Low Cost and Power LDPC for Commodity NAND Products

Eran Sharon, Idan Alrod, Ran Zamir, Omer Fainzilber, Ishai Ilani, Alex Bazarsky, Idan Goldenberg - Western Digital®

A. Abstract

Commodity NAND products such as USB drives, memory cards and IOT devices are highly cost and power sensitive. As such, these devices will usually utilize the densest memory available (e.g. 4 bit-per-cell memory based on the latest, most aggressively scaled technology node), which requires advanced ECC and DSP solutions in order to enable reliable storage over the "noisy" media. In order to minimize the ECC overprovisioning (and hence memory cost) needed for meeting the product reliability specs, state of the art Low Density Parity Check (LDPC) codes are used in such systems. These codes provide near Shannon limit performance with significantly higher correction capability compared to BCH codes, at the expense of larger silicon area and higher power consumption.
In this work, we describe a low cost and low power LDPC solution, based on a special subcode based structure, coupled with a multi-gear decoding algorithm. It enables >4X silicon area reduction with sub-BCH power consumption.

B. Background – LDPC and iterative decoding

LDPC codes can be defined using a sparse bipartite graph G, where the left nodes of the graph represent the codeword bits and the right nodes of the graph represent parity check constraints that the bits should satisfy in order to constitute a valid codeword (Fig. 1).

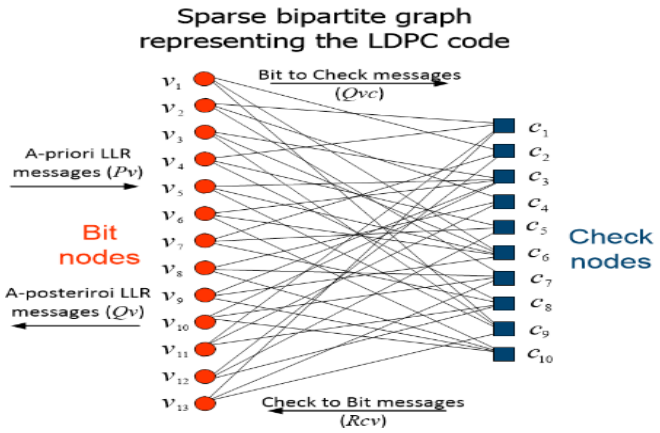

Fig. 1

They can be decoded iteratively by exchanging messages between the bit nodes and the check nodes, with performance that approaches the Shannon limit, when code is properly designed and when very large code length is used. However, the large code length incurs both a complexity and latency penalties. The iterative decoder complexity is proportional to the number of messages that need to be computed and stored, corresponding to the number of edges in the code's graph that grows linearly with the code length.

C. Subcodes based LDPC

As opposed to communication channels, in which an ECC solution is usually designed to withstand a fixed SNR condition inducing a narrow Bit Error Rate (BER) distribution, while meeting strict decoding throughput and latency, the situation is different for NAND memories, from two aspects. First, the memory channel is highly variable (process variations, temperature variations, wear along memory lifetime, disturb effects,…etc) [1]. This leads to a wide BER distribution, usually modeled as a log-normal distribution, with a very low median BER and a long tail that may reach high BER values, as shown in Fig.2.

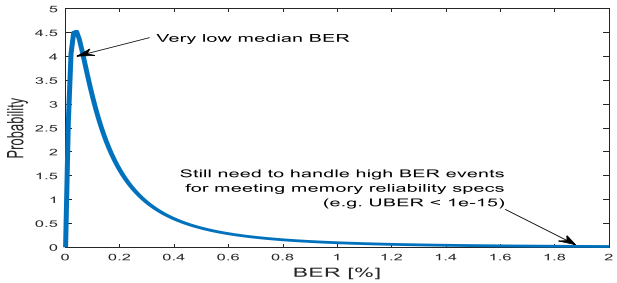

Fig. 2

Second, a memory system needs to meet an average decoding throughput, but can tolerate rare events of much higher decoding latency by stalling following read operations, especially in low end commodity products, which are not Quality-of-Service (QoS) sensitive.
In order to avoid over engineering and provide the lowest cost and power solution, an ECC system for a NAND memory based commodity product should be designed accordingly. Average decoding throughput for supporting the required read throughput should be met assuming that the vast majority of read pages exhibit very low BER. However, ECC should still handle high BER events in order to meet the memory reliability specs (e.g. UBER < 1E-15), while potentially compromising decoding latency in these rare events.
A tailored subcodes based LDPC solution can be designed exactly for these settings [3-8]. The subcodes are connected through a layer of joint parity bits, as shown in Fig. 3. This special structure allows each sub-codes to be individually decoded as a short LDPC code, via message passing over its sub-graph, which is sufficient in most cases, where BER is low. In addition, for rare cases where the BER of a read page is high (e.g. outliers of the process, extreme operational conditions, such as low/high temperatures, disturb effects,…etc), information can be exchanged between the sub-codes through the layer of joint parity bits. Such message passing decoding over the full graph provides significant boost in correction capability and hence a "safety net" for the high BER events, as it corresponds to decoding of a very long LDPC code.

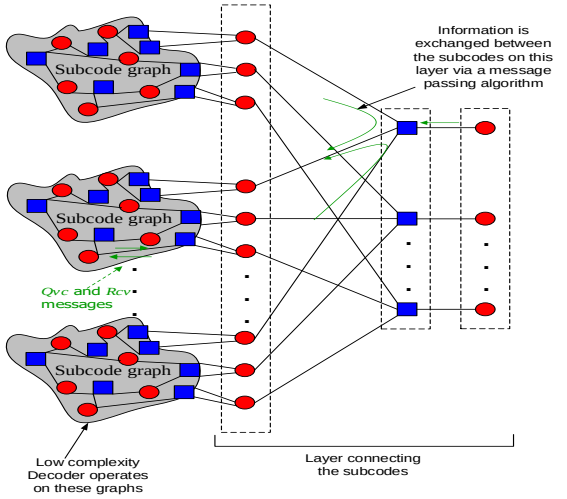

Fig. 3

The proposed subcode based structure has several benefits: 1. Near Shannon limit performance of a very long LDPC code, enabling minimal ECC overprovisioning for achieving the memory reliability specs. Thus, minimizing memory cost. 2. Low complexity LDPC implementation, scaled according to the short subcode length. The decoder memory is dimensioned according to the number of subocde's graph edges (for storing the messages exchanged during subcode decoding) and the small number of edges in the sparse connecting layer (for storing the messages exchanged between subcodes). 3. Low latency decoding in case of a random read scenario where only a single subcode information is requested and the BER is sufficiently low to succeed subcode decoding without joint parity usage ([2]).

The downside of the proposed structure is higher decoding latency in rare cases where individual subcode decoding fails and joint parity is used for information exchange between subcodes. This is due to reusing the same memory for decoding the different subcodes. Hence, each time we revisit a subcode, we start its decoding from scratch, though with improved extrinsic information conveyed through the connecting layer from its fellow subcodes. However, such rare performance hiccups are acceptable for commodity NAND products, which are not QoS sensitive.

D. Multi-gear Decoding

Given that during most of the memory lifetime the observed BER is very low, the cost and power of the LDPC engine can be further optimized by employing multiple decoding gears with different power, throughput and correction capability profiles. As an example, a 4x1KB subcodes based LDPC with the following two gears may be used:

- Bit-Flipping (BF) hard decoder for 1KB subcode (1 bit per message)
- Belief Propagation (BP) soft decoder for both 1KB subcode and 4KB full code modes (may use 6 bits per message)

The LDPC engine may estimate the read page BER as part of its initialization process by counting the number of unsatisfied parity checks and based on it automatically choose the appropriate decoding gear.

In addition, the parallelism of each decoding gear may be dimensioned according to its usage probability (given the memory BER distribution) and the overall required decoding throughput. Thus, the number of costly high resolution processing units instantiated for the BP decoder, which is rarely used ("safety net"), may be much lower than the number of simple BF processing units. This significantly reduces the ASIC footprint with negligible impact on decoding throughput.

Fig. 4 compares the correction capability of 1KB and 4KB conventional LDPC codes to a 4x1KB subcodes based LDPC code, under different decoding modes.

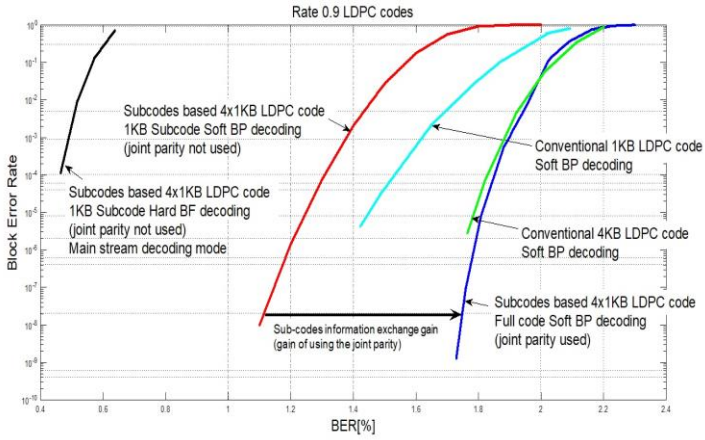

Fig. 4

E. Summary

The proposed subcodes based LDPC with multi-gear decoding provides a very efficient solution for commodity NAND products which are not sensitive to QoS. It allows for a significant reduction of >4X in the ASIC area compared to a conventional single gear LDPC solution of the same length. This is enabled by dimensioning the memory sizes according to the subcode size (e.g. 1KB vs 4KB) and instantiating significantly less costly BP processing units, while relying on low cost BF processing units to deliver the required throughput.

In addition, the solution provides a sub-BCH power profile, due to usage of a simple BF decoder as the mainstream mode.

References:

[1] V.Taranalli, H.Uchikawa, P.H.Siegel, "Channel Models for Multi-Level Cell Flash Memories based on Empirical Error Analysis," NVMW 2016
[2] E.Ram,Y.Casuto,"LDPC Codes with Local and Global Decoding", ISIT June, 2018
[3] E.Sharon, S.Litsyn, I.Alrod, "Memory-efficient LDPC decoder and method", US patent 8,291,279, May. 2008
[4] S. Litsyn, E.Sharon, I.Alrod, M.Lasser, "Method and Device for Multi Phase Error Correction", US patent 7,844,877, Sept. 2006
[5] E.Sharon, S.Litsyn, I.Alrod, O. Fainzilber, "Error-correction decoding with reduced memory and power requirements", US patent 9,270,297, July 2011
[6] O.Fainzilber, E.Sharon, I. Ilani, A. Bazarsky, "Multi-Stage decoder", US patent 9,614,547, June 2014
[7] I.Alrod, E.Sharon, S.Litsyn, "Error correction decoding by trial and error", US patent 8,042,029, Sept. 2006
[8] E.Sharon,I.Goldenberg,I.Ilani,I.Alrod,Y.Ryabinin,Y.Dumchin,M.Fiterman,R.Zamir, "ECC decoder with multiple decoding modes", US patent 10,218,384, Dec. 2016