Reduced-Complexity Offset Min-Sum Based Layered Decoding for 5G LDPC Codes

Vladimir L. Petrović, *Graduate Student Member, IEEE*, and Dragomir M. El Mezeni

Abstract — **This paper presents a novel approach for the reduced-complexity Min-Sum (MS) decoding of low density parity check (LDPC) codes in the partially parallel layered decoder architecture, which contains large number of serial check node processors. Reduced complexity is obtained by using the variant of the single-minimum Offset Min-Sum (smOMS) algorithm that approximates a second minimum with the addition of the variable weight parameter to the minimum value. Although the reduced-complexity MS algorithms primarily reduce hardware resources in fully parallel implementations, the results showed that a considerable reduction can be obtained if serial check node processors are used. Additionally, the paper proposes a better subminimum estimation for irregular codes from 5G new radio (5G NR). The method uses smaller subminimum estimation weights in check nodes with higher degree and higher weights in check nodes with smaller degree, which lead to the significant improvement in the SNR performance.**

Keywords — **5G new radio, layered decoding, low density parity check (LDPC) codes, min-sum (MS) algorithm.**

I. Introduction

Fifth generation (5G) New Radio (NR) uses low density parity check (LDPC) codes [1] for channel coding on the data channel [2]. These codes belong to the family of quasi cyclic (QC) LDPC codes whose parity check matrix (PCM) is consisted of circularly shifted identity sub matrices [3]. Such structure is convenient for achieving high parallelism in the decoder implementations, which is the key prerequisite for obtaining high throughput required by modern communications. 5G NR codes are constructed based on the base graph matrix, which determines the positions of the circularly shifted identity sub matrices [4]. Additionally, the code is defined by the size of the identity sub matrix (lifting size), and by shift values for each sub matrix in the code's PCM.

LDPC codes are usually decoded using the message passing algorithm [5], which provides good capacity approaching performance [6]. Multiple approximations to the original message passing algorithm are frequently applied in implementations of LDPC decoders in order to reduce the complexity of calculation. Some of the most frequent approaches are the min-sum (MS) algorithm [7] and, its variants, Offset MS (OMS) and Normalized MS (NMS) [8]. In the message passing algorithm, nodes in the LDPC codes' Tanner graph communicate using messages that are iteratively exchanged between variable and check nodes until the decoding is finished or until the maximum number of iterations is reached. In the MS algorithms check nodes calculate two minimum magnitude variable-to-check messages and return check-to-variable messages based on these two values.

Although MS algorithms give high complexity reduction comparing with the original message passing algorithm, a further reduction can be made if only one minimum is calculated and if the second is only approximated [9]–[13]. This can lead to significant hardware resource savings, especially when check node processes all messages in parallel [10]–[13]. A single minimum MS (smMS) is presented in [9], where the second minimum is estimated by an addition of a constant weight parameter to the minimum value. This is the simplest approach, but suffers from high error floors. A variable weight smMS (vwsmMS) is presented in [10], where the weight parameter is increased during the decoding, since the actual difference between the subminimum and the minimum in the MS algorithm increases in each new iteration of the decoding. Such approach gives significantly better SNR performance. If all messages are simultaneously available at the check node, even better results can be achieved by approximate finding of the subminimum value in the minimum finder tree [10]–[13].

Although fully parallel architectures give the highest throughput, the hardware complexity due to routing congestion can be high [14]. This paper presents a study on the reduced-complexity decoding for 5G NR in the partially parallel architectures. Such architectures have check node units (CNUs) that do not receive all messages in parallel, but serially [15], [16]. However, due to the high lifting size of the matrix, there are many of these serial CNUs. The chosen architecture performs layered decoding schedule which updates the variable nodes more frequently than the original message passing algorithm and provides faster convergence [17].

The paper presents a reduced-complexity OMS decoding, which is based on the vwsmMS, but is improved by different choice of weights specific for irregular 5G NR LDPC codes. Additionally, the paper presents the architecture of the serial check node unit that implements the improved reduced-complexity algorithm.

This work was developed during research funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

Vladimir L. Petrović and Dragomir M. El Mezeni are with the University of Belgrade – School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: petrovicv@etf.bg.ac.rs; elmezeni@etf.bg.ac.rs).

Digital Object Identifier: 10.1109/TELFOR51502.2020.9306590; Final, published article available at: https://ieeexplore.ieee.org/document/9306590

II. Decoding algorithms in layered LDPC decoders

A. Offset Min-Sum layered decoding

The iterative layered OMS decoding consists of the initialization, check node updates and variable node updates. Firstly, the variable nodes v are initialized by the input a priori log-likelihood ratio values calculated as

$$LLR_v = \log \frac{P(b_v = 0 \mid y_v)}{P(b_v = 1 \mid y_v)} = \frac{2y_v}{\sigma^2}, \quad (1)$$

where y_v are channel outputs, σ^2 is channel noise variance and $P(b_v = b \mid y_v)$ is the conditional probability that the bit b_v is equal to b, given that the y_v is received. After initialization variable nodes send messages to check nodes which in first iteration equal to the input LLRs ($M_{v2c} = LLR_v$). In layered schedule, not all messages are sent at the same time. The PCM is divided in layers and messages are sent firstly to the check nodes from the first layer (in QC-LDPC all check nodes that correspond to the first row in the base graph matrix). Check nodes calculate their response and send messages to variable nodes as

$$M_{c2v} = \max\left(\min_{v' \in V_c \setminus v}\left(\left|M_{v'2c}\right|\right) - \beta, 0\right) \times \prod_{v' \in V_c \setminus v} \operatorname{sgn}\left(M_{v'2c}\right), \quad (2)$$

where V_c is the set of all variable nodes connected to the check node c, and β is an offset parameter. Based on newly received check-to-variable messages and already sent variable-to-check message, variable nodes update LLR values as

$$LLR_v = M_{v2c} + M_{c2v}. \quad (3)$$

In the next sub iteration, a new layer is processed; variable nodes send new messages calculated as

$$M_{v2c} = LLR_v - M_{c2v}. \quad (4)$$

Check-to-variable messages in (4) are zero in first iteration, but in every other iteration they are check-to-variable messages from the previous iteration.

B. Reduced complexity offset Min-Sum

As can be seen from (2) the CNU should find two variable-to-check messages of least magnitude, select one of them, modify it with the offset parameter and pass it with the sign determined by the sign product as a new check-to-variable message. Reduced complexity can be achieved if only one minimum is calculated and the other estimated as

$$M_{c2v} = \max\left(M - \beta, 0\right) \times \prod_{v' \in V_c \setminus v} \operatorname{sgn}\left(M_{v'2c}\right), \text{ where}$$
$$M = \begin{cases} \min_{v' \in V_c}\left(\left|M_{v'2c}\right|\right) + w, & \text{if } \min_{v' \in V_c}\left(\left|M_{v'2c}\right|\right) = \left|M_{v'2c}\right| \text{ and } N_{min} = 1 \\ \min_{v' \in V_c}\left(\left|M_{v'2c}\right|\right), & \text{otherwise} \end{cases}, \quad (5)$$

where w is the weight factor for subminimum estimation and N_{min} is the number of message magnitudes that are equal to the minimum message magnitude [9]. If w is fixed during the decoding, the complexity is mostly reduced, but significantly better SNR performance is obtained if w is increased in each iteration or at least after every few iterations [10]. A method with fixed w is called single minimum OMS (smOMS), whereas the method with the variable w is called variable weight smOMS (vwsmOMS). These methods are the only that can lead to the reduced complexity for serial CNU since all other methods ([11]–[13]) calculate the subminimum in such way which gives reduction in complexity of the parallel minimum finder tree, but cannot reduce complexity of the serial CNU.

III. Improvement of the reduced-complexity OMS decoding for 5G NR

LDPC codes from 5G NR are highly irregular [4], which means that variable node degrees (d_v) and check node degrees (d_c) are highly different for different nodes. The general structure of the base graph matrix for 5G NR LDPC codes is show in Fig. 1. As can be seen, check node degrees vary from $d_c = 3$ to $d_c = 19$. Therefore, it can be expected that subminimum and minimum differences in the MS algorithm would be larger for smaller degree nodes, since the number of values that enter the minimum calculation is smaller. This means that the weight parameter w in the vwsmMS algorithm should be changed differently for check nodes with different check node degree. In order to confirm this assumption a fixed-point implementation of the OMS decoder for (16128, 8448) code from 5G NR is simulated for multiple code words in additive white Gaussian noise channel (AWGN). The average difference between subminimum and minimum for check nodes with different degrees is shown in Fig. 2. It is clear that higher degree check nodes have smaller differences between the subminimum and minimum values than the lower degree check nodes.

Interestingly, it is noticeable that high-degree check nodes' differences start decreasing near the end of the decoding. This is explained by the saturation of messages and LLRs in the fixed point implementation, because many messages have almost the same magnitude.

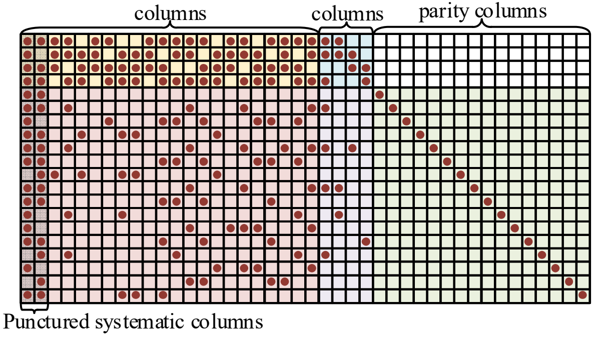

Fig. 1. The structure of the base graph 1 matrix of 5G NR LDPC codes. Dots represent the presence of the cyclically shifted identity matrix the PCM.

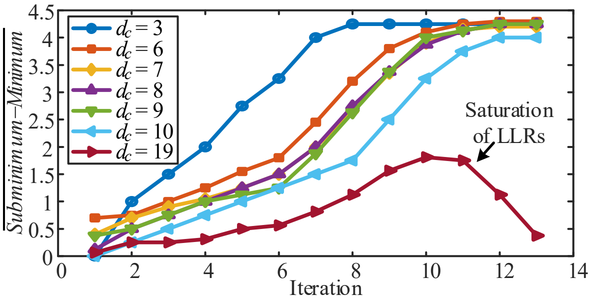

Fig. 2. The average difference of subminimum and minimum depending on the iteration number in fixed-point OMS layered decoding for 5G NR (16128, 8448) code.

Since the weight parameter w in the vwsmOMS algorithm is already changing during the decoding, which is done in hardware by external setting of the parameter to all CNUs by the control, no additional complexity would be induced if w is changed separately for each layer too. Therefore, in this paper a vwsmOMS is improved by setting the w parameter to smaller value for higher degree check nodes. This is further called enhanced vwsmOMS (EvwsmOMS).

IV. Reduced-complexity OMS check node unit architecture

The conventional offset Min-Sum serial check node unit architecture is shown in Fig. 3. The CNU is consisted of two parts separated by pipeline registers. The input part calculates the minimum, subminimum, index of minimum, and sign product of all variable-to-check messages' signs. When minimums are found, the aforementioned values are written to pipeline registers, and input part starts doing calculations for the next layer. The output part selects one of the minimum or subminimum values, based on the minimum index value (if the check node sends the message to the variable node whose variable-to-check message was of minimum magnitude, than the subminimum is used). The selected value is then reduced by the offset parameter and saturated if the resulting value is less than zero. The sign of the check-to-variable message is determined based on the calculated sign product and the sign of the variable-to-check message from the variable node to which the check-to-variable message is sent, as in (2).

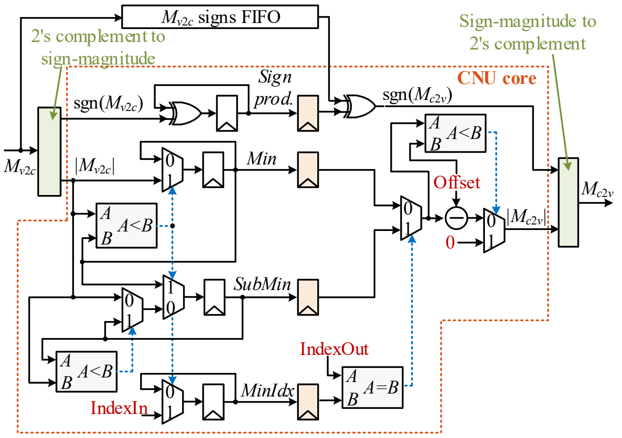

Fig. 3. Offset Min-Sum serial check node unit architecture

In the reduced complexity EvwsmOMS implementation, there is no need for calculation of the subminimum value. The proposed CNU architecture is shown in Fig. 4. The input part calculates the minimum, index of the minimum and a sign product. Besides that, it needs to provide information whether the number of messages that have the same magnitude as the minimum is only 1 or more than 1, as required in (5). This is done using the equality output of the comparator. The *Eq.* signal is set to "1" whenever the current minimum and the new message magnitude are equal and reset whenever a new minimum is found. This way, the *Eq.* signal always tells if there is another variable-to-check message that is equal to the minimum. The comparator with the "less than" and "equal" outputs uses significantly less resources than two comparators necessary for both the minimum and subminimum calculation. Additionally, the number of flip flops is also reduced since the subminimum value is not kept.

Conventionally, the output part would calculate the estimated subminimum, and then use the minimum value and estimated subminimum in the same way as in the OMS architecture. This would require two addition operations and possibly, unacceptable increase of hardware resources. The proposed architecture avoids this using the following procedure. The w parameter for subminimum estimation is already reduced by the offset parameter β, in order to avoid both the addition and the subtraction in cases when a subminimum should be passed. Instead of subtraction with the offset parameter, the output part performs a signed addition of the minimum and a negative offset ($-\beta$) or the already reduced weight parameter ($w-\beta$). If the resulting value is negative, the output should be saturated: to zero if the minimum is passed or to maximum integer value if the subminimum is passed, because in the second case, the negative result means that the overflow has happened. If w is smaller than β, the negative values are saturated to zero.

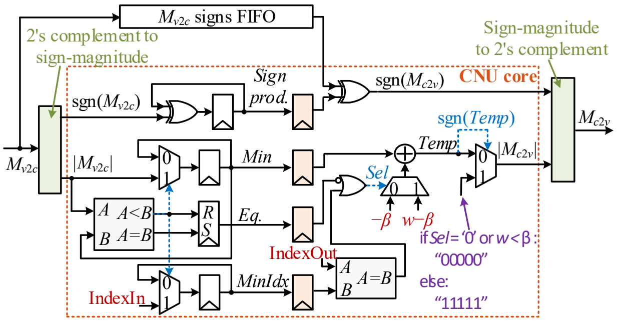

Fig. 4. EvwsmOMS serial check node unit architecture

V. Results

A. SNR performance

Monte Carlo simulation of the fixed point implementation of the multiple algorithms is performed for (16128, 8448) code from 5G NR, AWGN channel and QPSK modulation, with 6 bits used for messages and 8 for LLRs. Maximum iteration number was set to 20. The simulated algorithms were MS, OMS, smOMS, vwsmOMS and enhanced vwsmOMS (EvwsmOMS). In the reduced-complexity algorithms, weight parameters were changed as summarized in Table 1. In the smOMS, *w* was set to its best constant value. For vwsmOMS, *w* was changing linearly with the iteration number (*it*) starting from 0. The linear function from Table 1 gave the best results. Finally, the EvwsmOMS used various linear functions for nodes with different check node degree.

Fig. 5 shows the simulated block error rate (BLER) curves. The EvwsmOMS performance shows a significant improvement with respect to the vwsmOMS, although the loss compared with the original OMS is still considerable.

TABLE 1: WEIGHT PARAMETERS USED IN SINGLE-MINIMUM OMS ALGORITHMS.

Algorithm	d_c	*w*
smOMS	all	2.25
vwsmOMS	all	0.25 + 0.25*it*
(E)vwsmOMS	3	0.5 + 0.4*it*
	6, 7, 8, 9, 10	0.25 + 0.27*it*
	19	0.25*it*

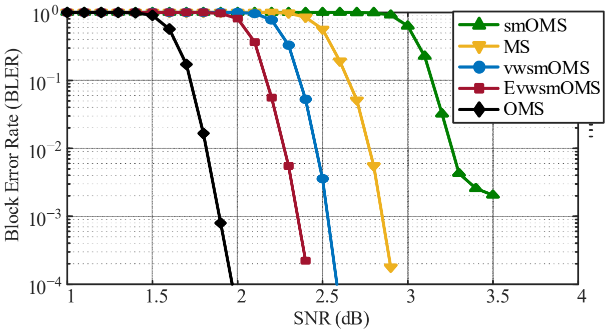

Fig. 5. SNR performance of various MS algorithms for rate 22/42 base graph 1 code from 5G NR (16128, 8448).

B. *Hardware complexity*

The proposed CNU core from section IV for the EvwsmOMS algorithm, CNU core for OMS algorithm and CNU core for MS algorithm are implemented on the Zynq UltraScale+ RF-SoC device (XCZU28DR). The implementation results for 384 (maximum lifting size in 5G NR) CNU cores are shown in Table 2. Resource savings achieved in the EvwsmOMS implementation compared with the OMS are about 35% for complex logic blocks (CLBs), 11% for look-up tables (LUTs), and 25% for flip flops (FFs). Therefore, even though the serial CNU architecture is used, a considerable reduction in complexity can be achieved.

TABLE 2: COMPARISON OF FPGA RESOURCES OF 384 CHECK NODE UNIT CORES FOR VARIOUS MS ALGORITHMS.

Algorithm	*CLBs*	*LUTs*	*FFs*
MS	1547	5379	12288
OMS	1595	6934	12288
(E)vwsmOMS	1032	6146	9216

VI. CONCLUSION

This paper showed that reduced-complexity OMS decoding can be achieved in partially-parallel LDPC decoder architecture with considerable SNR performance loss. The proposed check node unit architecture provides significant resources savings compared with the conventional OMS approach. The paper also presents a better method for choosing the weight parameter used for estimation of the second minimum in the reduced-complexity OMS decoding. The weight parameter is chosen separately for check nodes with different check node weights since 5G NR codes are highly irregular. Using this method the SNR performance loss is partially reduced.

REFERENCES

[1] R. Gallager, “Low-density parity-check codes,” *IRE Trans. Inform. Theory*, vol. 8, no. 1, pp. 21–28, Jan. 1962.

[2] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; Multiplexing and channel coding (Release 16)*, 3GPP TS 38.212 V16.1.0 (2020-03), 2020.

[3] R. G. Gallager, “Analysis of Number of Independent Decoding Iterations,” in *Low-Density Parity-Check Codes*, Cambridge, MA, USA, MIT Press, 1963, pp. 81–88.

[4] T. J. Richardson and S. Kudekar, “Design of low-density parity check codes for 5G new radio,” *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 28–34, Mar. 2018.

[5] D. J. C. MacKay and R. M. Neal, “Near Shannon limit performance of low density parity check codes,” *Electron. Lett.*, vol. 33, no. 6, pp. 457–458, Mar. 1997.

[6] D.J.C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 399–431, Mar. 1999.

[7] M. P. C. Fossorier, M. Mihaljevic, and H. Imai, “Reduced complexity iterative decoding of low density parity check codes based on belief propagation,” *IEEE Trans. Commun.*, vol. 47, no. 5, pp. 673–680, May 1999.

[8] J. Chen and M. P. C. Fossorier, “Density evolution for two improved BP based decoding algorithms of LDPC codes,” *IEEE Commun. Lett.*, vol. 6, no. 5, pp. 208–210, May 2002.

[9] A. Darabiha, A. Carusone, and F. Kschischang, “A bit-serial approximate min-sum LDPC decoder and FPGA implementation,” in *Proc. IEEE Int. Symp. Circuits and Syst. (ISCAS)*, May 2006, pp. 149–152.

[10] F. Angarita, J. Valls, V. Almenar, and V. Torres, “Reduced-complexity min-sum algorithm for decoding LDPC codes with low error-floor,” *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 61, no. 7, pp. 2150–2158, Jul. 2014.

[11] I. Tsatsaragkos and V. Paliouras, “Approximate algorithms for identifying minima on min-sum LDPC decoders and their hardware implementation,” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 8, pp. 766–770, Aug. 2015.

[12] C.-C. Cheng, J.-D. Yang, H.-C. Lee, C.-H. Yang, and Y.-L. Ueng, “A fully parallel LDPC decoder architecture using probabilistic min-sum algorithm for high-throughput applications,” IEEE Trans. Circuits Syst. I: Reg. Papers, vol. 61, no. 9, pp. 2738–2746, Sep. 2014.

[13] J. M. Català-Pérez, J. O. Lacruz, F. García-Herrero, and J. Valls, "Second Minimum Approximation for Min-Sum Decoders Suitable for High-Rate LDPC Codes", *Circuits Syst. Signal Process.*, vol. 38, pp. 5068–5080, Apr. 2019.

[14] A. J. Blanksby and C. J. Howland, “A 690-mW 1-Gb/s 1024-b, rate-1/2 low-density parity-check code decoder,” *IEEE J. Solid-State Circuits*, vol. 37, pp. 404–412, Mar. 2002.

[15] C. Marchand, L. Conde-Canencia, and E. Boutillon, “Architecture and finite precision optimization for layered LDPC decoders,” *J. Signal Process. Syst.*, vol. 65, pp. 185–197, 2011.

[16] V. L. Petrović, M. M. Marković, D. M. El Mezeni, L. V. Saranovac, and A. Radošević, “Flexible High Throughput QC-LDPC Decoder with Perfect Pipeline Conflicts Resolution and Efficient Hardware Utilization,” *IEEE Trans. Circuits Syst. I: Reg. Papers*, to be published, doi: 10.1109/TCSI.2020.3018048

[17] D. E. Hocevar, “A reduced complexity decoder architecture via layered decoding of LDPC codes,” in *Proc. IEEE Work. Signal Process. Syst.*, Austin, TX, USA, Oct. 2004.