Tables II and IV, we have the following comparisons for different labelings.

Theorem 5. *With channel transition matrix either $p_{MLC}^{E}(y|v)$ or $p_{MLC}^{L}(y|v)$, the rate regions satisfy $\mathscr{R}_G^{DS} \subset \mathscr{R}_G^{TIN}$, $\mathscr{R}_{NO}^{TIN} \subset \mathscr{R}_G^{TIN}$, and $\mathscr{R}_G^{SC} \subset \mathscr{R}_{NO}^{SC}$. For the sum rates, we have $r_{s(G)}^{TIN} > r_{s(G)}^{DS}$, $r_{s(G)}^{TIN} > r_{s(NO)}^{TIN}$, $r_{s(G)}^{TIN} > r_{s(EO)}^{TIN}$, and $r_{s(G)}^{SC} = r_{s(NO)}^{SC} = r_{s(EO)}^{SC}$.*

Remark 2 For the P/E cycling model (both early and late stages), from Theorem 5, among the 3 labelings, Gray labeling gives the largest sum rate for TIN decoding, which is larger than that of DS decoding. Moreover, compared to NO labeling, Gray labeling generates a larger uniform rate region for TIN decoding, but a smaller one for SC decoding.

For the early stage P/E cycling model, from Lemma 3, Theorem 5, and Table II, the sum rate of TIN decoding under Gray labeling is the same as that of SC decoding under any of the 3 labelings. For NO labeling or EO labeling, with SC decoding, rate $(R_1 = 1, R_2 = \lambda_5)$ can be achieved, which means that the lower page X_1 does not need coding. This gives us a very simple ECC solution for MLC flash. For encoding, we only need to apply a point-to-point capacity-achieving code, e.g., polar code [1], to the upper page X_2 to achieve rate $I(X_2; Y)$, and no coding is needed for the lower page X_1. For decoding, we first decode the upper page. Then, based on the decoded data from the upper page, binary labeling, channel transition matrix, and output Y, the data of the lower page can be determined.

For the late stage P/E cycling model, from Lemma 4, Theorem 5, and Table IV, the sum rate of TIN decoding under Gray labeling is strictly less than that of SC decoding. The gap Δ between the two is

$$\begin{aligned}\Delta =& r_{s(G)}^{SC} - r_{s(G)}^{TIN} = \tau_5 - \tau_3 \\ =& \frac{1}{4}\Big(f(3-\hat{a}_1-\hat{a}_2-\hat{c}_1) - f(2-\hat{a}_1-\hat{a}_2) - f(2-\hat{c}_1)\Big).\end{aligned}$$

For $0 < \hat{a}_1 + \hat{a}_2 < 1$ and $0 < \hat{c}_1 < 1$, we have $0 < \Delta < \frac{3\log_2 3 - 4}{4} = 0.1887$. If we impose constraints $\eta_{\hat{a}} \leqslant \hat{a}_1 + \hat{a}_2 < 1$ and $\eta_{\hat{c}_1} \leqslant \hat{c}_1 < 1$, we have $0 < \Delta \leqslant \frac{1}{4}\Big(f(3-\eta_{\hat{a}}-\eta_{\hat{c}_1}) - f(2-\eta_{\hat{a}}) - f(2-\eta_{\hat{c}_1})\Big)$ (see [5] for the proof). For example, with $\eta_{\hat{a}} = 0.95$ and $\eta_{\hat{c}_1} = 0.85$, we get $0 < \Delta \leqslant 0.00246$. In general, the gap Δ is very small. □

For the *data retention* model, the channel transition matrix $p_{MLC}^{DR}(y|v)$, for output $y \in \mathcal{Y}_{MLC}$ and voltage level $v \in \mathcal{V}_{MLC}$, reflects measured data and its structure is shown in Table V where $\tilde{a}_1$, $1-\tilde{a}_1$, $\tilde{b}_1$, $1-\tilde{b}_1$, $\tilde{c}_1$, and $1-\tilde{c}_1$ represent non-zero transition probabilities. Analysis of this model is very similar to the early stage P/E cycling model. We only give one result here.

Lemma 6. *For channel transition matrix $p_{MLC}^{DR}(y|v)$, using Gray labeling, we have $r_s^{TIN} = r_s^{SC}$ and $\mathscr{R}^{TIN}$=$\mathscr{R}^{SC}$. Using either NO labeling or EO labeling, we have $r_s^{TIN} < r_s^{SC}$.*

Next, we study the structure and property of all labelings. There exist a total of $4! = 24$ labelings. In order to categorize these 24 labelings, we consider a *labeling* σ as a *permutation* π in the *symmetric group* $\mathcal{S}_4$. This is the group whose elements are all the permutation operations that can be performed on 4 distinct elements in $\mathcal{T}_{MLC}$, and whose group operation, denoted as $*$, is the composition of such permutation operations, which are defined as bijective functions from the set $\mathcal{T}_{MLC}$ to itself. A labeling $\sigma = (w_0, w_1, w_2, w_3)$ corresponds to the permutation $\pi = (w_0, w_1, w_2, w_3)$ in $\mathcal{S}_4$, where the permutation vector $\pi = (w_0, w_1, w_2, w_3)$ (the w_i, $i = 0, 1, 2, 3$, represent the full set of possible 2-tuples) is defined to represent $\pi(11) = w_0$, $\pi(10) = w_1$, $\pi(01) = w_2$, and $\pi(00) = w_3$, e.g., $\pi = (11, 10, 01, 00)$ is the identity permutation in $\mathcal{S}_4$.

Lemma 7. *In the symmetric group $\mathcal{S}_4$, G_0=$\{(11,10,01,00)$, $(10,11,00,01)$, $(01,00,11,10)$, $(00,01,10,11)\}$ forms a normal subgroup (the Klein four-group).*

With subgroup G_0 in Lemma 7, we partition $\mathcal{S}_4$ into G_0 and its 5 cosets, each of size 4: $\mathcal{S}_4 = G_0 \cup G_1 \cup G_2 \cup \bar{G}_0 \cup \bar{G}_1 \cup \bar{G}_2$, where $G_1 = (11, 10, 00, 01) * G_0$; $G_2 = (11, 00, 01, 10) * G_0$; $\bar{G}_0 = (11, 01, 10, 00) * G_0$; $\bar{G}_1 = (11, 01, 00, 10) * G_0$; $\bar{G}_2 = (11, 00, 10, 01) * G_0$.

In the following, we will treat each vector in every coset as a labeling. For example, G_0 includes the NO labeling $\sigma_{NO} = (11, 10, 01, 00)$, G_1 includes $\sigma_G = (11, 10, 00, 01)$, and G_2 includes $\sigma_{EO} = (11, 00, 01, 10)$. The following two lemmas give properties of the uniform rate regions for different labelings. We assume an *arbitrary* channel transition matrix $p_{MLC}(y|v)$ is given.

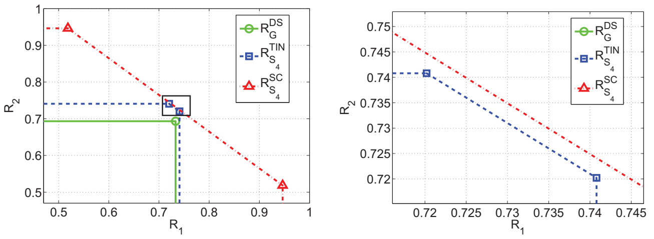
Lemma 8. *With channel transition matrix $p_{MLC}(y|v)$, for TIN decoding, the 4 labelings in each of G_0, G_1, G_2, $\bar{G}_0$, $\bar{G}_1$, and $\bar{G}_2$ give the same uniform rate region $\mathscr{R}^{TIN}$ and sum rate r_s^{TIN}. For SC decoding, the 4 labelings in each of G_0, G_1, G_2, $\bar{G}_0$, $\bar{G}_1$, and $\bar{G}_2$ give the same uniform rate region $\mathscr{R}^{SC}$, and all 24 labelings in $\mathcal{S}_4$ give the same sum rate r_s^{SC}.*

Lemma 9. *With channel transition matrix $p_{MLC}(y|v)$, for TIN decoding, if labelings in G_i, $i = 0, 1, 2$, give a uniform rate region: $0 \leqslant R_1 \leqslant \varphi_1$ and $0 \leqslant R_2 \leqslant \varphi_2$, then labelings in $\bar{G}_i$ give a uniform rate region: $0 \leqslant R_1 \leqslant \varphi_2$ and $0 \leqslant R_2 \leqslant \varphi_1$. For SC decoding, if labelings in G_i give a uniform rate region: $0 \leqslant R_1 \leqslant \psi_1$, $0 \leqslant R_2 \leqslant \psi_2$, and $R_1 + R_2 \leqslant \psi_3$, then labelings in $\bar{G}_i$ give a uniform rate region: $0 \leqslant R_1 \leqslant \psi_2$, $0 \leqslant R_2 \leqslant \psi_1$, and $R_1 + R_2 \leqslant \psi_3$.*

Remark 3 From Theorem 5, and Lemmas 8 and 9, under the P/E cycling model (both early and late stages), for TIN decoding, the 8 labelings in G_1 (including Gray labeling) and $\bar{G}_1$ among all 24 labelings produce the largest sum rate. For SC decoding, all the 24 labelings give the same sum rate. □

Last, we discuss the uniform rate region if we are allowed to use multiple labelings together for each codeword instead of one labeling in MLC flash. In current flash memory, only Gray labeling is used. More specifically, we study the uniform rate region achieved by *time sharing* of using labelings in $\mathcal{S}_4$.

Define $\mathscr{R}_{\mathcal{S}_4}^{TIN} = Conv\Big(\bigcup_{\sigma \in \mathcal{S}_4} \mathscr{R}_\sigma^{TIN}\Big)$, the convex hull of uniform rate regions of all 24 labelings for TIN decoding. Define $\mathscr{R}_{\mathcal{S}_4}^{SC} = Conv\Big(\bigcup_{\sigma \in \mathcal{S}_4} \mathscr{R}_\sigma^{SC}\Big)$, the convex hull of uniform rate regions of all 24 labelings for SC decoding. Through



(a) (b)

Fig. 1. (a) Uniform rate regions $\mathscr{R}_{S_4}^{TIN}$, $\mathscr{R}_{S_4}^{SC}$, and $\mathscr{R}_G^{DS}$ (baseline) with $\hat{a}_1 = 0.82$, $\hat{a}_2 = 0.1$, $\hat{b}_1 = 0.85$, and $\hat{c}_1 = 0.85$ for the late stage P/E cycling model, where the two curves (blue and red) in the black rectangle are shown in (b).

time sharing of different labelings, for TIN decoding, any point $(R_1, R_2) \in \mathscr{R}_{S_4}^{TIN}$ can be achieved. For SC decoding, any point $(R_1, R_2) \in \mathscr{R}_{S_4}^{SC}$ can be achieved (see [5] for the proof). Moreover, for the early stage P/E cycling model, $\mathscr{R}_{S_4}^{SC}$ can be determined explicitly as follows.

Theorem 10. *For the early stage P/E cycling model,* $\mathscr{R}_{S_4}^{SC}$ *is the set of all pairs* (R_1, R_2) *such that* $0 \leqslant R_1 \leqslant 1$, $0 \leqslant R_2 \leqslant 1$, *and* $R_1 + R_2 \leqslant I(X_1, X_2; Y) = 1 + \lambda_5$.

An example of uniform rate regions $R_{S_4}^{TIN}$, $R_{S_4}^{SC}$, and R_G^{DS} is shown in Fig. 1 for the late stage P/E cycling model. We can see $R_G^{DS} \subset R_{S_4}^{TIN} \subset R_{S_4}^{SC}$, where the inclusions are strict.

B. Quantization with Multiple Reads

In the above subsection, we use a quantizer with three reads. To improve decoding performance, we can progressively apply multiple reads to obtain more soft information.

For a channel $\mathcal{W}_{MLC}$, assume we already have output $\mathcal{Y}_{MLC}^q = \{s_0, s_1, \ldots, s_{q-1}\}$ obtained by a set of reads. The corresponding uniform rate regions for TIN and SC decodings are $\mathscr{R}^{TIN}$ and $\mathscr{R}^{SC}$, and the sum rates for TIN and SC decodings are r_s^{TIN} and r_s^{SC}. Now, we apply one more read (quantization) to split one of the outputs $s_0, s_1, \ldots, s_{q-1}$. Without loss of generality, we split s_0 into s_0^1 and s_0^2 to obtain $\hat{\mathcal{Y}}_{MLC}^{q+1} = \{s_0^1, s_0^2, s_1, s_2, \ldots, s_{q-1}\}$. The resulting uniform rate regions for TIN and SC decodings become $\hat{\mathscr{R}}^{TIN}$ and $\hat{\mathscr{R}}^{SC}$, and the corresponding sum rates become $\hat{r}_s^{TIN}$ and $\hat{r}_s^{SC}$. The following lemma shows such one-step progressive quantization will not decrease (in general, strictly increase) the performance.

Lemma 11. *For uniform rate regions, under TIN and SC decodings,* $\mathscr{R}^{TIN} \subseteq \hat{\mathscr{R}}^{TIN}$ *and* $\mathscr{R}^{SC} \subseteq \hat{\mathscr{R}}^{SC}$. *For sum rates, under TIN and SC decodings,* $r_s^{TIN} \leqslant \hat{r}_s^{TIN}$ *and* $r_s^{SC} \leqslant \hat{r}_s^{SC}$.

V. Extension to TLC Flash Memory

In this section, we extend our analysis from MLC to TLC flash memory. Similar to the model proposed for MLC flash memory, given a labeling σ and a quantizer Q, the TLC flash memory channel can be modeled as a 3-user discrete memoryless multiple access channel $\mathcal{W}_{TLC}$: $(\mathcal{X} \times \mathcal{X} \times \mathcal{X}, p(y|x_1, x_2, x_3), \mathcal{Y})$, where $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{s_0, s_1, \ldots, s_{q-1}\}$, and $p(y|x_1, x_2, x_3)$ is the transition probability for any $x_1, x_2, x_3 \in \mathcal{X}$ and $y \in \mathcal{Y}$. Define $P(Y = y|X_1 = x_1, X_2 = x_2, X_3 = x_3) \stackrel{\text{def}}{=} p_{BD(x_1,x_2,x_3),y}$.

For a channel $\mathcal{W}_{TLC}$, the uniform rate region $\mathscr{R}^{TIN}$ for TIN decoding is the set of all pairs (R_1, R_2, R_3) such that: $0 \leqslant R_i \leqslant I(X_i; Y)$ for all $i = 1, 2, 3$. In $\mathscr{R}^{TIN}$, the sum rate $r_s^{TIN} = \max\{\sum_{i=1}^{3} R_i : (R_1, R_2, R_3) \in \mathscr{R}^{TIN}\} = \sum_{i=1}^{3} I(X_i, Y)$. For SC decoding, its uniform rate region $\mathscr{R}^{SC}$ is the set of all pairs (R_1, R_2, R_3) such that: 1) $R_1 \leqslant I(X_1; Y|X_2, X_3)$, $R_2 \leqslant I(X_2; Y|X_1, X_3)$, $R_3 \leqslant I(X_3; Y|X_1, X_2)$; 2) $R_1 + R_2 \leqslant I(X_1, X_2; Y|X_3)$, $R_1 + R_3 \leqslant I(X_1, X_3; Y|X_2)$, $R_2 + R_3 \leqslant I(X_2, X_3; Y|X_1)$; 3) $R_1 + R_2 + R_3 \leqslant I(X_1, X_2, X_3; Y)$. In $\mathscr{R}^{SC}$, the sum rate $r_s^{SC} = \max\{\sum_{i=1}^{3} R_i : (R_1, R_2, R_3) \in \mathscr{R}^{SC}\} = I(X_1, X_2, X_3; Y)$.

Theorem 12. *For a channel* $\mathcal{W}_{TLC}$, *the sum rates* $r_s^{TIN} \leqslant r_s^{SC}$ *with equality if and only if*

$$\begin{aligned}
&(p_{7,s_j} + p_{6,s_j})(p_{1,s_j} + p_{0,s_j}) = (p_{5,s_j} + p_{4,s_j})(p_{3,s_j} + p_{2,s_j}),\\
&p_{7,s_j}(p_{4,s_j} + p_{2,s_j} + p_{0,s_j}) = p_{6,s_j}(p_{5,s_j} + p_{3,s_j} + p_{1,s_j}),\\
&p_{5,s_j}(p_{6,s_j} + p_{2,s_j} + p_{0,s_j}) = p_{4,s_j}(p_{7,s_j} + p_{3,s_j} + p_{1,s_j}),\\
&p_{3,s_j}(p_{6,s_j} + p_{4,s_j} + p_{0,s_j}) = p_{2,s_j}(p_{7,s_j} + p_{5,s_j} + p_{1,s_j}),\\
&p_{1,s_j}(p_{6,s_j} + p_{4,s_j} + p_{2,s_j}) = p_{0,s_j}(p_{7,s_j} + p_{5,s_j} + p_{3,s_j}),
\end{aligned} \tag{3}$$

for all $j = 0, 1, \ldots, q-1$. *If* $r_s^{TIN} = r_s^{SC}$, *then* $\mathscr{R}^{TIN} = \mathscr{R}^{SC}$ *and the rate region is a cube.*

We omit analysis of TLC flash memories here, due to the space limitation. For more details on TLC, please refer to the online paper [5].

VI. Conclusion

We analyzed different decoding schemes for flash memories from a multi-user perspective. In MLC flash memory, both TIN and SC decoding schemes outperform the current default decoding scheme in terms of either rate region or sum rate. For the P/E cycling model, for TIN decoding, 8 labelings which include Gray labeling give the largest sum rate among all 24 labelings. The sum rate of TIN decoding under Gray labeling equals that of SC decoding at the early stage of P/E cycling, and is smaller than but close to that of SC decoding at the late stage of P/E cycling.

Acknowledgment

The work was supported by the CMRR at UCSD and NSF Grants CCF-1116739 and CCF-1405119. The authors thank Dr. Lele Wang for discussions.

References

[1] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[2] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis," in *Proc. IEEE DATE*, Dresden, March 2012, pp. 521–526.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, New Jersey: John Wiley & Sons, 2012.

[4] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, UK: Cambridge University Press, 2011.

[5] P. Huang, P. H. Siegel, and E. Yaakobi, "Performance of multilevel flash memories with different binary labelings: a multi-user perspective," http://arxiv.org/pdf/1605.00738, 2016.

[6] T. Parnell, N. Papandreou, T. Mittelholzer, and H. Pozidis, "Modelling of the threshold voltage distributions of sub-20nm NAND flash memory," in *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 2351–2356.

[7] C. Schoeny, F. Sala, and L. Dolecek, "Analysis and coding schemes for the flash normal-Laplace mixture channel," in *Proc. IEEE ISIT*, Hongkong, June 2015, pp. 2101–2105.

[8] V. Taranalli, H. Uchikawa, and P. H. Siegel, "Error analysis and inter-cell interference mitigation in multi-level cell flash memories," in *Proc. IEEE ICC*, London, June 2015, pp. 271–276.

[9] J. Wang, K. Vakilinia, T.-Y. Chen, T. Courtade, G. Dong, T. Zhang, H. Shankar, and R. Wesel, "Enhanced precision through multiple reads for LDPC decoding in flash memories," *IEEE JSAC*, vol. 32, no. 5, pp. 880–891, May 2014.