All-Bit-Line MLC Flash Memories: Optimal Detection Strategies

Xiujie Huang*, Meysam Asadi†, Aleksandar Kavcic†, and Narayana (Prasad) Santhanam†
*Department of Computer Science, Jinan University, Guangzhou, GD 510632 China
†Department of Electrical Engineering, University of Hawaii, Honolulu, HI 96822 USA
Email: t_xiujie@jnu.edu.cn, {masadi, kavcic, nsanthan}@hawaii.edu

***Abstract*—We are concerned with the optimal detector design for the all-bit-line MLC flash memory. We provide a channel model of the MLC flash memory, where the channel parameters are mathematically tractable. Then we present an optimal maximum a-posteriori sequence detector. The optimal detector can be executed over a trellis whose branch metrics can be computed by using Fourier transforms of analytically computable characteristic functions (corresponding to likelihood functions). The soft-output detectors for both simple one-dimensional channel models and more realistic page-orientated two-dimensional channel models are derived. Simulation results show not only that the soft-output detector has the same hard-output bit-error-rate performance as some previously known detectors did, but that the soft-output detector outperforms previously known detectors by a gain of 0.23 dB.**

I. Introduction

As a class of non-volatile data storage system, multi-level cell (MLC) flash memories have been used in more and more consumer electronics and applications, from music players to solid-state disk drives. The growing usage of MLC flash memories is due to their low costs and high densities resulting from the continuous improvements in scaling technology that shrinks the sizes of transistors and MLC technology that stores more than one bit per cell. However, the scaling technology continues to increase the cell density, which dramatically enhances intercell interference (ICI). Moreover, the MLC technology narrows the width of threshold voltage for each level and reduces the margins between adjacent levels in a cell, which results in degradation of reliability. These two issues complicate detection in flash memories.

Up to date, some hard detection strategies including the *post-compensation* method [1] and the *coupling canceller* method [2] were presented. In particular, the post-compensation detector [1] and the coupling canceller detector [2] subtract estimates of ICI from the noisy observations of the channel output (i.e., the sensed voltage of each cell). Some "soft" detection strategies were also presented by using uniform and/or non-uniform channel output quantizations and then introducing the soft (decision) information in [3–5]. To the best of our knowledge, there exists no open literature providing an exact soft-output detector without channel output quantization in MLC flash memories. Actually, the optimal soft detection is unknown because there is no explicit and precise channel modeling and no knowledge of the exact statistics of the channel model.

In this paper, we are mostly concerned with the optimal detector design for the MLC flash memory. To this end, we first introduce the basics of the all-bit-line (ABL) MLC flash memory in Sec. II-A, and then provide channel models of the ABL MLC flash memory including the one-dimensional (1D) model with *causal output memory* and the two-dimensional (2D) model with *anticausal output memory* in Sec. II-B and Sec. II-C, respectively. Second, by investigating the exact statistics of the channel, we present a Viterbi-like sequence detector, i.e., the optimal *maximum a posteriori* (MAP) detector, for the 1D model in Sec. III. The exact statistics of the channel necessary for implementing the optimal MAP detector can be obtained by using the fast Fourier transform (FFT). Third, we derive a sub-optimal detector, called GA-MAP detector, by using a Gaussian approximation in Sec. IV. Both of the presented detectors can be extended for the 2D model. Fourth, in Sec. V, we utilize simulation results to show that the presented detectors outperform the existing detectors in the literature. Finally, we conclude this work in Sec. VI.

II. Channel Model

A. All-Bit-Line (ABL) MLC Flash Memory Basics

An MLC flash memory consists of lots of floating-gate transistors. Each transistor is a cell. By applying a certain voltage that falls into one of the multiple required voltage ranges to the floating gate of a cell, the charge is maintained and then a data is stored in the cell. All cells are hierarchically organized in arrays, blocks and page partitions. As the development of the reading and sensing techniques, the emerging ABL architecture of the organization is gaining more and more attention due to its advantages of low energy, low ICI and high read/verified speed [6, 7]. The ABL architecture allows all cells along a word-line to be read and programmed in parallel. Due to the parasitic capacitance coupling effects

978-1-4799-2003-7/14/$31.00 ©2014 IEEE

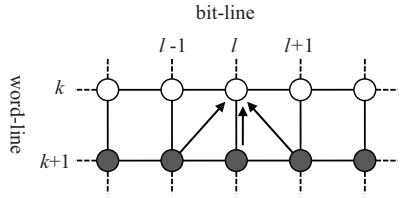

Fig. 1. The all-bit-line architecture and the ICI, where each cell is affected by three neighbors that are programmed after itself.

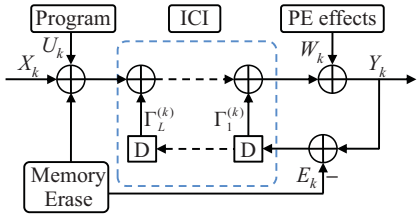

Fig. 2. 1D causal channel model.

(say ICI) [8], a (victim) cell is affected on its threshold voltage by the voltage shifts of those cells that are programmed after the victim cell. According to [1], each cell has three interfering neighbors that are programmed after itself, as shown in Fig. 1, and the coupling ratio is usually modeled by a (truncated) Gaussian distribution whose parameters depend on the distance between cells.

As the smallest unit that can be erased is a block, it is necessary to erase all cells in a block before being able to be programmed again. The threshold voltage of the erased cell tends to be Gaussian distributed [1]. The programming scheme is called *incremental step pulse program* (ISPP) [9–11], which is a one-way iterative technique that provide a series of program pulse steps and verify the amount of voltage carried at each cell after each programming step. The threshold voltage deviation of a programmed cell tends to be uniform distributed [10]. In addition, the threshold voltage in each cell is also affected by the program/erase (PE) cycling due to the device age, the number of PE cycles and the trapping and detrapping ability of the interface at the transistor gate. This effect is usually modeled by a Gaussian random variable with parameters dependent on the input voltage at the floating gate [10].

B. 1D Causal Channel Model with Output Memory

In this subsection, we provide a stochastic channel model that includes all the important features of the ABL MLC flash memory as aforementioned. We first present a simple *1D causal* channel model, as shown in Fig. 2, because it allows us to formulate the optimal detector in the universal manner, namely for a 1D causal channel.

Let $k \in \mathbb{Z}$. The channel input and output are denoted by X_k and Y_k, which are the intended voltage and the final voltage of the k-th cell, respectively. Let X_k take values from a finite alphabet $\mathcal{X}=\{v_0, v_1, \ldots, v_{m-1}\}$. Let $E_k \sim \mathcal{N}(\mu_e, \sigma_e^2)$ be the erased-state noise of the k-th cell. The 1D channel model[1] is

$$Y_k = X_k + \sum_{\ell=1}^{L} \Gamma_\ell^{(k)} \left(Y_{k-\ell} - E_{k-\ell}\right) + W_k + U_k, \quad (1)$$

where

a) $\Gamma_\ell^{(k)} \sim \mathcal{N}(\gamma_\ell, g_\ell)$ is a fading-like coefficient that models *causal* ICI effect from the $(k-\ell)$-th cell towards the k-th cell (victim cell).
b) L is the output memory, which implies that Y_k is affected by its L neighbors $Y_{k-1}, Y_{k-2}, \ldots, Y_{k-L}$.
c) $U_k \sim \mathcal{U}\left[-\Delta/2\,,\, \Delta/2\right]$ denotes the programming noise at the k-th cell, resulting from the ISPP scheme.
d) W_k is the observation noise due to the PE cycling, and is modeled to be Gaussian as $W_k \sim \mathcal{N}(0, \sigma_w^2)$.

We assume that all random variables $\Gamma_\ell^{(k)}, E_{k-\ell}, W_k$ and U_k are independent for all k and all ℓ. Note that several studies [10, 12] have shown that different inputs (levels) $v_0, v_1, \cdots, v_{m-1}$ give rise to different statistics of channel noises W_k and U_k. In this case of input-dependent noises, when the channel input $X_k = v_i$ is given, we denote the conditional distributions of W_k and U_k by $\mathcal{N}(0, \sigma_w^2(i))$ and $\mathcal{U}\left[\frac{-\Delta(i)}{2}, \frac{\Delta(i)}{2}\right]$, respectively.

Remark 1. *The channel in (1) is actually a general 1D causal ICI channel which belongs to the class of channels with memory. Note that such memory depends on the channel outputs Y_{k-L}^{k-1}. Although the output process Y_k (for a stationary memoryless (i.i.d.) input process X_k) is Markov, this channel does not belong to the well-known class of Markov finite-state channels [13, 14].*

Since the channel has memory, it is obvious that a capacity-achieving process X_k may also need to have memory. For this reason, we assume that X_k is a Markov

[1]The model in (1) does not take the same form as Eqn. (1) in [1], which is in terms of the "voltage shift". The difference $Y_{k-\ell} - E_{k-\ell}$ in our form (1) is equal to the "voltage shift" in [1]. Instead of calling explicit attention to the "voltage shift", we call explicit attention to the channel outputs Y_{k-L}^{k-1}, because for the purpose of detector design, it is much better to explicitly call attention to the channel outputs as the detector specifically operates on the channel outputs. The major differences of our model in (1) and the model in [1] are as follows:

i) We assume that $\Gamma_\ell^{(k)}$ are Gaussian, whereas Dong *et al.* [1] assumed the corresponding variables to be truncated Gaussian. Note that these two distributions almost have the same behavior in the common support range, but the Gaussian distribution is easier to track analytically.
ii) We assume that the observation noise $U_k + W_k$ is a mixture of uniform and Gaussian noises. The pdf of this mixture is actually very similar (though not identical) to the pdf shown in [1, Fig. 4].

process of order M. That is,

$$P_{X_k|X^{k-1}}(x_k|x^{k-1}) = P_{X_k|X_{k-M}^{k-1}}(x_k|x_{k-M}^{k-1}). \quad (2)$$

Note that the optimization problem of finding the actual optimal (capacity-achieving) Markov input process X_k is beyond the scope of this paper.

C. 2D Anticausal Page-Oriented Channel Model

An actual flash memory is not 1D, but rather 2D because the channel is page-oriented. In addition, the actual flash memory is not causal, but rather anticausal, because ICI is an anticausal effect as the victem cell is affected by those cells that are programmed *after* it. As discussed in Sec. II-A and shown in Fig. 1, a victim cell of the 2D ABL MLC flash memory is only affected by three anticausal neighbors. Let (k,ℓ) denote the location of a cell that locates at the k-th word-line and the ℓ-th bit-line. We denote by $\mathcal{I}_{(k,\ell)}$ the indices of the anticausal neighborhood for the (k,ℓ)-th victim cell[2], i.e.,

$$\mathcal{I}_{(k,\ell)} \triangleq \{(k+1,\ell-1),(k+1,\ell),(k+1,\ell+1)\}. \quad (3)$$

Therefore, an appropriate 2D channel model is given as

$$Y_{(k,\ell)} = X_{(k,\ell)} + \sum_{(a,b)\in\mathcal{I}_{(k,\ell)}} \Gamma_{(a,b)}^{(k,\ell)}\big(Y_{(a,b)} - E_{(a,b)}\big) + W_{(k,\ell)} + U_{(k,\ell)}. \quad (4)$$

As the channel does have 2D-memory, it is reasonable to expect that the capacity-achieving input process $X_{(k,\ell)}$ may have 2D-memory. However, optimizing the input processes (even 2D Markov processes) to maximize the information rate of a 2D channel with memory is a difficult problem, which is outside of the scope of this paper.

III. Viterbi-like 1D Sequence Detection

In this section, we address the Viterbi-like MAP detector of the input sequence x_1^n from output realizations y_1^n. Specifically, the MAP detection of x_1^n is the sequence $\hat{x}_1^n$ that maximizes the joint conditional pdf, i.e.,

$$\hat{x}_1^n = \arg\max_{x_1^n\in\mathcal{X}^n} f(x_1^n, y_1^n | x_{1-M}^0, y_{1-L}^0). \quad (5)$$

By factorizing the pdf $f(x_1^n, y_1^n|x_{1-M}^0, y_{1-L}^0)$, we find that the MAP detected sequence is equal to

$$\hat{x}_1^n = \arg\min_{x_1^n\in\mathcal{X}^n} \sum_{k=1}^{n} \underbrace{\big[-\ln\big(P\big(x_k|x_{k-M}^{k-1}\big) f\big(y_k|x_k, y_{k-L}^{k-1}\big)\big)\big]}_{\Lambda_{\rm MAP}(x_{k-M}^k, y_{k-L}^k)}. \quad (6)$$

This detector can be executed over the time-invariant trellis whose *state* and *branch metric* at time k are defined as $S_k \triangleq X_{k-M+1}^k$ and the term inside the summation $\Lambda_{\rm MAP}(x_{k-M}^k, y_{k-L}^k)$ in (6), respectively.

[2]For the marginal cases of the first bit-line, the last bit-line and the last word-line, we assume $\mathcal{I}_{(k,\ell)} = \{(k+1,\ell),(k+1,\ell+1)\}$, $\mathcal{I}_{(k,\ell)} = \{(k+1,\ell-1),(k+1,\ell)\}$ and $\mathcal{I}_{(k,\ell)} = \emptyset$, respectively.

If the channel input process X_k is i.i.d., then an equivalent of the MAP detector in (6) is, for all k,

$$\hat{x}_k = \arg\min_{x_k\in\mathcal{X}} \underbrace{\big\{-\ln\big(P(x_k) f\big(y_k|x_k, y_{k-L}^{k-1}\big)\big)\big\}}_{\Lambda_{\rm MAP}(x_k, y_{k-L}^k)}. \quad (7)$$

In this case, a trellis-based detector is not needed, and the optimal detector can be executed on a symbol-by-symbol basis. That is, the detection of the current input X_k can be done without the knowledge of prior channel input realizations x_{k-M}^{k-1}.

Remark 2. *The detectors in (6) and (7) can be slightly altered and applied to the 2D anticausal channel for the cases of 2D Markov channel inputs and i.i.d. channel inputs, respectively. However, the optimality is unknown because there exists no equivalents of the Viterbi detectors in 2D channel.*

A. Calculation of the Characteristic Function

From (6) and (7), it is clear that evaluating the branch metric $\Lambda_{\rm MAP}(\cdot,\cdot)$ requires evaluating the conditional pdf $f(y_k|x_k, y_{k-L}^{k-1})$. However, it is intractable to compute $f(y_k|x_k, y_{k-L}^{k-1})$ analytically. Instead, we calculate the conditional characteristic function of Y_k, denoted by $G_{Y_k|X_k,Y_{k-L}^{k-1}}(t)$, under the assumption that $X_k = x_k$ and $Y_{k-L}^{k-1} = y_{k-L}^{k-1}$ are given. Then the pdf $f(y_k|x_{k-M}^k, y_{k-L}^{k-1})$ for each realization y_k can be derived by taking the Fourier transform of the characteristic function as

$$f(y_k|x_k, y_{k-L}^{k-1}) = \int_{-\infty}^{\infty} G_{Y_k|X_k,Y_{k-L}^{k-1}}(t) e^{-iy_k t} dt. \quad (8)$$

To compute the conditional characteristic function $G_{Y_k|X_k,Y_{k-L}^{k-1}}(t)$, we rewrite the channel model

$$Y_k = \underbrace{X_k + W_k}_{R} + U_k + \sum_{\ell=1}^{L} \underbrace{\Gamma_\ell^{(k)}(Y_{k-\ell} - E_{k-\ell})}_{Z_\ell}. \quad (9)$$

When $X_k = x_k$ is given, R is Gaussian $\mathcal{N}\big(\mu_R, \sigma_R^2\big)$ where $\mu_R = \mathbb{E}[R|X_k = x_k] = x_k$ and $\sigma_R^2 = {\rm Var}[R|X_k = x_k] = \sigma_w^2$. Hence, the conditional characteristic function of R is

$$G_{R|X_k}(t) = \exp\left(-\frac{1}{2}\sigma_R^2 t^2 + i\mu_R t\right). \quad (10)$$

When $Y_{k-\ell} = y_{k-\ell}$ is given, Z_ℓ is the product of two Gaussians of $\Gamma_\ell^{(k)}$ and $\Omega \triangleq (y_{k-\ell} - E_{k-\ell})$. Then, the conditional characteristic function of Z_ℓ, denoted by $G_{Z_\ell|Y_{k-\ell}}(t)$ (see the Appendix in the extension [15] of this paper for detail, where the conventional even/odd bit-line MLC flash memory is considered), is

$$\begin{aligned} & G_{Z_\ell|Y_{k-\ell}}(t) \\ &= \frac{\exp\left(\frac{-t^2\left((y_{k-\ell}-\mu_e)^2 g_\ell + \gamma_\ell^2\sigma_e^2\right) + 2it\,(y_{k-\ell}-\mu_e)\gamma_\ell}{2\left(1+g_\ell\sigma_e^2 t^2\right)}\right)}{\sqrt{1+g_\ell\sigma_e^2 t^2}} \\ &\triangleq \frac{1}{\sqrt{1+g_\ell\sigma_e^2 t^2}} \exp\left(\Phi_\ell(t)\right). \end{aligned} \quad (11)$$

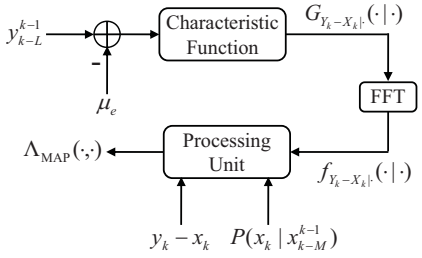

Fig. 3. Branch metric computation using the FFT.

Finally, utilizing the conditional independence of R and Z_ℓ (given y_{k-L}^{k-1} and x_k), we get

$$\begin{aligned} &G_{Y_k|X_k,Y_{k-L}^{k-1}}(t) \qquad (12)\\ &= G_{R|X_k}(t)\, G_{U_k}(t) \prod_{\ell=1}^{L} G_{Z_\ell|Y_{k-\ell}}(t)\\ &= \frac{\operatorname{sinc}(t\Delta/2)}{\sqrt{\prod_{\ell=1}^{L}(1+g_\ell\sigma_e^2t^2)}} \exp\!\left(-\frac{1}{2}\sigma_R^2t^2 + i\mu_R t + \sum_{\ell=1}^{L}\Phi_\ell(t)\right). \end{aligned}$$

B. FFT Implementation

An equivalent form of $\Sigma_{\ell=1}^{L}\Phi_\ell(t)$ in (12) is

$$\sum_{\ell=1}^{L}\Phi_\ell(t) = \frac{\sum_{\ell=1}^{2L} C_\ell(y_{k-L}^{k-1})t^\ell + q(t)}{2\prod_{\ell=1}^{L}(1+g_\ell\sigma_e^2t^2)}$$

where

$$C_\ell(y_{k-L}^{k-1}) = \begin{cases} \sum_{j=1}^{L}\alpha_j^{(\ell)}(y_{k-j}-\mu_e), & \text{if } \ell \text{ is odd} \\ \sum_{j=1}^{L}\beta_j^{(\ell)}(y_{k-j}-\mu_e)^2, & \text{if } \ell \text{ is even} \end{cases} \qquad (13)$$

and $q(t)$ is a polynomial (whose coefficients are independent of y_{k-L}^{k-1}). It is clear from (13) that the sufficient statistics to compute the characteristic function in (12) are the $2L$ outputs from two types of finite impulse response (FIR) filters where:

a) inputs are signals $(y_{k-j}-\mu_e)$, if ℓ is odd.
b) inputs are signals $(y_{k-j}-\mu_e)^2$, if ℓ is even.

Hence, there are $2L$ desired FIR filters.

Since the pdf is the Fourier transform of the characteristic function, as shown in (8), we are able to numerically compute the conditional pdf $f(y_k|x_k,y_{k-L}^{k-1})$ using FFT and then obtain the branch metric $\Lambda_{\text{MAP}}(\cdot,\cdot)$ in (6). Exactly speaking, we only need to compute one FFT for each trellis section. That is, the FFT is the same for all branches of the trellis section, but the actual branch metrics are obtained by sampling the FFT at different points as illustrated in Fig. 3.

IV. A Sub-Optimal Detector Using Gaussian-Approximation

In this section, we utilize a Gaussian approximation to give a simplified procedure to compute an estimate of $\Lambda_{\text{MAP}}(\cdot,\cdot)$, which requires only two FIR filters. To the end, we rewrite the channel model as

$$Y_k = \underbrace{X_k + \sum_{\ell=1}^{L}\Gamma_\ell^{(k)}(Y_{k-\ell}-E_{k-\ell}) + W_k}_{V_k} + U_k. \quad (14)$$

As discussed in [3], the Gaussian $\Gamma_\ell^{(k)}$ is usually approximated by a constant. In this case, when $X_k = x_k$ and $Y_{k-L}^{k-1} = y_{k-L}^{k-1}$ are given, then V_k is Gaussian since it is a linear combination of independent Gaussians. Hence, we assume that V_k can be approximated by a Gaussian $V_k^{(G)} \sim \mathcal{N}\big(\mu_G(k), \sigma_G^2(k)\big)$ where

$$\begin{aligned} \mu_G(k) &= \mathbb{E}\big[V_k | Y_{k-L}^{k-1} = y_{k-L}^{k-1}, X_k = x_k\big]\\ &= x_k + \sum_{\ell=1}^{L}\gamma_\ell(y_{k-\ell}-\mu_e), \qquad (15)\\ \sigma_G^2(k) &= \operatorname{Var}\big[V_k | Y_{k-L}^{k-1} = y_{k-L}^{k-1}, X_k = x_k\big]\\ &= \sum_{\ell=1}^{L}\big(g_\ell\big(\sigma_e^2 + (y_{k-\ell}-\mu_e)^2\big) + \sigma_e^2\gamma_\ell^2\big) + \sigma_w^2. \qquad (16) \end{aligned}$$

By convolving the Gaussian distribution of $V_k^{(G)}$ and the uniform distribution of U_k, the approximate conditional pdf $f^{(G)}\left(y_k|x_k, y_{k-L}^{k-1}\right)$ is obtained as follows

$$\begin{aligned} &f^{(G)}(y_k|x_k, y_{k-L}^{k-1})\\ &= \frac{1}{\Delta}\left[Q\left(\frac{y_k - \mu_G - \frac{\Delta}{2}}{\sigma_G}\right) - Q\left(\frac{y_k - \mu_G + \frac{\Delta}{2}}{\sigma_G}\right)\right], \quad (17) \end{aligned}$$

where $Q(\zeta) = \frac{1}{\sqrt{2\pi}}\int_\zeta^\infty \exp(-\frac{\eta^2}{2})\, d\eta$ is the Q-function.

So, clearly by examining (15)-(17), we conclude that the sufficient statistics for computing $\Lambda_{\text{MAP}}^{(G)}(\cdot,\cdot)$ are y_k and the outputs θ_k and ν_k from the two FIR filters as

$$\begin{aligned} \theta_k &= \textstyle\sum_{\ell=1}^{L}\gamma_\ell\,(y_{k-\ell}-\mu_e),\\ \nu_k &= \textstyle\sum_{\ell=1}^{L} g_\ell\,(y_{k-\ell}-\mu_e)^2. \end{aligned} \qquad (18)$$

Hence, instead of using $2L$ filters to compute $\Lambda_{\text{MAP}}(\cdot,\cdot)$ in Sec. III-B, we only need to use two filters to compute the approximation $\Lambda_{\text{MAP}}^{(G)}(\cdot,\cdot)$. Actually, the computation of $\Lambda_{\text{MAP}}^{(G)}(\cdot,\cdot)$ does not require generating FFTs, but can be implemented using sample DSP components such as multipliers and adders $\big($see (15)-(18)$\big)$.

It is interesting to note that some prior-art detectors can be obtained by further approximations of the Gaussian approximation given in this section. For example, the hard post-compensation detector in [1] can be obtained as a symbol-by-symbol detector (when X_k is i.i.d.) by computing the decision variables $y_k - \theta_k$ and heuristically determining the decision thresholds to achieve minimum BER. Similarly, the soft-decision detectors in [3, 7] can be obtained if instead of $\sigma_G^2(k)$ in (16), we use $\sigma_e^2\gamma_1^2 + \sigma_w^2$. Hence, the detectors in [3, 7]

TABLE I
PARAMETERS OF THE 4-LEVEL FLASH MEMORY

i	0	1	2	3
ith level v_i	1.1	2.7	3.3	3.9
$\Delta(i)$	0	0.3	0.3	0.3
$\sigma_w(i)$	0.35σ	0.03σ	0.03σ	0.03σ

are sub-optimal versions of the detector proposed here (particularly, if used as precursors to a soft-in-soft-out decoder for LDPC codes) as shown explicitly in Sec. V.

V. Simulation Results and Discussion

To show the performances of the present detectors, we utilize a 4-level flash memory channel, where the channel input X_k is i.i.d. with the pmf $P(X_k = v_i) = 0.25$ over the alphabet $\mathcal{X} = \{v_0, v_1, v_2, v_3\}$. The parameters of the 4-level flash memory (2D channel) with signal-dependent noises are given in Table I.

Following [1], we let s be the *intercell coupling strength factor* and assume the Gaussian random coupling ratios $\Gamma^{(k,\ell)}_{(a,b)}$ (see (4)) as $\Gamma^{(k,\ell)}_{(k+1,\ell)} \sim \mathcal{N}(\gamma_v, g_v)$, $\Gamma^{(k,\ell)}_{(k+1,\ell-1)} \sim \mathcal{N}(\gamma_d, g_d)$ and $\Gamma^{(k,\ell)}_{(k+1,\ell+1)} \sim \mathcal{N}(\gamma_d, g_d)$, where $\gamma_v = 0.08s$, $\gamma_d = 0.006s$, $g_i = 0.09\gamma_i^2$ for $i \in \{v, d\}$. Here, the subscripts v and d mean the vertical and diagonal interfering as shown in Fig. 1, respectively.

In the first simulation scenario, we fix $\sigma = 1$ (see Table I) and vary the coupling strength factor s. The bit-error-rate (BER) performances of the MAP detector, the GA-MAP detector, the post-compensation detector [1] and the raw detector [1] are shown in Fig. 4.

In the second simulation scenario, we fix $s = 0.75$ and vary the parameter σ. By varying σ, we effectively vary the signal-to-noise ratio (SNR), defined as

$$\text{SNR} \triangleq \frac{1}{\sum_i P(X_k = v_i)\sigma_w^2(i)}. \tag{19}$$

Fig. 5 shows the BER performances for varying SNRs of the MAP detector, the GA-MAP detector, the post-compensation detector [1] and the raw detector [1].

Figs. 4 and 5 reveal that if the BER is the figure of merit, neither the MAP detector nor the GA-MAP detector outperforms the post-compensation detector (originally disclosed in [1]). Hence, to get a better sense of the quality of each detector, we must compare the qualities of their *soft* outputs. Here, we measure the quality of a detector's soft output as follows.

Let $\mathcal{O}_k$ represent the soft output of a detector when X_k was written, which is the probability vector

$$\mathcal{O}_k = (P(X_k = v_0|y_1^n), P(X_k = v_1|y_1^n), \cdots, P(X_k = v_{m-1}|y_1^n)).$$

Then, we define the *soft information quality* (SIQ)[3] of a detector as

$$q = I(X_k; \mathcal{O}_k). \tag{20}$$

As explained in [16], SIQ is the capacity of random linear block codes. Therefore, this quantity is proved to be the highest information rate achievable by a random LDPC code. Moreover, SIQ allows us to compare performances of codes without going through the complicated task of simulating the actual codes. For example, if SIQ of detector A is 0.5 dB better than SIQ of detector B, then a random LDPC code using outputs from detector A will outperform the same random LDPC code using outputs from detector B by 0.5 dB. In other words, if we use detector A, we can afford to use a 0.5 dB weaker code and achieve the same overall system performance.

Fig. 6 shows the SIQ curves of the MAP detector and the GA-MAP detector when the coupling strength factor

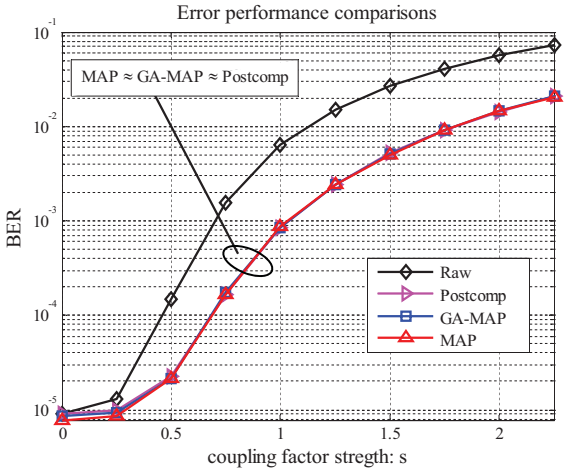

Fig. 4. BER comparisons for different detectors when the coupling factor strength is varying and $\sigma = 1$.

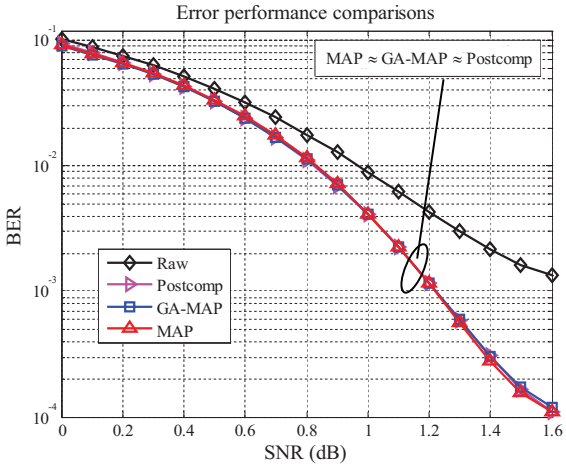

Fig. 5. BER comparisons for different detectors when the SNR is varying and the coupling factor strength is fixed at $s = 0.75$.

[3]The SIQ of the MAP detector is the *BCJR-once bound* in [16]. Furthermore, the SIQ can also be defined for a hard detector by defining $\mathcal{O}_k$ as a scalar estimate of the input $\mathcal{O}_k = \hat{X}_k \in \{v_0, v_1, \cdots, v_{m-1}\}$.

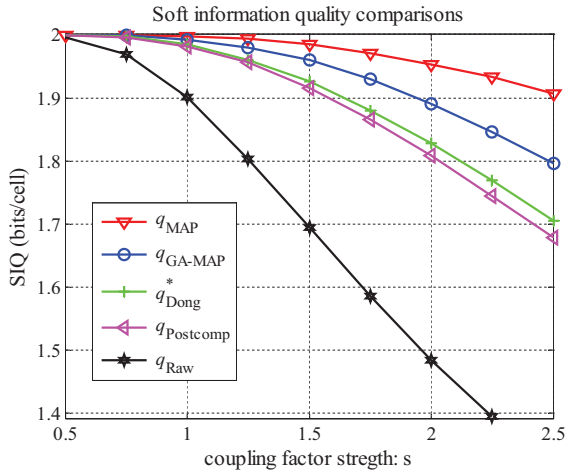

Fig. 6. SIQ comparisons for different detectors when the coupling factor strength is varying and $\sigma = 1$.

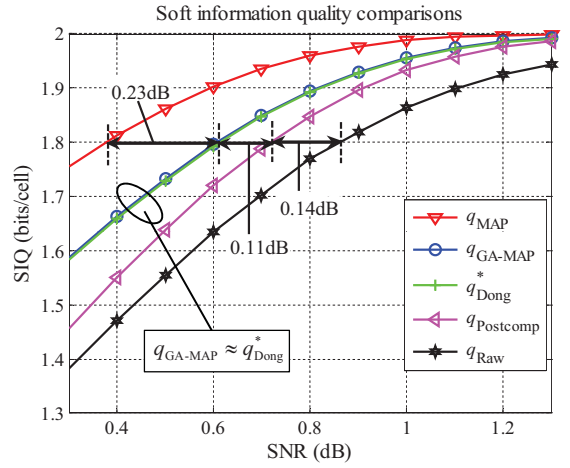

Fig. 7. SIQ comparisons for different detectors when the SNR is varying and the coupling factor strength is fixed at $s = 0.75$.

s varies for fixed $\sigma = 1$ (i.e., fixed SNR), while Fig. 7 shows the SIQ curves when the SNR varies for fixed $s = 0.75$. Also shown in Figs. 6 and 7 are SIQs of the post-compensation detector [1] and the raw detector [1]. Finally, the figures also show an upper bound on the SIQ of the soft-output detector presented in [3], denoted by q^*_{Dong}. At SIQ $= 1.8$ bits per cell (which corresponds to a code rate of 0.9 user bits per channel bit), though the GA-MAP detector has almost the same SIQ as q^*_{Dong}, the MAP detector outperforms known detectors by a gain of 0.23 dB, as shown in Fig. 7.

VI. Conclusion

We designed the optimal detector for the all-bit-line MLC flash memory. The optimal detector can be executed over the trellis, whose branch metrics are attainable using FFTs of analytically computable characteristic functions. At a small performance loss, a sub-optimal detector using a Gaussian approximation is also derived. Both detectors can be applied to the 1D causal channel with output memory and the 2D anticausal channel. Under the assumption of memoryless (i.i.d.) channel inputs, we implemented the presented detectors for the 2D anticausal channel and demonstrated their superior performances (particularly if executed as soft-output detectors) through simulation results.

Acknowledgment

This work was supported by LSI Corporation, by NSF Grants CCF-1018984 and EECS-1029081, and by Guangdong Leading Talent Program Grants 50502664 and 50502665. Parts of this work were performed while A. Kavcic was visiting the Institute of Network Coding at the Chinese University of Hong Kong.

References

[1] G. Dong, S. Li, and T. Zhang, "Using data postcompensation and predistortion to tolerate cell-to-cell interference in MLC NAND flash memory," *IEEE Trans. Circuits Syst.–I: Reg. Papers*, vol. 57, no. 10, pp. 2718–2728, Oct. 2010.

[2] D. Park and J. Lee, "Floating-gate coupling canceller for multi-level cell NAND flash," *IEEE Trans. Magn.*, vol. 47, no. 3, pp. 624–628, Mar. 2011.

[3] G. Dong and *et al.*, "On the use of soft-decision error-correction codes in NAND flash memory," *IEEE Trans. Circuits Syst.–I: Reg. Papers*, vol. 58, no. 2, pp. 429–439, Feb. 2011.

[4] J. Kim and W. Sung, "Low-energy error correction of NAND flash memory through soft-decision decoding," *EURASIP Journal on Advances in Signal Processing 2012*, 2012:195.

[5] J. Wang and *et al.*, "Soft information for LDPC decoding in flash: mutual-information optimized quantization," in *Proc. IEEE GLOBECOM 2011*, Houston, Texas, USA, Dec. 2011.

[6] L. Crippa and R. Micheloni, "Sensing circuits," in *Inside NAND Flash Memories*, R. Micheloni, L. Crippa, and A. Marelli, Eds. Springer, 2010, ch. 8, pp. 197–233.

[7] X. Wang, G. Dong, L. Pan, and R. Zhou, "Error correction codes and signal processing in flash memory," in *Flash Memories*, I. S. Stievano, Ed. InTech, Sept. 2011, ch. 3, pp. 57–82.

[8] 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*, vol. 23, no. 5, pp. 264–266, May 2002.

[9] B. Ricco and *et al.*, "Nonvolatile multilevel memories for digital applications," *Proceedings of the IEEE*, vol. 86, no. 12, pp. 2399–2420, Dec. 1998.

[10] R. Bez and *et al.*, "Introduction to flash memory," *Proceedings of the IEEE*, vol. 91, no. 4, pp. 489–502, Apri. 2003.

[11] A. Jiang and *et al.*, "On the capacity and programming of flash memories," *IEEE T-IT*, vol. 58, no. 3, pp. 1549–1564, Mar. 2012.

[12] X. Huang, A. Kavcic, X. Ma, G. Dong, and T. Zhang, "Optimization of achievable information rates and number of levels in multilevel flash memories," in *IARIA ICN 2013*, Seville, Spain, Jan. 27-Feb. 1 2013, pp. 125–131.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, Inc, 1968.

[14] J. Chen and P. H. Siegel, "Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels," *IEEE T-IT*, vol. 54, no. 3, pp. 1295–1303, Mar. 2008.

[15] M. Asadi, X. Huang, A. Kavcic, and N. P. Santhanam, "Optimal detector for multilevel NAND flash memory channels with intercell interference," IEEE J-SAC, Appear in May 2014.

[16] A. Kavčić and *et al.*, "Binary intersymbol interference channels: Gallager codes, density evolution, and code performance bounds," *IEEE T-IT*, vol. 49, no. 7, pp. 1636–1652, July 2003.