Optimizing Write Voltages to Achieve Equal Reliability for All Pages in Flash Memory

Semira Galijasevic
Department of Electrical and Computer Engineering
University of California
Los Angeles, USA
semiragali@g.ucla.edu

Richard D. Wesel
Department of Electrical and Computer Engineering
University of California
Los Angeles, USA
wesel@ucla.edu

I. Introduction and Motivation

Multilevel cell techniques store multiple bits in each cell. For example, triple-level-cell (TLC) Flash stores three bits per cell by using eight levels of write voltage. However, to minimize the read latency in practical systems, each bit in the same cell is mapped to a different page. Thus, the bits corresponding to the same cell are encoded independently. Recent work comparing the independent encoding used in practice with the information theoretically superior joint encoding include [1], [2], and [3]. It is generally accepted that when independent encoding is applied, some bit positions and hence some pages are more reliable than others.

This paper introduces a mutual information (MI) maximization paradigm that adapts the locations and probabilities of write levels to iteratively increase the mutual information of the weakest bit channel and hence improve the reliability of its corresponding page. In this way, we seek a constellation of write levels that delivers the same amount of mutual information to the bit channel for each page, so that all pages are equally reliable. For simplicity, we consider the example of TLC Flash with an additive white Gaussian noise (AWGN) channel model, but the principle may be applied to denser cells and more realistic channel models. The following sections are shortened due to limited space. The full paper with detailed background, equations, and algorithms is given in [4].

Consider a TLC Flash memory with eight voltage levels. As shown in Fig. 1, the three bits B_1, B_2 and B_3 written to a cell for the three independent pages together cause the threshold voltage X to be written to the Flash cell. As noted in [2], this is analogous to a multiple access channel (MAC) with three users. When the cell voltage is read, distortion causes the actual threshold voltage at the time of reading to be $Y = X+Z$ where $Z \sim \mathcal{N}(0, N)$. In this paper, the noise Z is assumed to be independent of the signal X, but actual Flash noise is signal dependent. We can model the Flash write levels as M-ary pulse amplitude modulation (M-PAM). For TLC, we use 8-PAM constellations to store 3 bits per cell. We will consider equally spaced equally likely (ESEL) 8-PAM as illustrated in the Table I in [4] as a baseline for comparison, although practical Flash write levels are not equally spaced.

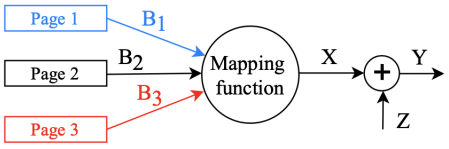

Fig. 1. Flash Memory Layout as Multiple Access Channel (MAC). The noise Z is drawn i.i.d. from a Gaussian distribution with variance N.

II. The Mutual Information Perspective

Page mutual information rates $I(B_1;Y), I(B_2;Y)$ and $I(B_3;Y)$ for independent encoding are calculated as follows:

$$I(B_j;Y) = \int_y f(B_j=0,y)\log_2\left(\frac{f(B_j=0,y)}{P(B_j=0)f(y)}\right)dy + \int_y f(B_j=1,y)\log_2\left(\frac{f(B_i=1,y)}{P(B_i=1)f(y)}\right)dy$$

where $P(B_j=0)=p_j$, $P(B_j=1)=1-p_j$, $j=1,2,3$, and $f(\cdot,\cdot)$ is a joint probability density function.

The penalty for encoding independently rather than jointly is calculated as follows, assuming B_1, B_2, B_3 are independent:

$$I(B_1,B_2,B_3;Y) - I(B_1;Y) - I(B_2;Y) - I(B_3;Y) = I(B_1;B_2|Y) + I(B_1,B_2;B_3|Y) \geq 0 \quad (1)$$

Sec. III in [4] provides detailed derivations.

III. Maximizing the Minimum Rate

We formulate following optimization problem to maximize the minimum page MI subject to power constraint P_c, i.e. we seek a solution for which all three page MIs in TLC Flash memory are equal:

$$\max_{\mathcal{X},p_2,p_3} \min_j I(B_j;Y) \quad j=1,2,3. \quad (2)$$

$$\text{s.t.} \quad \sum_{i=0}^{M-1} p(x_i)x_i^2 = P_c \quad (3)$$

$$x_k = -x_{M-1-k}, \qquad k=0,\dots M/2-1 \quad (4)$$

$$p(x_k) = p(x_{M-1-k}), \quad k=0,\dots M/2-1 \quad (5)$$

$$p(x_i) = P_{B_1}(b_1(i))P_{B_2}(b_2(i))P_{B_3}(b_3(i)) \quad (6)$$

With $M=8$ for TLC Flash and $\mathcal{X}=\{x_0,x_1,...,x_7\}$ having the ESEL values $x=\{-7,-5,-3,-1,1,3,5,7\}$, the average

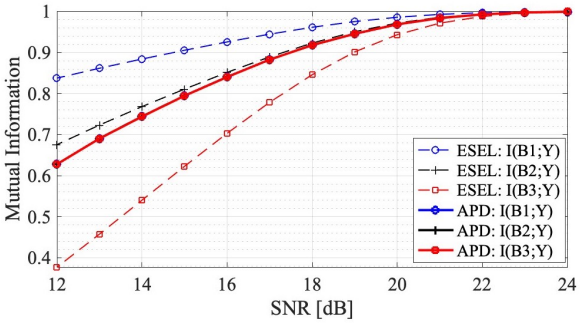

Fig. 2. Independent MI rates for Equally Spaced Equally Likely (ESEL) Constellations and Adaptive Positions and Distribution (APD) 8-PAM Independent rates for the optimized constellation as a function of SNR.

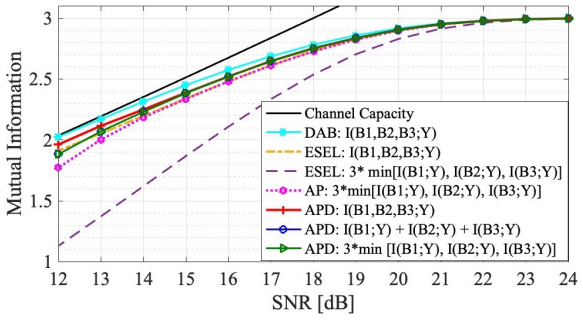

Fig. 3. Three times the minimum independent rate for AP and APD, i.e. $3 * \min\{I(B_1 : Y), I(B_2 : Y), I(B_3 : Y)\}$ is compared to the joint mutual information rate $I(B_1, B_2, B_3; Y)$ for AP and APD, the sum of independent rates, three times the minimum ESEL rate, and the MI for DAB optimized constellations. Equal-rate constraint loss and independent-encoding loss are nearly zero. The most evident loss comes from shaping.

power for equally spaced equally likely constellation points given by $P_c = \frac{\sum_{i=0}^{7} x_i^2}{8} = 21$. We present two algorithms in [4] as a solution to optimization problem above. Alg. 1 enumerates the steps for maximizing the minimum rate through dynamic assignment of the write levels, i.e. adaptive positions (AP). Alg. 2 enumerates the steps for maximizing the minimum rate by optimizing both positions and probabilities of the write levels, i.e. adaptive positions and distribution (APD). In this paper we have modified Alg. 1 and Alg. 2 in [4] as follows:

1) *Adjust Write-Level Positions $\mathcal{X}$ in Alg. 1*

A single scalar parameter α replaces diagonal matrix $\tilde{d}$ in [4] scaling all points to satisfy the power constraint as follows:

$$\sum_{i=0}^{M-1} p(x_i)\alpha^2 x_i^2 = P_c \tag{7}$$

2) *Adjust Write-Level Positions $\mathcal{X}$ in Alg. 2*

In Alg. 2 instead of scaling the constellation points in Step 1, we adjust the *probabilities* of either p_2 or p_3 to maintain the power constraint. That is , we either scale p_2 by β or p_3 by γ so that as a point pair is moved, $p(x_i)$ in (6) is adjusted so that the power constraint in (3) is satisfied.

3) *Optimize PMFs p_2 and p_3 in Alg. 2*

Gradient descent is replaced with line search (fminbnd in MATLAB) to find optimal p_2 and p_3 while scaling the constellation to maintain the power constraint as in Eq. 7.

Fig. 2 compares the (essentially equal) APD independent rates achieved by the optimized point locations and PMFs to the unequal ESEL rates as a function of SNR. Observing Fig. 3, three losses can be examined for flash system:

1) Shaping loss: $I(B_1, B_2, B_3; Y)_{DAB} - I(B_1, B_2, B_3; Y)$
2) Independent-encoding loss: $I(B_1, B_2, B_3; Y) - I(B_1 : Y) - I(B_2; Y) - I(B_3; Y)$
3) Equal-rate constraint loss: $I(B_1; Y) + I(B_2; Y) + I(B_3; Y) - 3 * \min_j\{I(B_j : Y)\}$

DAB refers to the capacity achieving constellations of [5]. We note that equal rate constraint loss and independent encoding loss are nearly zero for both AP and APD rates. However, APD rates provide significant improvement with respect to shaping loss.

IV. Conclusion

Optimizing the positions and probabilities of the write levels in the constellation to maximize the minimum mutual information rate of a bit channel significantly increases the minimum page rate. The independent encoding loss and equal rate constraint loss are negligible (essentially zero) for operational SNRs. Shaping loss is smaller for APD as compared to AP optimization and may be further reduced by an additional optimization step similar to DAB, which is the subject of future research. We note that 8-PAM constellations with points that are not equally likely can be supported by coded modulation techniques such as probabilistic amplitude shaping [6]. We look forward to applying this approach to realistic Flash read channels with signal-dependent noise, peak rather than average power constraints, and asymmetric constellations.

Acknowledgment

We thank Derek Xiao, an author of paper referenced in [5], for providing DAB algorithm MATLAB code and DAB optimized PMFs and points needed for DAB rate in Fig. 3.

References

[1] N. Wong, E. Liang, H. Wang, S. V. S. Ranganathan, and R. D. Wesel, "Decoding flash memory with progressive reads and independent vs. joint encoding of bits in a cell," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[2] P. Huang, P. H. Siegel, and E. Yaakobi, "Performance of multilevel flash memories with different binary labelings: A multi-user perspective," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 9, pp. 2336–2353, 2016.

[3] D. N. Bailon, S. Shavgulidze, and J. Freudenberger, "Cell-wise encoding and decoding for tlc flash memories," in *2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin)*, 2022, pp. 1–6.

[4] S. Galijasevic and R. D. Wesel. Optimizing write voltages for independent, equal-rate pages in flash memory. [Online]. Available: http://www.seas.ucla.edu/csl/files/publications/165.pdf

[5] D. Xiao, L. Wang, D. Song, and R. D. Wesel, "Finite-support capacity-approaching distributions for awgn channels," in *2020 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–5.

[6] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. on comm.*, vol. 63, no. 12, pp. 4651–4665, 2015.