This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2011 proceedings.

# Soft Information for LDPC Decoding in Flash: Mutual-Information Optimized Quantization

Jiadong Wang and Thomas Courtade
Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, California 90024
Email: {wjd,tacourta}@ee.ucla.edu

Hari Shankar
Inphi Corporation
112 S Lakeview Canyon Road
Westlake Village, California 91362
Email: hshankar@inphi.com

Richard D. Wesel
Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, California 90095
Email: wesel@ee.ucla.edu

***Abstract*—High-capacity NAND flash memory can achieve high density storage by using multi-level cells (MLC) to store more than one bit per cell. Although this larger storage capacity is certainly beneficial, the increased density also increases the raw bit error rate (BER), making powerful error correction coding necessary. Traditional flash memories employ simple algebraic codes, such as BCH codes, that can correct a fixed, specified number of errors. This paper investigates the application of low-density parity-check (LDPC) codes which are well known for their ability to approach capacity in the AWGN channel. We obtain soft information for the LDPC decoder by performing multiple cell reads with distinct word-line voltages. The values of the word-line voltages (also called reference voltages) are optimized by maximizing the mutual information between the input and output of the multiple-read channel. Our results show that using this soft information in the LDPC decoder provides a significant benefit and enables the LDPC code to outperform a BCH code with comparable rate and block length over a range of block error rates.**

## I. Introduction

Flash memory can store large quantities of data in a small device that has low power consumption and no moving parts. The original NAND flash memories used only two levels. This was called single-level-cell (SLC) flash because there is only one nonzero charge level. Devices currently available use multiple levels and are referred to as multiple-level cell (MLC) flash. Four and eight levels are currently in use, and the number of levels will increase further to provide more storage capability [1][2].

Error control coding for flash memory is becoming more important in a variety of ways as the storage density increases. The increasing number of levels (and smaller distance between levels) means that variations in cell behavior from cell to cell (and over time due to wear-out) lower the signal-to-noise ratio of the read channel making a stronger error-correction code necessary. Reductions in feature size make inter-cell interference more likely, adding an equalization or interference suppression component to the read channel [3]. Also, the wear-out effect is time varying, introducing a need for adaptive coding to maximize the potential of the system.

Low-density parity-check (LDPC) codes are well-known for their capacity-approaching ability in the AWGN channel [4]. LDPC codes have typically been decoded with soft reliability information while flash systems have typically only provided hard reliability information to their decoders. This paper demonstrates that at least some soft information is crucial to successfully reaping the benefits of LDPC coding in flash memory. We also explore how much soft information is necessary to provide most of the benefits and how flash systems could be engineered to provide the needed soft information without an unnecessary penalty in complexity or processing time.

This research was supported by a gift from Inphi Corp.

This paper uses pulse-amplitude modulation (PAM) with Gaussian noise to model Flash cell threshold voltage levels, and investigates how to optimize the word-line voltages by maximizing the mutual information between the input and the output of the equivalent read channel. After choosing the word-line voltage for each of the reads, the multiple-read channel can be represented by a probability transition matrix and the data can be decoded with a standard belief-propagation algorithm.

Section II introduces the basics of the NAND flash memory model and LDPC codes. Section III studies three cases covering SLC and MLC with different quantization choices. Section III also shows how to obtain word-line voltages by maximizing the mutual information of the equivalent read channel. Section IV provides simulation results demonstrating the benefits of using soft information with word-line voltages selected as described in Section III, and Section V delivers the conclusions.

## II. Background

This section introduces the basics of NAND flash memory and LDPC codes.

### A. Basics of NAND Flash Memory

This paper focuses on the NAND architecture for flash memory, which is the most prevalent architecture today. Each memory cell in the NAND architecture features a transistor with a control gate and a floating gate. To store information, a charge level is written to the cell by adding a specified amount of charge to the floating gate through Fowler-Nordheim tunneling by applying a relatively large voltage to the control gate [5].

978-1-4244-9268-8/11/$26.00 ©2011 IEEE

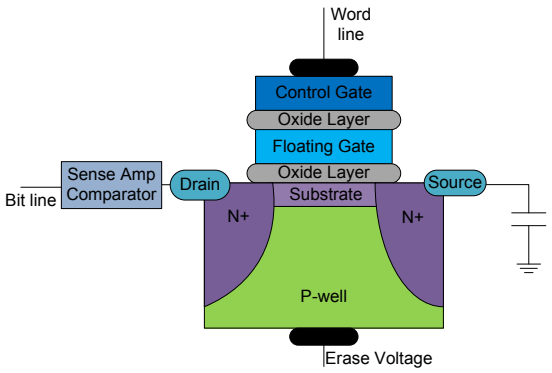


Fig. 1: A NAND flash memory cell.

Figure 1 shows the configuration of a NAND flash memory cell. To read a memory cell, the charge level written to the floating gate is detected by applying a specified word-line voltage to the control gate and measuring the transistor drain current. The drain current is compared to a threshold by a sense amp comparator. If the drain current is above the comparator threshold, then the word-line voltage was sufficient to turn on the transistor, indicating that the charge written to the floating gate was insufficient to prevent the transistor from turning on. If the drain current is below the threshold, the charge added to the floating gate was sufficient to prevent the applied word-line voltage from turning on the transistor. The sense amp comparator only provides one bit of information about the charge level present in the floating gate. A bit error occurring at this threshold-comparison stage is called a *raw bit error*.

The word-line voltage required to turn on a particular transistor (called the threshold voltage) can vary from cell to cell for a variety of reasons. For example, the floating gate can be overcharged during the write operation, the floating gate can lose charge due to leakage in the retention period, or the floating gate can receive extra charge when nearby cells are written [6].

The probability density function of the variation of threshold voltage from its intended value is usually modeled by a Gaussian distribution. In this paper, we assume an i.i.d. Gaussian threshold voltage for each level of an MLC flash memory cell. Therefore an $m$-level flash cell is equivalent to an $m$-PAM communication system with AWGN noise, except that the threshold voltage cannot be directly observed. Rather, one bit of information about the threshold voltage may be obtained by each cell read.

More precise models such as the model in [6] in which the lowest and highest threshold voltage distributions have a higher variance and the model in [7] in which the lowest threshold voltage (the one associated with zero charge level) is Gaussian and the other threshold voltages have Gaussian tails but a uniform central region are sometimes used. The model in [8] is similar to [7], but is derived by explicitly accounting for inter-cell interference. Despite its limitations, the simple Gaussian model is sufficient to motivate the proposed investigation of soft information. Furthermore, the techniques presented in this paper can easily be extended to other probability distributions.

### B. Basics of LDPC codes

LDPC codes are linear block codes defined by sparse parity-check matrices. By optimizing the degree distribution, it is well-known that LDPC codes can approach the capacity of an AWGN channel [4]. Several algorithms have been proposed to generate LDPC codes for a given degree distribution, such as the ACE algorithm [9], and the PEG algorithm [10].

Designing LDPC codes with low error-floors is crucial for applications to flash memory since storage systems usually require block-error-rates lower than $10^{-15}$. This topic has generated a significant amount of recent research including [11] [12] [13] [14] [15] [16].

In addition to their powerful error-correction capabilities, another appealing aspect of LDPC codes is the existence of low-complexity iterative algorithms used for decoding. These iterative decoding algorithms are called belief-propagation algorithms. Belief-propagation decoders commonly use soft reliability information about the received bits, which can greatly improve performance. Conversely, a quantization of the received information which is too coarse can degrade the performance of an LDPC code.

Traditional algebraic codes, such as BCH codes, use bounded distance decoding and can only correct a specified, fixed number of errors. Unlike these traditional codes, for LDPC codes it can be difficult to guarantee a specified number of correctable errors. However the average bit-error-rate performance can often outperform that of BCH codes in Gaussian noise.

The remainder of this paper studies how quantization during the read process affects the performance of LDPC decoding for flash memory. In the next section, we present a general quantization approach for selecting word line voltages for reading the flash memory cells in both the SLC and the MLC cases.

## III. Illustrative case study on SLC and MLC Flash Memory

Since the sense amp comparator only provides one bit of information about the threshold voltage (or equivalently the amount of charge present in the floating gate), decoders for error control codes in flash have historically relied on hard bit decisions from the sense-amp comparator. However, soft information can be obtained either by reading from the same sense amp comparator multiple times with different word line voltages (as is already done to read multi-level flash cells) or by equipping a flash cell with multiple sense amp comparators on the bit line, which is essentially equivalent to replacing the sense amp comparator (a one-bit A/D converter) with a higher precision A/D converter.

These two approaches are not completely interchangeable. The real goal is to detect soft information about the threshold

voltage. Each additional read of a single sense amp comparator can provide additional useful soft information about the threshold voltage if the word line voltages are well-chosen. However, multiple comparators may not give much additional information if the drain current vs. word-line-voltage curve (the classic I-V transistor curve) is too nonlinear. If the drain current has saturated too low or too high, the outputs from more sense-amp comparators are not useful in establishing precisely how much charge is in the floating gate. However, if the word line voltage and floating gate charge level place the transistor in the linear gain region, then some valuable soft information is provided by multiple sense amp comparators. Our work focuses on soft information obtained from multiple reads using the same sense-amp comparator with different word line voltages.

This section investigates the potential improvement of increasing the resolution beyond one bit and studies how best to obtain this increased resolution. In [8], the use of soft information was explored and the poor performance of uniformly spaced word-line voltages was clearly established. This paper takes an information-theoretic perspective on optimizing the word-line voltages. We study quantization models with different numbers of reads for both SLC and MLC flash memory. In the course of our analysis, we choose the word-line voltages for each quantization by maximizing the mutual information between the input and output of each equivalent read channel. Theoretically, this choice of word-line voltages maximizes the amount of information provided by the quantization. The next subsection provides an example of SLC with just one additional read to provide extra soft information. After that, the section looks at the benefit of additional reads for SLC and MLC. Numerical results are given in Section IV.

### A. SLC Flash Memory with 2 reads

For SLC flash memory, each cell can store 1 bit of information. Figure 2 shows a simplistic model of the threshold voltage distribution as a mixture of two Gaussian random variables. In particular, if a "0" is written to the cell, the threshold voltage is modeled as a Gaussian random variable with mean $-\sqrt{E_s}$ and variance $\sqrt{N_0/2}$. Similarly, if a "1" is written to the cell, we model the threshold voltage as a Gaussian random variable with mean $+\sqrt{E_s}$ and variance $\sqrt{N_0/2}$. Using this model, the read channel is equivalent to a 2-PAM signal with AWGN noise, where the noise power is $N_0/2$ and the symbol energy is $E_s$.

If we read twice with two different word line voltages (equivalent to $q$ and $-q$ in Figure 2), the threshold voltage can be quantized to one of three regions. This quantization model is shown in Figure 2 and an equivalent channel model is given in Figure 3.

Suppose the input and output of the equivalent channel are $X \in \{0,1\}$ and $Y \in \{0,e,1\}$ respectively, and the various crossover probabilities are as shown in Figure 3. Assuming $X$ is equally likely to be 0 or 1, the mutual information between

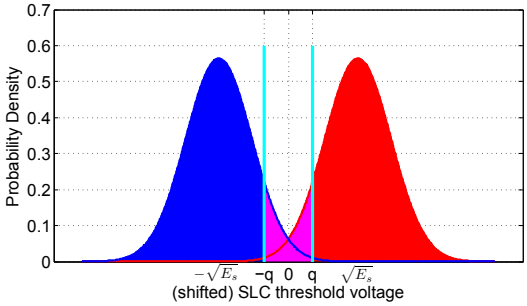


Fig. 2: Quantization model for SLC with 2 reads.

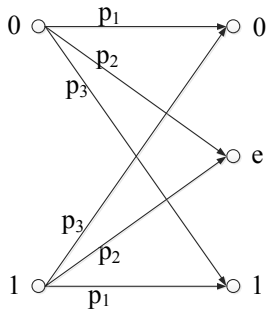


Fig. 3: Channel model for SLC with 2 reads.

the input and output can be calculated as

$$I(X;Y) = H(Y) - H(Y|X) = H\left(\frac{p_1+p_3}{2}, p_2, \frac{p_1+p_3}{2}\right) - H\left(p_1, p_2, p_3\right), \quad (1)$$

where the crossover probabilities are computed as

$$p_1 = 1 - Q\left(\frac{\sqrt{E_s}-q}{\sqrt{N_0/2}}\right),$$

$$p_2 = Q\left(\frac{\sqrt{E_s}-q}{\sqrt{N_0/2}}\right) - Q\left(\frac{\sqrt{E_s}+q}{\sqrt{N_0/2}}\right), \text{ and}$$

$$p_3 = Q\left(\frac{\sqrt{E_s}+q}{\sqrt{N_0/2}}\right).$$

For a fixed $E_s/N_0$, the mutual information in equation (1) can be maximized numerically to find the parameter $q$ that yields the largest mutual information $I(X;Y)$. Note that choosing $q$ to maximize the mutual information should provide approximately optimal LDPC decoding performance for a given level of quantization, and that the optimum $q^*$ is a function of $E_s/N_0$. For example, if $E_s/N_0 = 3.241$ dB, $q^* = 0.2188\sqrt{E_s}$, and if $E_s/N_0 = 6.789$ dB, $q^* = 0.1253\sqrt{E_s}$.

Figure 4 shows that the mutual information with 2 reads is larger than the mutual information with just 1 read, and recovers most of the gap to the capacity with full soft information.

Note that even if the probability density function of the threshold voltage is different from our assumption, the analysis can be easily extended to find the word-line voltages that

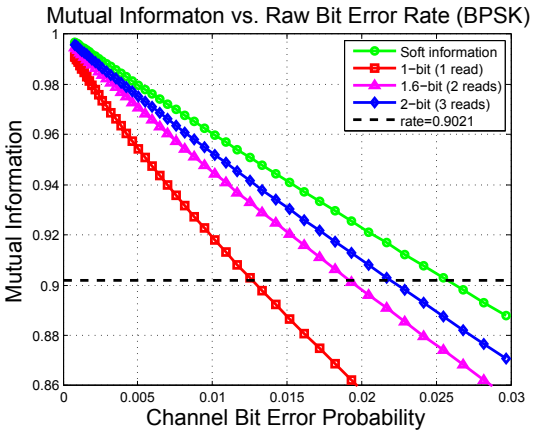


Fig. 4: Mutual information provided by different quantizations for SLC. The dashed horizontal line indicates the operating rate of our simulations. When a mutual information curve is below the dashed line, the read channel with that quantization cannot possibly support the attempted rate.

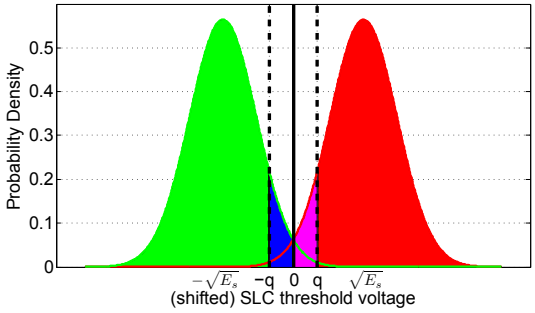


Fig. 5: Quantization model for SLC with 3 reads.

maximize the mutual information between the input and output of the corresponding equivalent channel.

A similar analysis can be applied if there are 3 reads per cell, which we describe next.

### B. SLC Flash Memory with 3 reads

Suppose we can have 3 reads for each cell, and each read corresponds to checking a comparator at a given word-line voltage. By symmetry, the word-line voltages should be chosen symmetrically as shown in Figure 5. An equivalent channel model with labeled crossover probabilities is given in Figure 6.

Similar to the analysis of Section III-A, the mutual information between the input and output can be calculated as

$$\begin{aligned} I(X;Y) =& H(Y) - H(Y|X) \\ =& H\left(\frac{p_1+p_4}{2}, \frac{p_2+p_3}{2}, \frac{p_3+p_2}{2}, \frac{p_4+p_1}{2}\right) \\ & - H(p_1, p_2, p_3, p_4), \end{aligned} \tag{2}$$

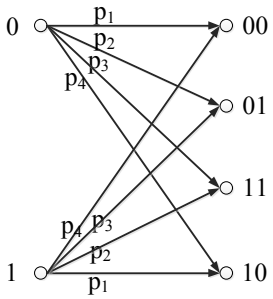


Fig. 6: Channel model for SLC with 3 reads.

where the crossover probabilities are computed as

$$\begin{aligned} p_1 &= 1 - Q\left(\frac{\sqrt{E_s} - q}{\sqrt{N_0/2}}\right), \\ p_2 &= Q\left(\frac{\sqrt{E_s} - q}{\sqrt{N_0/2}}\right) - Q\left(\frac{\sqrt{E_s}}{\sqrt{N_0/2}}\right), \\ p_3 &= Q\left(\frac{\sqrt{E_s}}{\sqrt{N_0/2}}\right) - Q\left(\frac{\sqrt{E_s} + q}{\sqrt{N_0/2}}\right), \text{ and} \\ p_4 &= Q\left(\frac{\sqrt{E_s} + q}{\sqrt{N_0/2}}\right). \end{aligned}$$

Note that the optimum $q^*$ is again a function of $E_s/N_0$. Figure 4 shows that the mutual information with 3 reads has a even larger than the mutual information with 2 reads, and is a little closer to the capacity with full soft information.

The above analysis can be easily extended to MLC flash memory. This is described in the next section for the case of 4-level MLC flash.

### C. 4-level MLC Flash Memory with 6 reads

For 4-level MLC flash memory, each cell can store 2 bits of information. Extending the previously introduced SLC model in the natural way, we model the MLC read channel as a 4-PAM signal with AWGN noise. To minimize the raw bit error rate, we also use the Gray labeling $(00, 01, 11, 10)$ for these four levels. Typically in 4-level MLC flash, each cell is compared to 3 word-line voltages and thus the output of the comparator has 4 values (i.e., four distinct quantization regions). If we consider three additional word-line voltages (for a total of six), the threshold voltage can be quantized to seven distinct values as shown in Figure 7. An equivalent channel model is given in Figure 8. (Although not shown, the crossover probabilities for the channel model are defined symmetrically in the lower half of the figure.)

Similar to the analysis of Section III-A, the mutual infor-

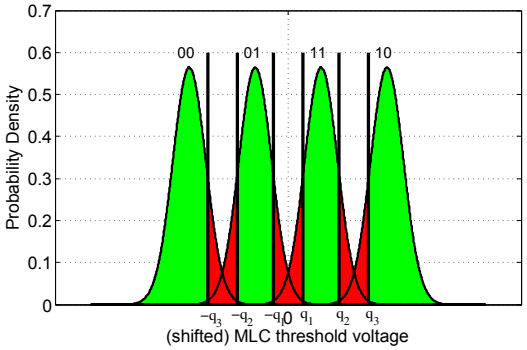


Fig. 7: Channel model for 4-MLC with 6 reads.

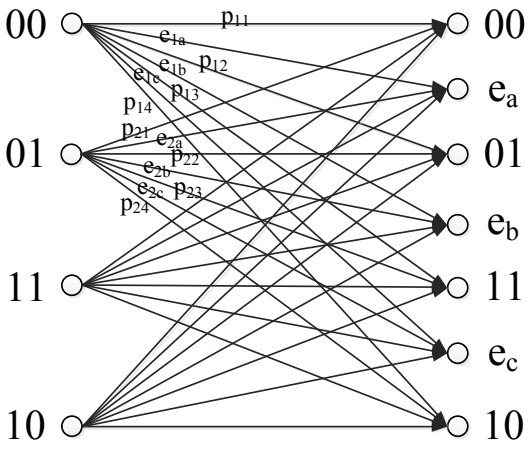


Fig. 8: Quantization model for 4-MLC with 6 reads.

mation between the input and output can be calculated as

$$
\begin{aligned}
&I(X;Y)\\
=&H(Y)-H(Y|X)\\
=&H\left(\frac{p_{11}+p_{21}+p_{24}+p_{14}}{4},\frac{p_{11}+p_{22}+p_{23}+p_{13}}{4},\right.\\
&\frac{p_{13}+p_{23}+p_{22}+p_{12}}{4},\frac{p_{14}+p_{24}+p_{21}+p_{11}}{4},\\
&\frac{e_{1a}+e_{2a}+e_{2c}+e_{1c}}{4},\frac{e_{1b}+e_{2b}+e_{2b}+e_{1b}}{4},\\
&\left.\frac{e_{1c}+e_{2c}+e_{2a}+e_{1a}}{4}\right)\\
&-\frac{1}{2}H(p_{11},p_{12},p_{13},p_{14},e_{1a},e_{1b},e_{1c})\\
&-\frac{1}{2}H(p_{21},p_{22},p_{23},p_{24},e_{2a},e_{2b},e_{2c}),
\end{aligned}
\tag{3}
$$

where all of the crossover probabilities can be calculated in the same manner as those in Sections III-A and III-B. Thus, in order to choose the optimal quantization levels $q_1, q_2$, and $q_3$ for a fixed $E_s/N_0$, we maximize the mutual information given in equation (3).

Figure 9 shows that the mutual information with 6 reads is much closer to the capacity with full soft information, than the mutual information with 3 reads.

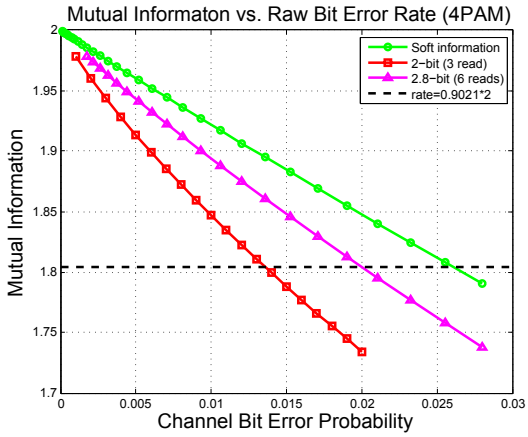


Fig. 9: Mutual information provided by different quantizations for MLC. The dashed horizontal line indicates the operating rate of our simulations. When a mutual information curve is below the dashed line, the read channel with that quantization cannot possibly support the attempted rate.

## IV. Simulation Results

In this section we demonstrate the benefits of LDPC decoding using soft information provided through multiple reads. A rate-0.9021 BCH code with block length $n = 9152$ and dimension $k = 8256$ provides a baseline for comparison. For our simulations, we use a rate-0.9021 irregular LDPC code with block length $n = 9118$ and dimension $k = 8225$. This LDPC code is designed with an optimal degree distribution for the additive white Gaussian noise channel [4]. Moreover, the code was designed using the ACE algorithm [9], and the stopping-set check algorithm [17] to optimize the LDPC matrix while maintaining the prescribed degree distribution. All of the simulations were performed using a sequential belief propagation decoder.

Frame error rate (FER) is plotted vs. channel bit error probability (raw bit error probability). The frame sizes are the block lengths, $k = 8256$ for BCH and $k = 8225$ for LDPC.

Since the BCH decoder is limited to using hard decisions from the comparator, we first simulate the LDPC decoder using only hard decisions in order to make a fair baseline comparison. The BCH and LDPC 1-bit curves in Figures 10 and 11 show that the LDPC code outperforms the BCH code in this range of page error rates, but not significantly so. The red dashed vertical line gives the Shannon limit for operating at rate 0.9021 on this channel with a single bit of reliability information.

Providing an additional bit of reliability information to the LDPC decoder through increased quantization resolution improves performance significantly, recovering almost all of the performance available with full soft information. This can be observed by comparing the Shannon limits corresponding to varying levels of soft information with their respective simulations in Figures 10 and 11.

We also plot the frame error rate versus the traditional

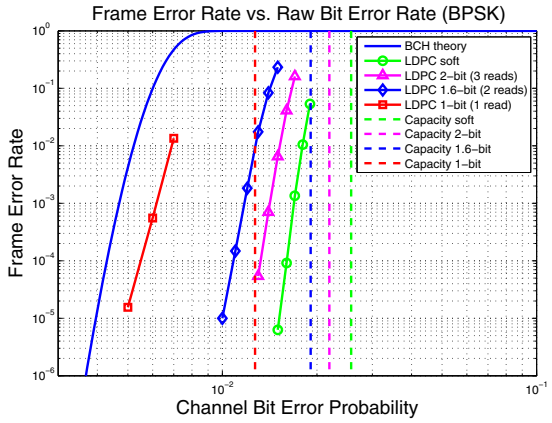


Fig. 10: Simulation results for SLC.

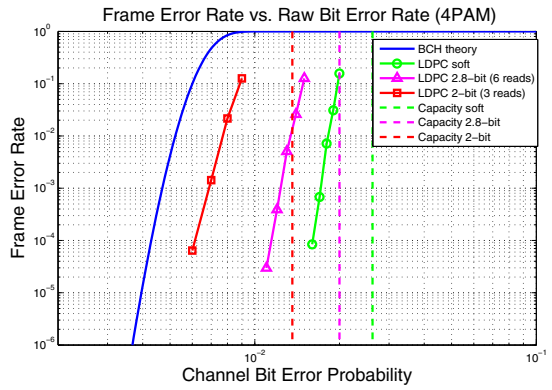


Fig. 11: Simulation results for 4-level MLC.

signal-to-noise ratio $E_s/N_0$ in Figure 12 for SLC, where each $E_s/N_0$ corresponds to an equivalent raw bit error rate in Figure 10.

Of course the BCH code will also benefit from the use of soft information. However, we were unable to perform simulations of a BCH decoder utilizing soft information (such as erasures) for inclusion in this paper.

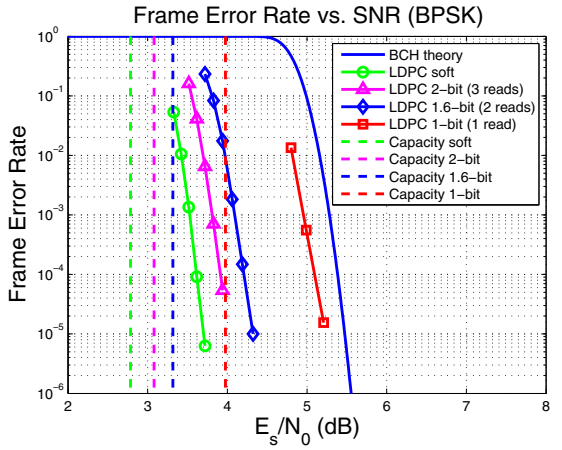


Fig. 12: Simulation results for SLC.

## V. Conclusion

This paper explores the benefit of using soft information in an LDPC decoder for flash memory. Using a small amount of soft information improves the performance of LDPC codes significantly and demonstrates a clear performance advantage over conventional BCH codes. In order to maximize the performance benefit of the soft information, we develop a word-line-voltages-selection method that maximizes the mutual information between the input and output of the equivalent read channel. Possible directions for future research include extending these results to more precise channel models, the design of better high-rate LDPC codes for flash memory, and the analysis of the corresponding error-floor properties.

## References

[1] Y. Li, S. Lee, and et al. A 16 Gb 3b/cell NAND Flash Memory in 56nm With 8MB/s Write Rate. In *Proc. of ISSCC*, pages 506–632, Feb. 2008.

[2] C. Trinh, N. Shibata, and et al. A 5.6MB/s 64 Gb 4b/Cell NAND Flash Memory in 43nm CMOS. In *Proc. of ISSCC*, page 246, Feb. 2009.

[3] J.-D. Lee, S.-H. Hur, and J.-D. Choi. Effects of floating-gate interference on NAND flash memory cell operation. *IEEE Electron Device Letters*, 23(5):264–266, May 2002.

[4] T. Richardson, M. Shokrollahi, and R. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Trans. Inform. Theory*, 47(2):616–637, Feb. 2001.

[5] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti. Introduction to Flash Memory. *Proc. IEEE*, 91(4), April 2003.

[6] Y. Maeda and K. Haruhiko. Error Control Coding for Multilevel Cell Flash Memories Using Nonbinary Low-Density Parity-Check Codes. In *24th IEEE Int. Symp. on Defect and Fault Tolerance in VLSI Systems*, Chicago, IL, Oct. 2009.

[7] S. Li and T. Zhang. Improving Multi-Level NAND Flash Memory Storage Reliability Using Concatenated BCH-TCM Coding. *IEEE Trans. VLSI Systems*, 18(10):1412–1420, Oct. 2010.

[8] G. Dong, N. Xie, and T. Zhang. On the Use of Soft-Decision Error-Correcting Codes in NAND Flash Memory. *IEEE Trans. Circ. and Sys.*, 58(2):429–439, Feb. 2011.

[9] T. Tian, C. Jones, J. D. Villasenor, and R. D. Wesel. Selective Avoidance of Cycles in Irregular LDPC Code Construction. *IEEE Trans. Comm.*, 52(8):1242–1247, Aug. 2004.

[10] X.-Y. Hu, E. Eleftheriou, and D.-M. Arnold. Progressive edge-growth Tanner graphs. In *Proc. IEEE GLOBECOM*, San Antonio, TX, Feb. 2001.

[11] T. Richardson. Error-floors of LDPC codes. In *Proc. 41st Annual Allerton Conf.*, Monticello, IL, Oct. 2003.

[12] J. Wang, L. Dolecek, and R.D. Wesel. Controlling LDPC Absorbing Sets via the Null Space of the Cycle Consistency Matrix. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, Kyoto, Japan, June. 2011.

[13] J. Wang, L. Dolecek, and R. D Wesel. LDPC Absorbing Sets, the Null Space of the Cycle Consistency Matrix, and Tanner's Constructions. In *Proc. Info. Theory and Appl. Workshop*, San Diego, CA, Feb. 2011.

[14] M. Ivkovic, S. K. Chilappagari, and B. Vasic. Eliminating trapping sets in low-density parity-check codes by using Tanner graph covers. *IEEE Trans. Inform. Theory*, 54(8):3763–3768, Aug. 2008.

[15] D. V. Nguyen, B. Vasic, and M. Marcellin. Structured LDPC Codes from Permutation Matrices Free of Small Trapping Sets. In *Proc. IEEE Info. Theory Workshop (ITW)*, Dublin, Ireland, Sept. 2010.

[16] Q. Huang, Q. Diao, S. Lin, and K. Abdel-Ghaffar. Cyclic and quasi-cyclic LDPC codes: new developments. In *Proc. Info. Theory and Appl. Workshop*, San Diego, CA, Feb. 2011.

[17] A. Ramamoorthy and R. D. Wesel. Construction of Short Block Length Irregular LDPC Codes. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, Paris, France, June. 2004.