High-Throughput QC-LDPC Decoder with Cost-Effective Early Termination Scheme for Non-Volatile Memory Systems

Yu-Min Lin, Yu-Hao Chen, Ming-Han Chung and An-Yeu (Andy) Wu
Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan (R.O.C.)
{yumin,zerobigtree,cmh}@access.ee.ntu.edu.tw, andywu@cc.ee.ntu.edu.tw

Abstract—This paper presents a high-throughput layered min-sum quasi-cyclic LDPC (QC-LDPC) decoder for non-volatile memory systems (NVMs). A cost-effective column-based early termination (CB-ET) scheme is proposed to early terminate decoding process within iteration. The throughput improvement is 37.7\% compared to the state-of-the-art early termination scheme when raw bit error rate of flash memory is $3 \times 10^{-3}$. The QC-LDPC decoder with proposed early termination scheme is synthesized by TSMC 90nm CMOS technology, and the area overhead is only 2.20\%.

Keywords—Early termination scheme; layered LDPC decoder; Non-volatile memory;

I. INTRODUCTION

Flash memories become the prevailing storage element of non-volatile memory systems (NVMS) and are widely used in many consumer electronic products, including mobile devices, computers, and the promising solid-state drives (SSDs). The storage density of flash memories increases dramatically with the advanced manufacturing technology and multi-level cell (MLC) techniques which allow more bits to be stored in one cell. However, the high-capacity flash memories suffer severe degradation of reliability and endurance due to higher error rate of MLC [1]. Recently, low-density parity-check (LDPC) codes with soft-decision have been proposed as error-correcting codes (ECCs) to solve the reliability issues of NVMs [2]. Compared to the conventional algebraic codes, LDPC codes provide much superior performance. However, the iterative process of LDPC will not terminate until reaching the set maximum iteration. The redundant calculation of decoder leads to decreased throughput and increased power consumption. Hence, LDPC decoder with early termination is necessary for NVMs.

LDPC codes, which were invented by Gallager in 1962 [3], have received great attention for their outstanding error correcting performance. Quasi-cyclic LDPC (QC-LDPC) codes [5] are an essential branch of LDPC codes since their regular cyclic structures significantly reduce hardware requirements. On the other hand, turbo-decoding message passing (TDMP) [4], which is also called layered decoding, allows updating both check nodes and variable nodes concurrently based on the min-sum algorithm (MSA) for LDPC decoding. Moreover, the decoding convergence of layered min-sum algorithm is twice as fast as the standard two-phase message passing (TPMP). Therefore, we follow [5] to implement the layered min-sum high-throughput QC-LDPC decoder design.

Early termination scheme is a hot topic in LDPC decoding field since it can increase the throughput tremendously. Several research studies have been proposed to reduce the number of decoding iteration. Several hard-decision-aided (HDA) criterion is proposed by [6] [7] to early terminate the decoding process if two consecutive iteration decoding results are the same. However, HDA criterion doesn’t guarantee the correctness of decoding result while incorrect termination is forbidden for NVMs. To solve incorrect termination problem, [8] and [9] proposed several criterions that terminate decoding processes without degrading bit error rate. However, the early termination schemes proposed by [8] [9] are unable to terminate decoding operations within iteration.

In this paper, we proposed an early termination scheme with cost-effective hardware design and implemented a high-throughput layered min-sum QC-LDPC decoder for NVMs. The proposed early termination scheme is able to terminate decoding operations within iteration with low hardware overhead.

The rest of the paper is organized as follows. Section II introduces the layered min-sum decoding algorithm and block-serial decoding schedule. Section III presents the proposed a column-based early termination scheme and cost-effective hardware design. Section IV presents the simulation and implementation results. At last, Section V concludes this paper.

II. BLOCK-serial LAYERED MIN-SUM DECODING

A. Layered Min-Sum Decoding Algorithm

The iterative decoding algorithm of LDPC codes are based on belief propagation, in which messages form bit nodes and check nodes are updated and passed along the edges iteratively. Layered decoding regards the parity check matrix (PCM) H as L layers. Decoding process executes the updating operations in a layer-by-layer fashion. Hence, an iteration decoding is composed of several sub-iterations in layer decoding. In each sub-iterations, the check node updating and the bit node updating are executed locally within the corresponding row. The decoding flow of the layer decoding algorithm is shown in Fig. 1.

Fig. 1. Decoding flow of the layered min-sum decoding algorithm.
The iterative updating process of k-th layer of our work is formulated below:

$$\rho^{(k)} = r^{(k-1)}(I_j) - \Lambda_j^{(k-1)}$$  \hspace{1cm} (1)

$$\Lambda_j = 0.5 \cdot \{\text{XOR}[\text{sign}(\rho^{(l)})] \cdot \min[1, \rho^{(l)}]\}$$  \hspace{1cm} (2)

$$r^{(k)}(I_j) = \rho^{(k)} + \Lambda_j$$  \hspace{1cm} (3)

Where $\rho = [\rho_1, ..., \rho_N]$ represent prior messages, $\gamma(I_i)$ denotes posterior messages for row i and $A_j$ denotes the vector of extrinsic messages where $j = 1, ..., r$.

The Layered min-sum decoding allows updating both check nodes and variable nodes concurrently. Moreover, the layer-by-layer updating leads to significant memory reduction and the refined posterior messages form earlier rows help the decoding convergence twice as fast as the standard two-phase message passing (TPMP).

B. Block-Serial Decoding Schedule

The parity check matrix H of QC-LDPC codes are an $M \times N$ array of circulants with $M = m/b$ and $N = n/b$. Each sub-matrix $A_{ij}$ is a $b \times b$ square matrix, where each row is the cyclical right-shift of previous row, and the first row is cyclically right shifted form the last row. Decoding QC-LDPC codes with layered min-sum decoding algorithm, the b rows of non-zero entries in each sub-matric of circulant are non-overlapping. Since the non-zero entries in a row correspond to the bit nodes to be updated, the non-overlapping rows of non-zero entries represent that there is no data confliction among the bit node updating in a row of circulants. Therefore, the b rows of circulant in a layer can be parallel processed in each sub-iteration.

Fig. 2 illustrates the block-serial decoding schedule of layered QC-LDPC decoder. The decoder processes one circulant sub-matrix at a time. A sub-iteration is finished when $N$ circulants in a layer are all processed. Besides, it takes $M$ sub-iterations to finish single iteration. The architecture of block-serial layered min-sum decoder is shown in Fig. 3. The permuter cyclically right shift $b$ posterior messages according to $d_i$ shift value. The b-parallel check node updates (CNUs) and b-parallel bit node updates (BNUs) execute updating operations in parallel.

III. PROPOSED EARLY TERMINATION SCHEME

The redundant calculation of decoder leads to decreased throughput and increased power consumption. Hence, LDPC decoder with efficient early termination scheme is necessary for NVMs.

A. Proposed Column-Based Early Termination

Several studies have been proposed to early terminate the decoding process for LDPC codes. However, the correctness of LDPC results should be guaranteed for NVMs. The parity check equation (PCE) is an ideal stopping criterion for decodable codeword because it reduces latency without incorrect termination. If a codeword has been corrected, the hard-decision bits of the valid codeword satisfy the parity check equations, which can be formulated as $Hv^T = 0$.

Layered decoding updates posterior messages every layer, which infers that there are updated $v^T$ every sub-iteration. In order to avoid throughput degradation, we proposed a column-based early termination (CB-ET) scheme which is able to terminate decoding process with extra sub-iteration checking. Fig. 4 shows the checking schedule of proposed CB-ET scheme. The parity check of the proposed column-based early termination is performed in a column-by-column manner. The parity check can be regarded as checking whether $H_{ij}v^T = 0$ or not every cycle, where $i = 1, ..., N$. Furthermore, the column-based checking process is working concurrently with block-serial decoding process. Therefore, it takes only single extra sub-iteration to do the PCE checking, which make it possible to early terminate within iteration. The timing diagram of proposed CB-ET scheme in the block-serial decoder is shown in Fig. 5.
the value of XOR-register pair in the decoding process.

The details of the proposed CB-ET can be summarized in the following pseudo code.

```plaintext
// Initialization
check_all_zero ← 1;
δ1,...,δM ← 0;
C1,...,CM ← 0;
// Early termination checking
for i form 0 to M do
  for j form 0 to N do
    // Input: updated posterior message vector γ = [γ0, γ1,...,γN]
    // Output: check_all_zero
    // Initialization
    check_all_zero ← 1;
    δ1,...,δM ← 0;
    C1,...,CM ← 0;
    // Early termination checking
    for i form 0 to max_iteration-1 do
      for j form 0 to M do
        // XOR operation between the accumulated register and permuted data
        δi ← permuter(d1, sign(γ));
        δM ← permuter(dM, sign(γ));
        C1 ← XOR(C1, δi);
        CM ← XOR(CM, δM);
        end for
    if (check_all_zero == 0) then
      early_terminated_decoding();
    end for
end for
```

### Column-Based Early Termination Scheme Algorithm

A cost-effective CB-ET design is proposed to implement parity check equation \( H v_i = 0 \), which is matrix multiplication. Fig. 6 shows a simple example of checking process for QC-LDPC decoder. Because the elements of parity check matrix and hard-decision of posterior messages are all binary, we can use XOR operation to implement the matrix multiplication. Furthermore, due to the structure of QC-LDPC, it is convinced that there is only one input for each block-serial decoding. Therefore, we can simply use a XOR-register pair to accumulate parity check equation.

The block diagram of layered decoder with proposed CB-ET scheme is shown in Fig. 7. Every \( b \) sign bits of posterior messages from BNU processor are forwarded to permuters. The permuter cyclic shifts \( b \) sign bits according to the difference of shift value between decoding block and checking block. Then, the accumulators do the XOR operation between accumulated register and permuted data. At last, OR-tree checks if the content of the registers are all-zero after all the block columns are processed. If the output of the OR-trees is zero, the parity check is satisfied and the decoding procedure can be early terminated. The details of the proposed CB-ET can be summarized in the following pseudo code. \( δ_i \) denotes the permuted sign bits vector of \( i \)th layer with the cyclic shift value \( d_{shift} \); \( C_i \) denotes the value of XOR-register pair in \( i \)th layer. The result of OR-tree is regards as check_all_zero. The proposed CB-ET scheme checks parity equation after each layer is processed.

### IV. SIMULATION AND IMPLEMENTATION RESULT

#### A. Simulation Results

To verify the proposed column-based early termination scheme, we follow [10] to construct a QC-LDPC code with \((M, N, b) = (6, 58, 160)\) for NVMs. The maximum iteration is set to 8, and each iteration consists of 6 sub-iterations. Layered minimum decoder with proposed column-based ET scheme is implemented with Xilinx ML605 for emulation. Each point of Fig. 8 collects 100 uncorrected codewords. The figure shows that there is no occurrence of incorrect termination of decoding during the emulation. Therefore, the proposed CB-ET scheme early terminates decoding process without degrading bit error rate, which is called coded bit error rate. Fig. 9 shows the average sub-iteration numbers of proposed CB-ET scheme compared with LC-ICA which is proposed by [9]. The quantity of emulation data is \( 10^2 \) codewords for each scale of raw bit
error rate in NVMs. Blue line and green line represent results of LC-ICA and our work, respectively. Our proposed CB-ET performs fewer sub-iteration number than LC-ICA and the improvement is more significant under small raw bit error rate. The comparison of throughput is presented in TABLE I with several chosen raw bit error rate points. It shows that the throughput of proposed scheme is 37.7% higher than LC-ICA when raw bit error rate of flash memory is 3x10^{−3}. Moreover, the positive relation between throughput improvement and low raw bit error rate of flash memory is beneficial to NVMs.

\[ \text{TABLE I. THROUGHPUT ANALYSIS} \]

<table>
<thead>
<tr>
<th>Raw Bit Error Rate</th>
<th>LC-ICA [9] (Mb/s)</th>
<th>Proposed CB-ET (Mb/s)</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>6x10^{−3}</td>
<td>551</td>
<td>631</td>
<td>1.4%</td>
</tr>
<tr>
<td>5x10^{−3}</td>
<td>751</td>
<td>917</td>
<td>21.8%</td>
</tr>
<tr>
<td>4x10^{−3}</td>
<td>937</td>
<td>1211</td>
<td>29.5%</td>
</tr>
<tr>
<td>3x10^{−3}</td>
<td>1123</td>
<td>1549</td>
<td>37.7%</td>
</tr>
</tbody>
</table>

The QC-LDPC decoder with proposed early termination scheme is synthesized by TSMC 90nm 1P9M process. The gate count overhead is shown in TABLE II. The proposed ET checking circuit achieves 2.2% area overhead. However, the proposed early termination scheme achieves great throughput improvement with low area overhead.

\[ \text{TABLE II. SYNTHESIS RESULT OF LAYERED MIN-SUM DECODER} \]

<table>
<thead>
<tr>
<th>No ET Scheme</th>
<th>Proposed CB-ET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>TSMC 90nm</td>
</tr>
<tr>
<td>Frequency</td>
<td>200 MHz</td>
</tr>
<tr>
<td>Gate Count</td>
<td>730 K</td>
</tr>
<tr>
<td>Area Overhead</td>
<td>2.20%</td>
</tr>
</tbody>
</table>

V. CONCLUSION

A high-throughput layered min-sum QC-LDPC decoder with cost-effective CB-ET for NVMs has been presented. The proposed CB-ET scheme is able to early terminate decoding process within iteration, which increases the throughput tremendously. The throughput improvement is 37.7% compared to the state-of-the-art early termination scheme when raw bit error rate of flash memory is 3x10^{−3}. The decoder is implemented in TSMC 90nm process with operation frequency of 200MHz. The implemented result shows the checking circuit only achieves 2.2% area overhead.

REFERENCES


