

High-Throughput and Energy-Efficient VLSI Architecture for Ordered Reliability Bits GRAND

Syed Mohsin Abbas, Thibaud Tonnellier, Furkan Ercan, *Member, IEEE*
 Marwan Jalaleddine and Warren J. Gross, *Senior Member, IEEE*

Abstract—Ultra-reliable low-latency communication (URLLC), a major 5G New-Radio use case, is the key enabler for applications with strict reliability and latency requirements. These applications necessitate the use of short-length and high-rate codes. Guessing Random Additive Noise Decoding (GRAND) is a recently proposed Maximum Likelihood (ML) decoding technique for these short-length and high-rate codes. Rather than decoding the received vector, GRAND tries to infer the noise that corrupted the transmitted codeword during transmission through the communication channel. As a result, GRAND can decode any code, structured or unstructured. GRAND has hard-input as well as soft-input variants. Among these variants, Ordered Reliability Bits GRAND (ORBGRAND) is a soft-input variant that outperforms hard-input GRAND and is suitable for parallel hardware implementation. This work reports the first hardware architecture for ORBGRAND, which achieves an average throughput of up to 42.5 Gbps for a code length of 128 at a target FER of 10^{-7} . Furthermore, the proposed hardware can be used to decode any code as long as the length and rate constraints are met. In comparison to the GRANDAB, a hard-input variant of GRAND, the proposed architecture enhances decoding performance by at least 2 dB. When compared to the state-of-the-art fast dynamic successive cancellation flip decoder (Fast-DSCF) using a 5G polar (128, 105) code, the proposed ORBGRAND VLSI implementation has $49\times$ higher average throughput, $32\times$ times more energy efficiency, and $5\times$ more area efficiency while maintaining similar decoding performance.

Index Terms—Area efficiency, Energy efficiency, Error Correcting Code (ECC), Guessing Random Additive Noise Decoding (GRAND), Maximum Likelihood Decoding (MLD), Ordered Reliability Bits GRAND (ORBGRAND), VLSI architecture

I. INTRODUCTION

FOLLOWING Shannon's seminal 1948 paper [1], much work was directed toward developing practical coding schemes that could approach channel capacity (named “the Shannon limit”). Early proposed channel coding techniques mainly focused on designing error correcting codes where the number of correctable errors is selected by design. This approach towards channel coding brought rise to Hamming codes [2] and Bose–Chaudhuri–Hocquenghem codes [3], [4]. Recently, researchers have been trying to discover new capacity achieving and capacity-approaching codes. Most notably, Turbo codes [5] and LDPC codes [6], were proposed over time as capacity-approaching codes. On the other hand, Polar

S. M. Abbas, T. Tonnellier, M. Jalaleddine and W. J. Gross are with the Department of Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada. F. Ercan is affiliated with Octasic Inc. Canada. (email: syed.abbas@mail.mcgill.ca, thibaud.tonnellier@mail.mcgill.ca, furkan.ercan@mail.mcgill.ca, marwan.jalaleddine@mail.mcgill.ca, warren.gross@mcgill.ca.). A part of this work has been presented in ICASSP 2021.

codes [7], proposed in 2008, are the first proven class of codes that asymptotically reach the Shannon limit for binary-input symmetric memory-less channels, as well as discrete and continuous memory-less channels [8]. However, designing codes that perform well in the short-to-medium block length regime is challenging [9].

The Guessing Random Additive Noise Decoding (GRAND) [10] is a recently proposed Maximum Likelihood (ML) decoding technique for short-length and high-rate codes. Furthermore, when used with random-codebooks, GRAND achieves capacity [10]. GRAND is highly useful for emerging applications requiring short-length and high-rate codes, such as ultra-reliable low-latency communication [11], augmented and virtual reality, intelligent transportation systems [12], the internet of things (IoT) [13], [14], machine-to-machine communication [15], and many more. GRAND is code-agnostic and noise-centric, which implies that, unlike other decoding techniques that leverage the structure of the underlying code to decode, it attempts to guess the noise that corrupted the codeword during transmission across the communication channel. As a result, GRAND can be used with any codebook, structured or unstructured. In comparison to other universal decoders, such as brute-force ML decoding and ordered statistic decoding (OSD) [16], GRAND provides a low complexity solution [17].

GRAND generates test error patterns, which are successively applied to the vector of channel observation values and the resulting vector is evaluated for codebook membership. GRAND can be used with any codebook as long as there is a method for validating a vector's codebook membership. For linear codebooks (\mathcal{C}), the codebook membership for a vector can be verified using the underlying code's parity check matrix \mathbf{H} . For other non-structured codebooks, stored in a dictionary, the codebook membership of a vector can be checked with a dictionary lookup. For the rest of the discussion, we restrict ourselves to (n, k) linear block codes, where n is the code length and k is the code dimension.

The order in which these test error patterns are generated is the primary distinction between GRAND variants. GRAND with ABandonment (GRANDAB) [10] is a hard decision input version of GRAND that produces test error patterns in increasing Hamming weight order up to weight AB . GRAND Markov Order (GRAND-MO) [18] is designed specifically for channels with memory which are susceptible to burst noise. Symbol Reliability GRAND [19] is another variant that distinguishes between reliable and unreliable bits of the received vector by using thresholding on the input channel observation values.

Ordered Reliability Bits GRAND (ORBGRAND) [20] and Soft GRAND (SGRAND) [21] are soft-input variants that efficiently utilize soft information (channel observation values), resulting in improved decoding performance over SRGRAND and GRANDAB. SGRAND outperforms other GRAND variants in terms of decoding performance by generating test error patterns in ML order [21],[22]. The generated test error patterns are interdependent and distinct for each received vector of channel observation values. As a consequence, SGRAND is unsuitable for parallel hardware implementation. ORBGRAND, on the other hand, is highly parallelizable and is highly suitable for a parallel hardware implementation.

In this paper, we aim to present a hardware architecture for a universal decoder based on the ORBGRAND algorithm. Our contributions are summarized as follows:

- We propose to modify ORBGRAND algorithm to reduce the worst case computational complexity (maximum number of codebook membership queries). ORBGRAND's complexity is directly proportional to the number of putative test error patterns. We investigate the parameters that affect the decoding performance as well as the computational complexity of the ORBGRAND decoding process, and we suggest changes to the ORBGRAND algorithm to aid hardware implementation and reduce complexity.
- ORBGRAND is primarily focused on generating integer partitions of a particular logistic weight (LW), from which test error patterns are generated. Based on an arrangement of shift registers and XOR gates, we propose a simple method for generating the integer partitions of a particular LW . The suggested technique makes the ORBGRAND algorithm easier to implement on hardware.
- Finally, we demonstrate how the simplified ORBGRAND algorithm can be implemented in hardware. All of the proposed simplifications are incorporated into the proposed hardware. We also provide further simplifications, such as partitioning the sorter to reduce the area overhead for the proposed ORBGRAND hardware, and we investigate the correlation of sorter partitioning on the ORBGRAND algorithm's decoding performance.

The VLSI implementation results show that, considering a code of length 128 and a target FER of 10^{-7} , the proposed hardware architecture can achieve an average throughput of 42.5 Gbps. Furthermore, the proposed hardware can be used to decode any code as long as the length and rate constraints are met. When compared to the state-of-the-art fast dynamic successive cancellation flip decoder (Fast-DSCF) using a 5G polar (128, 105) code, the proposed ORBGRAND implementation results in $49\times$ more average throughput, $32\times$ more energy efficiency, and $5\times$ more area efficiency while retaining similar decoding performance.

It should be noted that a part of this work was previously discussed in [23], where numerical simulation results for decoding Polar codes with the ORBGRAND algorithm and the ORBGRAND VLSI architecture were presented. This paper builds on earlier work [23] by presenting numerical simulation results for Bose-Chaudhuri-Hocquenghem (BCH)

codes, Cyclic Redundancy Check (CRC) codes, Random Linear Codes (RLCs) and Polar codes. The proposed ORBGRAND test error pattern generating scheme is explained with examples, graphics, and mathematical proof. Furthermore, the proposed ORBGRAND hardware is expanded to support additional parameters, and it is evaluated in terms of hardware efficiency, energy efficiency, and error decoding performance gain. A comprehensive analysis of worst-case latency and throughput for choosing different parameters for the proposed ORBGRAND hardware is also presented.

The remainder of this work is structured as follows: Preliminaries GRAND, ORBGRAND and VLSI architecture for GRAND are given in Section II. Section III investigates ORBGRAND parameters and proposes modifications for simple hardware implementation as well as complexity reduction of ORBGRAND. The proposed hardware architecture for ORBGRAND is detailed in Section IV, and the implementation results are compared to the Fast-DSCF decoder and GRANDAB decoder. Finally, in Section V, concluding remarks are made.

II. PRELIMINARIES

A. Notations

Matrices are denoted by a bold upper-case letter (M), while vectors are denoted with bold lower-case letters (v). The transpose operator is represented by \top . The number of k -combinations from a given set of n elements is noted by $\binom{n}{k}$. $\mathbf{1}_n$ is the indicator vector where all locations except the n^{th} are 0 and the the n^{th} is 1. All the indices start at 1. For this work, all operations are restricted to the Galois field with 2 elements, noted \mathbb{F}_2 . The symbols \therefore and \because denote *therefore* and *because* respectively.

B. GRAND decoding of linear block codes

A linear block code is a linear mapping $g : \mathbb{F}_2^k \rightarrow \mathbb{F}_2^n$, where $k < n$. In this mapping, a vector \mathbf{u} of size k maps to a vector \mathbf{c} of size n and the ratio $R \triangleq \frac{k}{n}$ is called the code-rate. For every linear block code, there exists a $k \times n$ matrix \mathbf{G} called generator matrix and a $(n - k) \times n$ matrix \mathbf{H} called parity check matrix. The set of the 2^k vectors \mathbf{c} is called a code \mathcal{C} , whose elements \mathbf{c} are called codewords and each codeword verifies the following property:

$$\forall \mathbf{c} \in \mathcal{C}, \mathbf{H} \cdot \mathbf{c}^\top = \mathbf{0}. \quad (1)$$

Consider the case where \mathbf{c} was sent via a noisy channel and \mathbf{r} was received at the channel's output. Because of the channel noise, \mathbf{r} and \mathbf{c} might differ. As a result, the relationship between \mathbf{r} and \mathbf{c} may be deduced as follows: $\mathbf{r} = \mathbf{c} \oplus \mathbf{e}$, where \mathbf{e} is the *error pattern* caused by the channel noise.

GRAND sequentially generates the test error patterns (e), and applies them to \mathbf{r} and checks for codebook membership of \mathbf{r} by verifying that

$$\mathbf{H} \cdot (\mathbf{r} \oplus \mathbf{e})^\top \quad (2)$$

is equal to zero. If so, \mathbf{e} is the guessed noise and $\hat{\mathbf{c}} \triangleq \mathbf{r} \oplus \mathbf{e}$ is the estimated codeword. GRAND's focus is on noise, thus it can be used with any codebook as long as there is a method

Algorithm 1: ORBGRAND Algorithm

```

Input:  $\mathbf{y}$ ,  $\mathbf{H}$ ,  $\mathbf{G}^{-1}$ ,  $LW_{\max}$ 
Output:  $\hat{\mathbf{u}}$ 
1  $ind \leftarrow \text{sortChannelObservationValues}(\mathbf{y})$ 
2  $e \leftarrow \mathbf{0}$ 
3 for  $i \leftarrow 0$  to  $LW_{\max}$  do
4    $S \leftarrow \text{generateAllIntPartitions}(i)$ 
5   forall  $j$  in  $S$  do
6      $e \leftarrow \text{generateErrorPattern}(j, ind)$ 
7     if  $\mathbf{H} \cdot (\hat{\mathbf{y}} \oplus e)^T == \mathbf{0}$  then
8        $\hat{\mathbf{u}} \leftarrow (\hat{\mathbf{y}} \oplus e) \cdot \mathbf{G}^{-1}$ 
9       return  $\hat{\mathbf{u}}$ 

```

for validating a vector's codebook membership. For linear codebooks, the codebook membership for a vector can be verified using \mathbf{H} .

C. ORBGRAND decoding

Algorithm 1 summarizes the steps of the ORBGRAND given the log-likelihood ratios (LLRs) for the communication channel. The inputs to the algorithm are the LLRs (\mathbf{y}) of size n , a $(n-k) \times n$ parity check matrix of the code \mathbf{H} , an $n \times k$ matrix \mathbf{G}^{-1} such that $\mathbf{G} \cdot \mathbf{G}^{-1}$ is the $k \times k$ identity matrix, with \mathbf{G} a generator matrix of the code, and the maximum logistic weight considered LW_{\max} ($LW_{\max} \leq \frac{n(n+1)}{2}$). The logistic weight (LW) corresponds to the sum of the indices of non zero elements in the test error patterns (e) [20].

ORBGRAND begins by sorting \mathbf{y} in ascending order of absolute value of LLRs ($|\mathbf{y}|$), and the corresponding indices are recorded in a permutation vector denoted by ind (line 1). Following that, all integer partitions are generated for each logistic weight (line 4). The test error patterns (e) are generated using the integer partition and ind (line 6). The generated test error patterns are then applied sequentially to the hard decision vector ($\hat{\mathbf{y}}$), which is obtained from the input soft channel observation values (\mathbf{y}). The resulting vector is then queried for codebook membership (line 7).

If the codebook membership criterion (2) is met, then e is the guessed noise and $\hat{\mathbf{c}} \triangleq \hat{\mathbf{y}} \oplus e$ is the estimated codeword. Otherwise, either larger logistic weights or the remaining error patterns for that logistic weight are considered. Finally, using \mathbf{G}^{-1} (line 8), the original message ($\hat{\mathbf{u}}$) is retrieved from the estimated codeword, and the decoding process is terminated.

The frame error rate (FER) performance of ORBGRAND, a soft decision decoder, is compared to a hard decision variant GRANDAB for cyclic redundancy check (CRC) codes [24] and Random Linear Codes (RLCs) [25], [26] in Fig. 1. CRC codes [24] are typically used to detect errors in communication systems and to assist list-based channel code decoders in selecting the final candidate codeword. On the other hand, CRC codes can also be used for error correction using the GRAND algorithm. The concept of using CRC codes for error correction with GRAND decoding was presented in [27] and expanded on in [28]. RLCs [25], [26] are linear block codes that are theoretically high-performing [25], [26] but are not

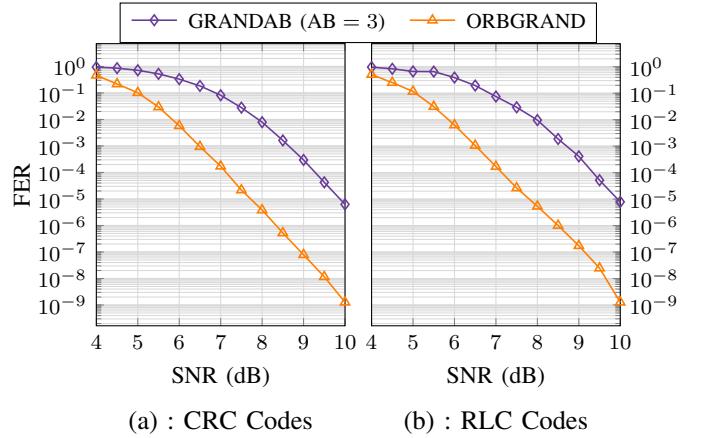


Fig. 1. Comparison of the GRANDAB ($AB = 3$) and ORBGRAND decoding performance using (a) Cyclic Redundancy Check (CRC) codes and (b) Random Linear Codes (RLCs) for $n = 128$ and $k = 104$.

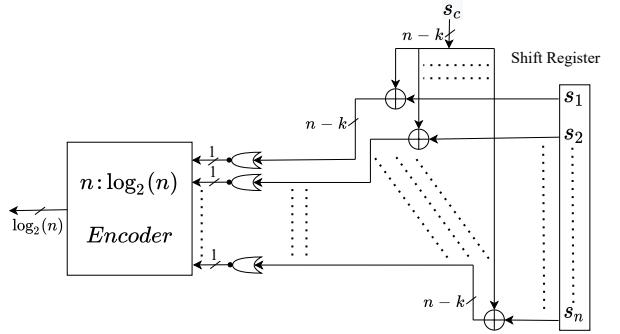


Fig. 2. VLSI architecture for checking error patterns with Hamming weight of 1 ($s_i = \mathbf{H} \cdot \mathbf{1}_i^T$, $i \in [\mathbb{1} \dots n]$).

considered realistic in terms of decodability. For the numerical simulation results presented in Fig. 1, a BPSK modulation and an AWGN channel with variance σ^2 are considered. The SNR in dB is defined as $\text{SNR} = -10 \log_{10} \sigma^2$. For CRC code (128,104), generator polynomial 0xB2B117 is used. As seen in Fig. 1, ORBGRAND outperforms GRANDAB by at least 2 dB for a target FER of 10^{-5} for both CRC codes and RLCs.

To conclude, both GRAND and ORBGRAND can be used to decode any linear block code (n, k) , structured or unstructured, as long as the underlying code's parity check matrix (\mathbf{H}) is provided. Furthermore, as a soft-input decoder, ORBGRAND outperforms hard-input GRANDAB.

D. GRAND VLSI architecture

In [27], a VLSI architecture for GRANDAB ($AB=3$) was proposed. Without going through the details, we briefly explain the technique used in [27], to test error patterns with Hamming weight ≥ 1 . For a linear (n, k) block code, given parity check matrix (\mathbf{H}), the vector $\mathbf{r} \oplus \mathbf{e}$ is a valid codeword if (2) is equal to zero, where \mathbf{e} represents the generated test error pattern.

For the error patterns with Hamming weight of 1 ($\mathbf{1}_i$, with $i \in [\mathbb{1} \dots n]$) using the distributivity rule, code membership can be checked as

$$\mathbf{H} \cdot (\mathbf{r} \oplus \mathbf{1}_i)^T = \mathbf{H} \cdot \mathbf{r}^T \oplus \mathbf{H} \cdot \mathbf{1}_i^T, \quad (3)$$

where $\mathbf{H} \cdot \mathbf{r}^\top$ (denoted as \mathbf{s}_c) is the $(n - k)$ -bits syndrome associated with the received vector \mathbf{r} and $\mathbf{H} \cdot \mathbf{1}_i^\top$ (denoted as \mathbf{s}_i) is the $(n - k)$ -bits syndrome associated with the error pattern with Hamming weight of 1 ($\mathbf{1}_i$).

In the VLSI architecture for GRANDAB (AB=3) [27], a $n \times (n-k)$ -bit shift register is used to store all the n syndromes associated with the error patterns with Hamming weight of 1 (\mathbf{s}_i). To check all the error patterns with Hamming weight of 1, each row of the shift register is combined with the syndrome of the received vector (\mathbf{s}_c) to compute (3) in one time step. Each of the n test syndromes is NOR-reduced, to feed an n -to- $\log_2 n$ priority encoder as depicted in Fig. 2. The output of each NOR-reduce is 1 if and only if all the bits of the syndrome computed by (3) are 0.

Similarly, the test error patterns with Hamming weight of t , $\mathbf{1}_{1,2,\dots,t}$, can be checked for code membership with

$$\mathbf{H} \cdot (\mathbf{r} \oplus \mathbf{1}_{1,2,\dots,t})^\top = \mathbf{s}_c \oplus \mathbf{s}_1 \oplus \mathbf{s}_2 \dots \oplus \mathbf{s}_t. \quad (4)$$

To understand the details of the method used to check error patterns with Hamming weight ≥ 1 , we refer the reader to [27].

III. ORBGRAND DESIGN CONSIDERATIONS

A. Generating the integer partitions

ORBGRAND is centered around generating distinct integer partitions of a particular logistic weight (LW) and these integer partitions are then used for generating test error patterns (e). An integer partition λ of a positive integer m , noted $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_P) \vdash m$ where $\lambda_1 > \lambda_2 > \dots > \lambda_P$, is the multiset of positive integers λ_i ($\forall i \in [1, P]$) that sum to m . If all λ_i are different, the partition is called distinct. Furthermore, the Hamming weight of the generated test error pattern (e) obtained from an integer partition with P elements is P .

In [29] a hardware implementation for the generation of integer partitions was proposed. However, since the generated partitions are not distinct, their approach cannot be directly applied to our proposed ORBGRAND architecture. Furthermore, their partition generation is sequential, which is unsuitable for use in a parallelized, high-throughput hardware architecture. As a result, we suggest a method to generate integer partitions of a specific logistic weight that is based on an arrangement of shift registers and XOR gates.

For generating integer partitions of a specific logistic weight m , we noticed that a breakdown of m generates convenient patterns. For example, for $m = 12$ the distinct integer partitions are $\lambda = \{(12); (11, 1); (10, 2); (9, 3); (8, 4); (7, 5); (9, 2, 1); (8, 3, 1); (7, 4, 1); (6, 5, 1); (7, 3, 2); (6, 4, 2); (5, 4, 3); (6, 3, 2, 1); (5, 4, 2, 1)\}$. If a listing order is followed for $P = 2$ (i.e. the subset $\{(11, 1); (10, 2); (9, 3); (8, 4); (7, 5)\}$), the first integer descends while the second ascends. Therefore for a particular logistic weight m , integer partitions of size 2 ($P = 2$) can be generated as $(\lambda_1, \lambda_2) \vdash m$ where $\lambda_2 \in [1, \lfloor \frac{m}{2} \rfloor - 1]$ and $\lambda_1 = m - \lambda_2$.

Similar trends can be observed for higher-order partitions such as $P = 3$ (i.e. the subset $\{(9, 2, 1); (8, 3, 1); (7, 4, 1); (6, 5, 1); (7, 3, 2); (6, 4, 2); (5, 4, 3)\}$), the first integer descends while the second ascends as the third integer remains

fixed until all iterations for the first two integers are complete. Hence, integer partitions of size 3 ($P = 3$) can be generated as $(\lambda_1, \lambda_2, \lambda_3) \vdash m$ where, $\lambda_3 \in [1, \lambda_3^{\max}]$, $\lambda_2 \in [\lambda_3 + 1, \lambda_{2,\lambda_3}^{\max}]$ and $\lambda_1 = m - \lambda_2 - \lambda_3$. Moreover, λ_3^{\max} is the maximum value of λ_3 , and $\lambda_{2,\lambda_3}^{\max}$ is the maximum value of λ_2 for a specific value of λ_3 ($\lambda_3 \in [1, \lambda_3^{\max}]$).

In general, an integer partition of size P can be generated as $(\lambda_1, \lambda_2, \dots, \lambda_P) \vdash m$ where $\lambda_P \in [1, \lambda_P^{\max}]$, $\lambda_i \in [\lambda_{i+1} + 1, \lambda_{i,\lambda_{i+1},\dots,\lambda_{P-1}}^{\max}] \forall i \in [2, P-1]$ and $\lambda_1 = m - \sum_{i=2}^{P-1} \lambda_i$. Moreover, λ_P^{\max} is the maximum value of λ_P , and $\lambda_{i,\lambda_{i+1},\dots,\lambda_{P-1}}^{\max}$ is the maximum value of λ_i for specific values of λ_j ($\forall j \in [i+1, P-1]$). For simplicity, we will denote $\lambda_{i,\lambda_{i+1},\dots,\lambda_{P-1}}^{\max}$ as λ_i^{\max} . The maximum value for each λ_i ($\forall i \in [2, P]$) is bounded by (5).

Lemma III.1. *If a positive integer m is partitioned into P distinct parts, $\forall i \in [2, P]$, and assuming that λ_i are ordered, the maximum value for each λ_i is bounded by*

$$\lambda_i^{\max} < \frac{2 \times m - (i \times (i-1)) + 2 - 2 \times \sum_{j=i+1}^P \lambda_j}{2 \times i}. \quad (5)$$

whereas the first value of λ is given as $\lambda_1 = m - \sum_{j=2}^P \lambda_j$.

Proof. The proof is provided in Appendix A. \square

B. Analysis of ORBGRAND parameters

As seen in the Algorithm 1, LW_{\max} is an important parameter for the ORBGRAND decoding procedure. LW_{\max} affects not only the decoding performance of ORBGRAND but also the maximum number of codebook membership queries (and hence the worst-case complexity). To reduce the worst case complexity, the LW_{\max} can be appropriately chosen to have minimal impact on the decoding performance of ORBGRAND. In addition to restricting the LW_{\max} , the number of elements (P) in λ can be limited, suggesting that the Hamming weight of the generated test error pattern can be restricted to $\leq P$.

Fig. 3(a) depicts the impact of different parameters (LW_{\max} , P) on ORBGRAND decoding performance for 5G CRC-aided polar code (128,105+11) [30], [31] with BPSK modulation over an AWGN channel. As seen in Fig. 3(a), the FER performance of ORBGRAND outperforms the GRANDAB (AB=3) [10] decoder by at least 2 dB for target FERs $\leq 10^{-5}$. For LW_{\max} values of 128, 96, and 64, the maximum number of codebook membership queries is 5.33×10^7 , 3.69×10^6 , and 1.5×10^5 respectively. When LW_{\max} is decreased from 128 to 64, a performance degradation of 0.2 dB is observed at $\text{FER} = 10^{-7}$, as shown in Fig. 3(a). The complexity, on the other hand, is reduced by $355\times$ as a result of this reduction.

Similarly, the degradation in ORBGRAND decoding performance for the considered polar code with $LW_{\max} = 64$ is negligible when $P = 6$ is used instead of an unbounded P . As a consequence, with $LW_{\max} = 64$ and $P = 6$, the maximum number of queries is limited to 1.16×10^5 , and the ORBGRAND performs similarly to the state-of-the-art

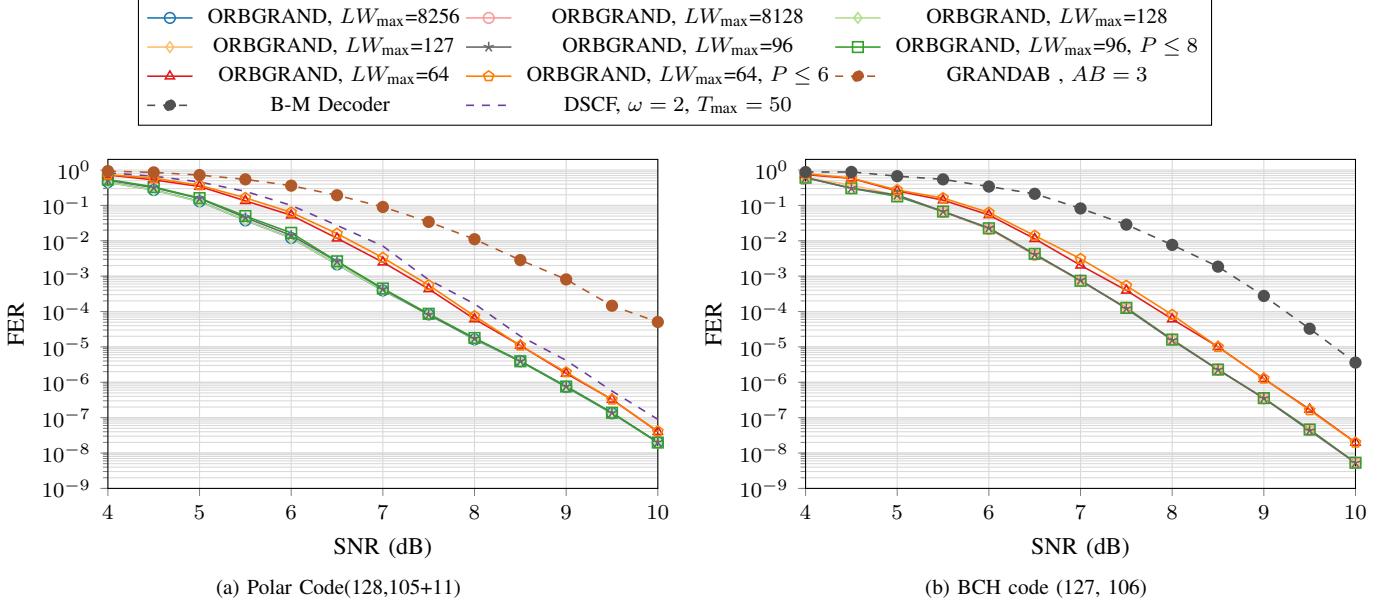


Fig. 3. Comparison of decoding performance of ORBGRAND decoding with different parameters (LW_{\max}, P) for 5G CRC-aided Polar Code (128,105+11) and BCH code (127, 106).

TABLE I
DISPLACEMENT OF LLR ELEMENTS y_i ($\forall i \in [1, n]$) FROM THEIR CORRECT LOCATIONS

Displacement	# of segments for bitonic sorter			
	$S = 2$	$S = 4$	$S = 8$	$S = 16$
$= 0$	10.31%	5.98%	3.87%	2.59%
≤ 1	29.40%	17.42%	11.40%	7.67%
≤ 2	45.50%	28.18%	18.67%	12.65%
≤ 3	58.76%	38.05%	25.67%	17.50%
≤ 5	77.84%	54.62%	38.65%	26.85%
≤ 10	96.89%	81.64%	63.82%	47.68%
≤ 20	99.99%	98.34%	90.09%	75.95%
≤ 30	100%	99.94%	98.10%	90.58%

Dynamic SC-Flip (DSCF) polar code decoder [32], which is also an iterative decoder. The number of DSCF attempts (T_{\max}) parameter is set to 50, and the maximum bit-flipping order (ω) is set to 2.

Figure 3(b) shows FER performance of ORBGRAND decoding of BCH code (127, 106) versus Berlekamp-Massey (B-M) decoder [33], [34]. As compared to the hard decision B-M decoder counterpart, ORBGRAND decoding of BCH code (127, 106) results in a 1.8dB performance gain at a target FER of 10^{-5} . As illustrated in Fig. 3(b), by varying ORBGRAND parameters (LW_{\max}, P), similar trends in decoding performance for ORBGRAND decoding of BCH code (127,106) can be observed.

To conclude, the appropriate selection of ORBGRAND parameters LW_{\max} and P results not only in a reduction in complexity but also in the design of simple hardware, as seen in the section IV.

C. Implementing sorter in ORBGRAND

The ORBGRAND decoding procedure, as described in Algorithm 1, begins by sorting channel LLRs (\mathbf{y}) in ascending order of their absolute value ($|\mathbf{y}|$). Any sorter [35] (*insertion*

sorter, merge sorter, bubble sorter) may be used to sort \mathbf{y} for the proposed ORBGRAND hardware. The sorter choice is determined by the target application's budget in terms of decoding latency and hardware overhead. On one end of the spectrum, we have sequential sorters with high latency but low hardware implementation cost, while on the other, we have parallel sorters with low latency but high hardware implementation cost.

We used a *bitonic sorter* [36] of length n that is pipelined to $\log_2(n)$ stages for the proposed ORGRAND VLSI architecture. As a result, the sorting phase requires just $\log_2(n)$ time-steps (clock cycles). Furthermore, we propose segmenting a *bitonic sorter* of length n into S segments to reduce the area overhead of a parallel *bitonic sorter* implementation.

The segmented *bitonic sorter* is depicted in Fig. 4, which employs four sorters ($S = 4$) of length $\frac{n}{4}$, each of which receives an unique subset of channel LLRs (\mathbf{y}). To generate the final LLRs, the sorted LLRs from individual sorters are concatenated. The first four elements of the final sorted LLRs are comprised of the first element of the output of each sorter. Similarly, the second element of each sorter's output occupies the following four positions of the final sorted LLRs. This procedure is continued until the last elements of each sorter's output are placed in the last four positions of the sorted LLRs, as shown in Fig. 4.

A non-segmented sorter will sort the LLR elements y_i ($\forall i \in [1, n]$) to their correct location. However, compared to a non-segmented sorter, the sorted LLRs using a segmented sorter ($S > 1$) will have elements that are displaced from their correct locations. To investigate the effect of segments on the displacement of LLR elements from their correct locations, we performed Monte-Carlo simulations and measured the percentage of LLR elements that were within a specified distance of their correct location.

Table I compares the displacement of LLR elements from

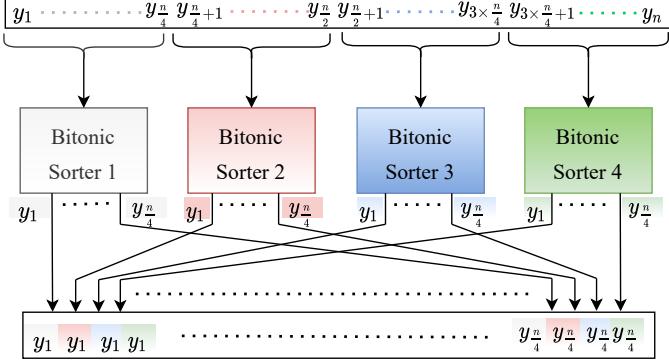


Fig. 4. Proposed segmented sorter ($S = 4$) for ORBGRAND.

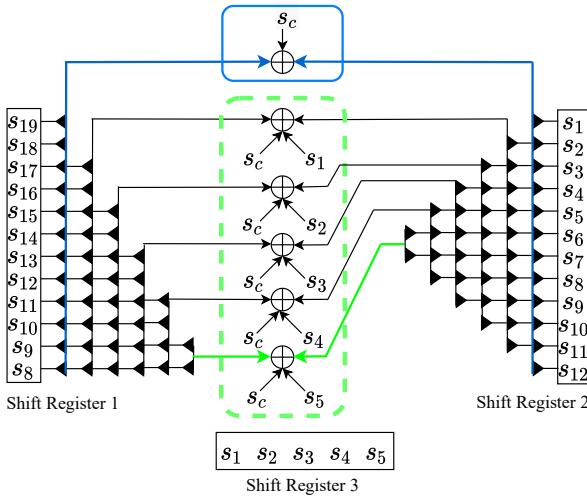


Fig. 5. Example of the shift registers content and interconnection for logistic weight $m = 20$ for checking error patterns of Hamming weights 2 and 3.

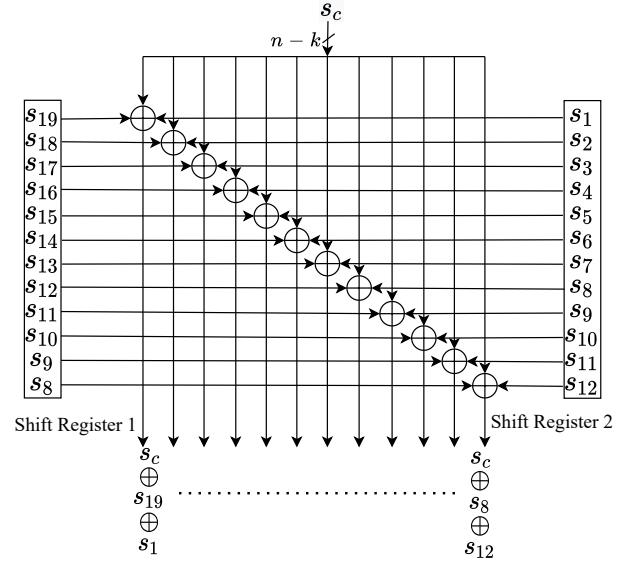
their correct locations with varying numbers of segments employed in the segmented-sorter approach. Table I shows that as the number of segments decreases, more elements are concentrated closer to their correct locations, but as the number of segments increases, LLR elements are concentrated further away from their correct locations. The number of segments influences both the decoding performance and the area overhead of the proposed ORBGRAND VLSI implementation, as detailed in section IV-C.

IV. VLSI ARCHITECTURE FOR ORBGRAND

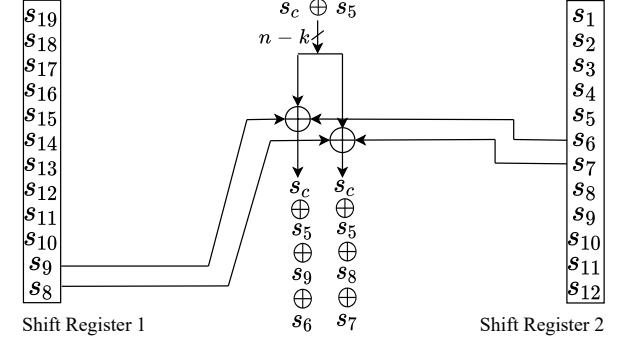
The GRANDAB VLSI architecture in [27] can only generate test error patterns with Hamming weights ≤ 3 . As a result, significant improvements are needed to cater to soft-inputs, to generate error patterns in increasing order of their logistic weight, and to consider larger Hamming weights as required by ORBGRAND.

A. Scheduling and details

As explained in section III, ORBGRAND is based on generating test error patterns corresponding to integer partitions of a specific logistic weight m ($\forall m \in [3, LW_{\max}]$).



(a) Interconnections and the associated XOR gates for the first bus for checking error patterns of Hamming weight of 2 ($P = 2$).



(b) Interconnections and the associated XOR gates for the last (6^{th}) bus for checking error patterns of Hamming weight of 3 ($P = 3$).

Fig. 6. Example of interconnections and the associated XOR gates for the first and last bus for logistic weight $m = 20$.

Moreover, for each m , integer partitions are generated with size P ($\forall P \in [2, P_{\max}]$). We propose to generate these integer partitions in ascending order of their size. This modification does not impact the FER performance, however, it helps in designing a simpler hardware implementation.

The size and number of shift registers used in [27] have a direct impact on the Hamming weight of the error patterns that can be evaluated in parallel. For example, in [27], two $n \times (n - k)$ shift registers are used to evaluate n test error patterns in parallel with a Hamming weight of 2. However, if more shift registers are added, the number of interconnections becomes a problem. As a result, for the proposed ORBGRAND architecture, we choose three shift registers that correspond to an integer partition of size 3 ($P = 3$).

In the proposed ORBGRAND VLSI architecture, λ_1 , λ_2 , λ_3 ($(\lambda_1, \lambda_2, \lambda_3) \vdash m$) are mapped to first, second and third shift register respectively. The third shift register is a $\lambda_3^{\max} \times (n - k)$ bit shift register, where λ_3^{\max} value is given by (5) corresponding to $P = 3$. Whereas the first and second shift registers are each $2 \times (\lambda_3^{\max} + 1) \times (n - k)$ bits in size. Since

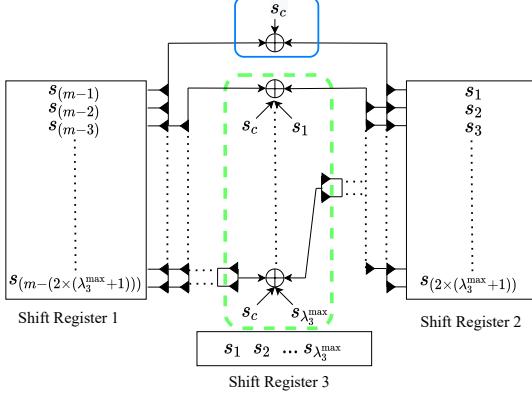


Fig. 7. Shift registers contents for checking error patterns corresponding to a Hamming weight of 2 and 3 for any logistic weight m .

we have $\lambda_1 = m - \sum_{i=2}^3 \lambda_i$, corresponding to $P = 3$, s_{m-i} is stored at the i^{th} index of the first shift register, while for the second and third shift registers s_i is stored at the i^{th} index.

Fig. 5 shows an example of the content and interconnection of three shift registers for logistic weight $m = 20$. The elements of the three shift registers are syndromes (s_i) of the error pattern with Hamming weight of 1. These syndromes (s_i) of the error pattern with Hamming weight of 1 are combined using an array of $(n - k)$ -wide XOR gates to check for error patterns with Hamming weights 2 and 3. A collection of these connections is defined as a *bus*. Since there are numerous connections and XOR gates involved, we used a single XOR gate and a signal *bus* symbol to illustrate these interconnections in Fig. 5. As seen in Fig. 5, there are 6 buses ($\lambda_3^{\max} + 1$, where $\lambda_3^{\max} = 5$ for $P = 3$ (5)) for $m = 20$. The first bus (highlighted by solid rectangle) is used to check error patterns with Hamming weight of 2, and the remaining buses (highlighted by the dashed rectangle in Fig. 5) are used to check for error patterns of Hamming weight 3.

To check the error patterns corresponding to a Hamming weight 2 ($P = 2$), the first bus (highlighted by solid rectangle) is used to combine all the elements of shift register 1 with all the elements of the shift register 2 using an array of XOR gates. These results are again combined with the syndromes of the received vector (s_c) to check for the error patterns with Hamming weight of 2. The detailed interconnections and the associated XOR gates for the first bus are shown in Fig. 6a.

Similarly, to check the error patterns corresponding to a Hamming weight of 3 ($P = 3$), the selected elements of the shift register 1 and 2 are again combined with s_c , but also with the elements of the shift register 3. We use a signal bus and a single XOR gate to illustrate these interconnections, which are depicted in Fig. 5 by the dashed rectangle. The detailed interconnections and the associated XOR gates for the last (6^{th}) bus are shown in Fig. 6b.

Due to the described arrangement and interconnection of the shift registers and XOR gates, all the error patterns corresponding to an integer partition of sizes 2 and 3 for a specific any logistic weight m are checked in one time-step. In general, to check the error patterns corresponding to an

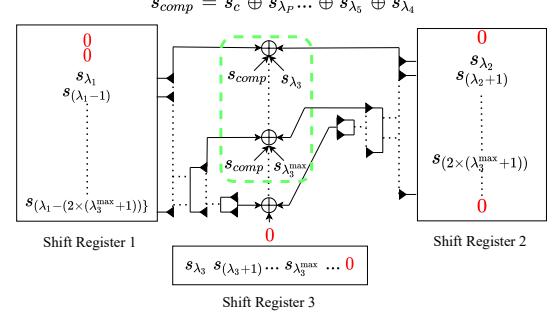


Fig. 8. Shift registers contents for checking error patterns corresponding to $P > 3$ for any logistic weight m ($\lambda_1 = m - \sum_{i=2}^P \lambda_i$).

integer partition of sizes 2 and 3 for any logistic weight m , the content and the interconnection of the three shift registers are depicted in Fig. 7.

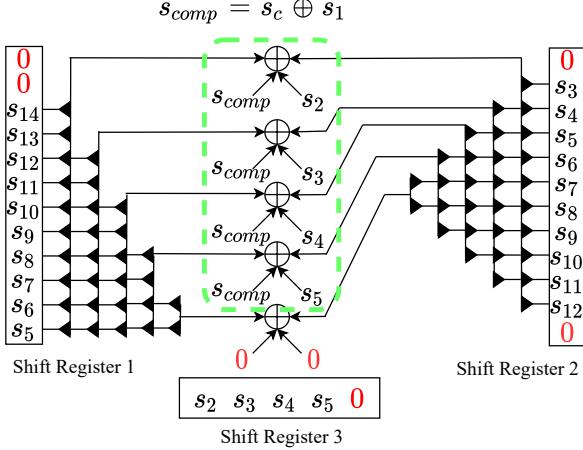
To check all the test error patterns corresponding to integer partitions of sizes $P > 3$, a controller is used in conjunction with the shift registers. The controller combines $P_{\max} - 3$ syndromes together with the syndromes of the received vector, noted s_{comp} . Hence, when s_{comp} is fixed, only one time-step is required to generate all possible combinations of $\{\lambda_1, \lambda_2, \lambda_3\}$ using the shift registers with adequately chosen shift values.

The content and the interconnection of the three shift registers, which are used to check the test error patterns corresponding to integer partitions of sizes $P > 3$, are depicted in Fig. 8. Since the first bus is only used to check error patterns with Hamming weight of 2 ($P = 2$), it is disabled for $P > 3$ and not shown in Fig. 8. A **0** corresponds to a disabled connection, which means the respective elements of the bus, do not take part in the final computations. Fig. 9 illustrates testing error patterns corresponding to $P = 4$. At each time step, the controller outputs $s_{comp} = s_c \oplus s_{\lambda_4}$ ($\lambda_4 \in [1, \lambda_4^{\max}]$) and $\{\lambda_1, \lambda_2, \lambda_3\}$ are computed and mapped to their corresponding shift registers.

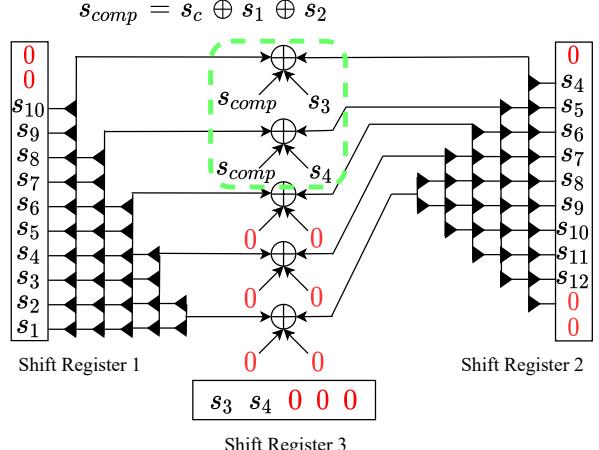
At the first time step, having received $s_{comp} = s_c \oplus s_1$ ($\lambda_4 = 1$) as an output from the controller, λ_3 ($\lambda_3 \in [2, \lambda_3^{\max}]$ where $\lambda_{3,\lambda_4}^{\max} = 5$ with $\lambda_4 = 1$ (5)) is computed and mapped to the third shift register. Similarly, λ_2 ($\lambda_2 \in [\lambda_3 + 1, \lambda_{(2\times(\lambda_3^{\max}+1))}]$) and λ_1 ($\lambda_1 = m - \sum_{i=2}^4 \lambda_i$) are computed and mapped to their corresponding shift registers. The test error patterns with $\lambda_4 = 1$ are checked as shown in Fig. 9a.

At the next time step, the controller outputs $s_{comp} = s_c \oplus s_2$ ($\lambda_4 = 2$) and λ_3 ($\lambda_3 \in [3, \lambda_{3,\lambda_4}^{\max}]$ where $\lambda_{3,\lambda_4}^{\max} = 5$ with $\lambda_4 = 2$ (5)) is computed. Shift register 2 is shifted up by 1 position and shift register 1 outputs λ_1 ($\lambda_1 = m - \sum_{i=2}^4 \lambda_i$) as shown in Fig. 9b. Hence, the test error patterns with $\lambda_4 = 2$ are checked in the second time-step. Similarly, at third time step, the controller outputs $s_{comp} = s_c \oplus s_3$, ($\lambda_4 = 3$) λ_3 ($\lambda_3 \in [4, \lambda_{3,\lambda_4}^{\max}]$ where $\lambda_{3,\lambda_4}^{\max} = 4$ with $\lambda_4 = 3$ (5)) is computed as shown in Fig. 9c. Therefore, a total of 3 time steps ($\lambda_4^{\max} = 3$, Eq. 5), are required to check for error patterns corresponding to $P = 4$ and $m = 20$.

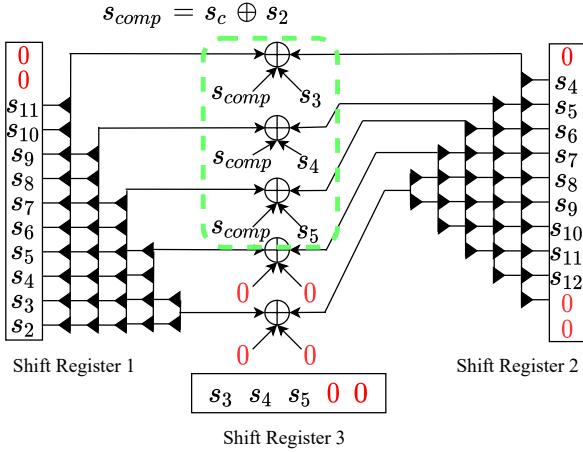
Fig. 10 depicts the use of shift registers to check the error patterns corresponding to $P = 5$ and $m = 20$. At each time step, the controller outputs $s_{comp} = s_c \oplus s_{\lambda_5} \oplus s_{\lambda_4}$. For



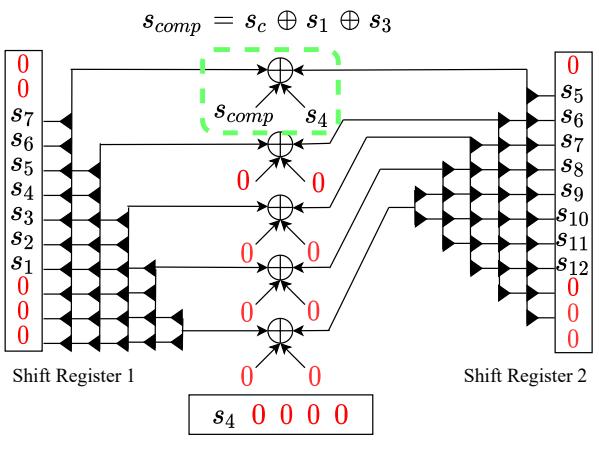
(a) Shift registers contents for checking test error patterns corresponding to $P = 4$ at time step 1.



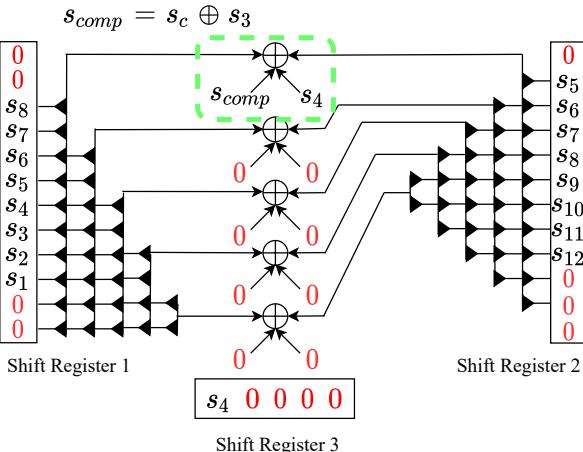
(a) Shift registers contents for checking test error patterns corresponding to $P = 5$ at time step 1.



(b) Shift registers contents for checking test error patterns corresponding to $P = 4$ at time step 2.



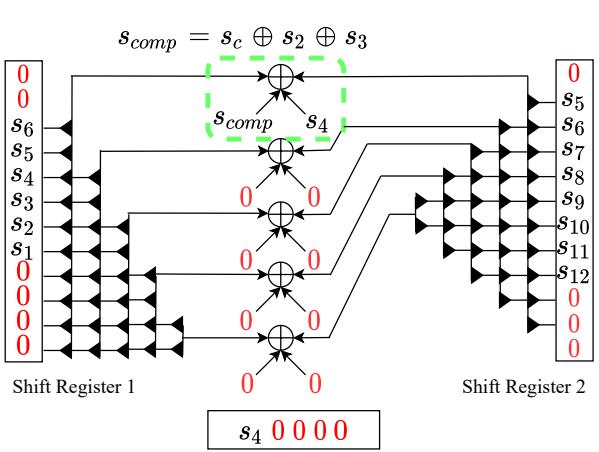
(b) Shift registers contents for checking test error patterns corresponding to $P = 5$ at time step 2.



(c) Shift registers contents for checking test error patterns corresponding to $P = 4$ at time step 3.

Fig. 9. Shift registers contents for checking test error patterns corresponding to $P = 4$ ($m = 20$).

each value of λ_5 ($\lambda_5 \in [1, \lambda_5^{max}]$), λ_4 ($\lambda_4 \in [\lambda_5 + 1, \lambda_4^{max}]$) is computed. Similarly, for each value of λ_4 , λ_3 ($\lambda_3 \in [\lambda_4 + 1, \lambda_3^{max}]$), λ_2 ($\lambda_2 \in [\lambda_3 + 1, \lambda_{(2 \times (\lambda_3^{max} + 1))}]$) and λ_1



(c) Shift registers contents for checking test error patterns corresponding to $P = 5$ at time step 3.

Fig. 10. Shift registers contents for checking test error patterns corresponding to $P = 5$ ($m = 20$).

($\lambda_1 = m - \sum_{i=2}^5 \lambda_i$) are computed and mapped to their corresponding shift registers. Hence, a total of 3 time steps

$(\sum_{\lambda_5=1}^{\lambda_5^{\max}} \left(\sum_{\lambda_4=\lambda_5+1}^{\lambda_4^{\max}} (1) \right)$, where $\lambda_5^{\max} = 2$, $\lambda_{4,\lambda_5=1}^{\max} = 3$ and $\lambda_{4,\lambda_5=2}^{\max} = 3$ (5)) are required to check for error patterns corresponding to $P = 5$ and $m = 20$ as shown in Fig. 10

In general, the number of time steps required to generate all integer partitions of size $P > 3$ for a specific logistic weight (LW) is given by:

$$\sum_{\lambda_P=1}^{\lambda_P^{\max}} \left(\sum_{\lambda_{P-1}=\lambda_P+1}^{\lambda_{P-1,\lambda_P}^{\max}} \left(\dots \sum_{\lambda_4=\lambda_5+1}^{\lambda_{4,\lambda_5,\dots,\lambda_P}^{\max}} (1) \right) \right). \quad (6)$$

B. Proposed VLSI architecture

Figure 11 depicts the proposed VLSI architecture for ORBGRAND which can be used to decode any linear block code of length n . For clarity, the control and clock signals are not shown. To support different codes and rates, any H matrix can be loaded into the H memory of size $(n - k) \times n$ -bit at any time. The hard decided vector \hat{y} is subjected to a syndrome check (2) in the first phase of decoding. Decoding is assumed to be successful if the syndrome is verified. Otherwise, the LLRs values are sorted in ascending order of their absolute value $|y|$.

As depicted in Fig. 11, the sorted syndromes of error patterns with Hamming weight of 1 (s_i) are passed to the *decoder core*, while the indices of the sorted LLRs are forwarded to the multiplexers for later use by the *word generator* module. Following the sorting process, all syndromes of error patterns with Hamming weight of 1 (s_i) are tested for codebook membership (3) in a single time-step. Following that, error patterns are tested for codebook membership in ascending logistic weight (LW) order as explained in section II-C.

The test error pattern syndromes corresponding to integer partitions of a given logistic weight m are generated using the shift register and XOR gate arrangement proposed in section IV-A. The rows of shift registers are combined with the controller's output (s_{comp}), and the resulting test syndromes are NOR-reduced and fed to a 2D priority encoder. Each NOR-reduce output is 1 if and only if all of the bits of the symptom computed by (4) are 0. If any of the tested syndrome combinations satisfy the parity check constraint (NOR-reduced output is 1), the 2D priority encoder is used in conjunction with the *controller* module to forward the respective indices to the word generator module, where P multiplexers are used to convert the sorted index values to their appropriate bit-flip locations.

C. Implementation results

The proposed ORBGRAND VLSI architecture with parameters ($LW \leq 64$, $P \leq 6$) and ($LW \leq 96$, $P \leq 8$), has been implemented in Verilog HDL and synthesized using Synopsys Design Compiler with general-purpose TSMC 65 nm CMOS technology. The design has been verified using test benches generated via the bit-true C model of the proposed hardware. Table II presents the synthesis results for the proposed decoder with $n = 128$ and the proposed architecture can support code rates between 0.75 and 1. Input channel LLRs are quantized on 5 bits, including 1 sign bit and 3 bits for the fractional

part. To ensure accuracy in power measurements, switching activities from real test vectors are extracted for all of the VLSI architectures presented in Table II.

The maximum frequency supported by the ORBGRAND implementation is 454 MHz. Since there is no pipelining technique for the decoder core, one clock cycle corresponds to one time-step. At $\text{SNR} = 10$ dB (target FER of 10^{-7}), the average latency is only 2.47ns, resulting in an average decoding information throughput of 42.5 Gbps for a (128,105) 5G-NR CRC-Aided polar code. However, the worst-case (W.C.) scenario needs 4224 cycles with $n = 128$ and parameters $LW \leq 64$ and $P \leq 6$, culminating in a W.C. latency of $9.3\mu\text{s}$. Similarly, parameters $LW \leq 96$ and $P \leq 8$ result in $205.76\mu\text{s}$ W.C. latency. Fig. 12 depicts the W.C. latency and W.C. information throughput ($k = 105$) for the proposed ORBGRAND VLSI architecture with parameters LW and P .

As seen in Fig. 3 (a), ORBGRAND with parameters $LW \leq 64$ and $P \leq 6$ performs similar to the Dynamic SC-Flip (DSCF) [32] decoder for (128,105) 5G-NR CRC-Aided (CA) polar code. The proposed ORBGRAND VLSI implementation ($LW \leq 64$ and $P \leq 6$) is compared to VLSI architecture for DSCF polar code decoder ($\omega = 2$, $T_{\max} = 50$) [37], which employs 6 and 7 bit internal and channel LLR quantizations, respectively. Compared to DSCF [37], ORBGRAND ($LW \leq 64$, $P \leq 6$) has a $8\times$ area overhead, as well as a 52% increase in the worst-case latency. However, at a target FER of 10^{-7} , the proposed ORBGRAND results in $49\times$ lower average latency and higher average throughput than the DSCF [37]. Furthermore, as compared to DSCF [37], ORBGRAND ($LW \leq 64$, $P \leq 6$) is $5\times$ more area efficient and $32\times$ more energy efficient. Moreover, the proposed ORBGRAND hardware is code and rate compatible, while the DSCF [37] can only decode polar codes.

In comparison to the hard-input GRANDAB decoder ($AB=3$) [27], ORBGRAND ($LW \leq 64$, $P \leq 6$) has a $7\times$ area overhead, as well as a 13.5% higher W.C. and a 23.5% higher average latency. This leads to a 13.3% decrease in W.C. and a 23.5% decrease in average decoding throughput, respectively. Furthermore, as compared to [27], ORBGRAND ($LW \leq 64$, $P \leq 6$) is $2\times$ less energy efficient and $9\times$ less area efficient. However, as seen in Fig. 3, the FER performance of ORBGRAND ($LW \leq 64$, $P \leq 6$), a soft decision decoder, outperforms the GRANDAB ($AB=3$) decoder by at least $1.4 \sim 2$ dB for target FERs $\leq 10^{-5}$.

As shown in Fig. 3, ORBGRAND with parameters $LW \leq 96$ and $P \leq 8$ results in a $0.2 \sim 0.3$ dB gain in decoding performance when compared to ORBGRAND with parameters $LW \leq 64$ and $P \leq 6$ at a target FER of $\leq 10^{-7}$. As presented in Table II, the ORBGRAND implementation with parameters $LW \leq 96$ and $P \leq 8$ incurs a 23.6% area overhead as compared to the ORBGRAND implementation with parameters $LW \leq 64$ and $P \leq 6$. To reduce this area overhead, the *bitonic sorter* in the proposed ORBGRAND VLSI implementation, as discussed in section III-C, can be segmented into S segments. This number of segments (S) influences both the decoding performance and the area overhead of the proposed ORBGRAND VLSI implementation.

Fig. 13 depicts the FER performance of implementing a

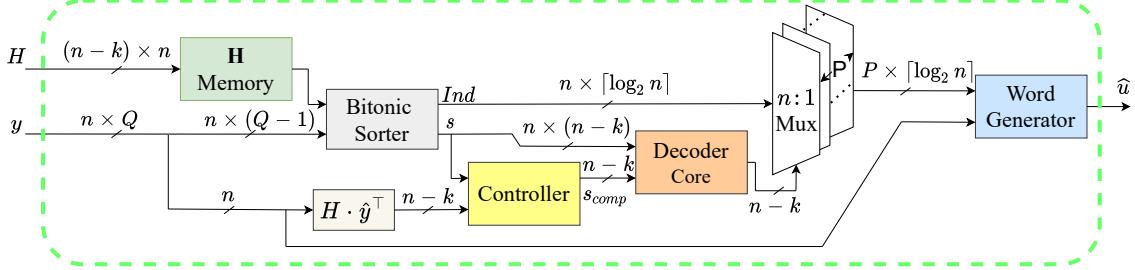


Fig. 11. Proposed VLSI Architecture for ORBGRAND.

TABLE II
TSMC 65 NM CMOS IMPLEMENTATION COMPARISON FOR ORBGRAND WITH GRANDAB AND DSCF FOR $n = 128$.

	GRANDAB [27]			ORBGRAND			DSCF [37]
Parameters	AB=3	LW ≤ 64 , P ≤ 6	LW ≤ 96 , P ≤ 8	LW ≤ 96 , P ≤ 8 , S = 2	LW ≤ 96 , P ≤ 8 , S = 4		$\omega = 2$, $T_{max} = 50$
Technology (nm)	65	65	65	65	65		65
Supply (V)	0.9	0.9	0.9	0.9	0.9		0.9
Max. Frequency (MHz)	500	454	454	454	454		426
Area (mm 2)	0.25	1.82	2.25	2.08	1.85		0.22
W.C. Latency (μ s)	8.196	9.30	205.76	205.76	205.76		6.103
Avg. Latency (ns)	2	2.47	2.47	2.47	2.47		122
W.C. T/P (Mbps)	12.8	11.3	0.512	0.512	0.512		17.2
Avg. T/P (Gbps)	52.5	42.5	42.5	42.5	42.5		0.86
Power (mW)	46	104.3	133	131.3	130		68.51
Energy per Bit (pJ/bit)	0.87	2.45	3.13	3.09	3.0		79.6
Area Efficiency (Gbps/mm 2)	210	23.3	18.9	20.4	23		3.9
Code compatible	Yes	Yes	Yes	Yes	Yes		No

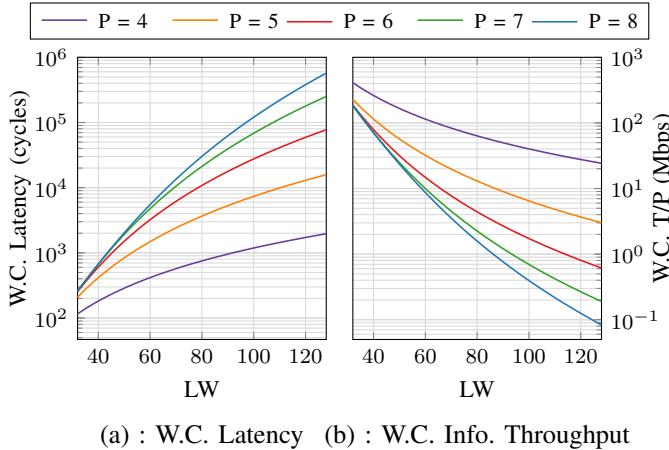


Fig. 12. Worst-Case (W.C.) latency and W.C. information throughput for the proposed ORBGRAND architecture with various parameters (LW, P).

segmented sorter approach for ORBGRAND ($LW \leq 96$, $P \leq 8$) decoding of 5G CA (128,105) polar code. As seen in the Figure 13, the ORBGRAND using the segmented sorter with $S = 2$ and $S = 4$ suffers from a FER performance degradation of 0.06dB and 0.3dB respectively, at the target FER of 10^{-5} , as compared to ORBGRAND with a non-segmented sorter. As presented in Table II, compared to the ORBGRAND ($LW \leq 96$, $P \leq 8$) using the non-segmented sorter, the proposed ORBGRAND ($LW \leq 96$, $P \leq 8$) with sorter parameters $S = 2$ and $S = 4$ results in 8% and 21.6% less area overhead and a 8.2% and 21.6% increase in area efficiency respectively.

To conclude, the parameters of ORBGRAND (LW , P and

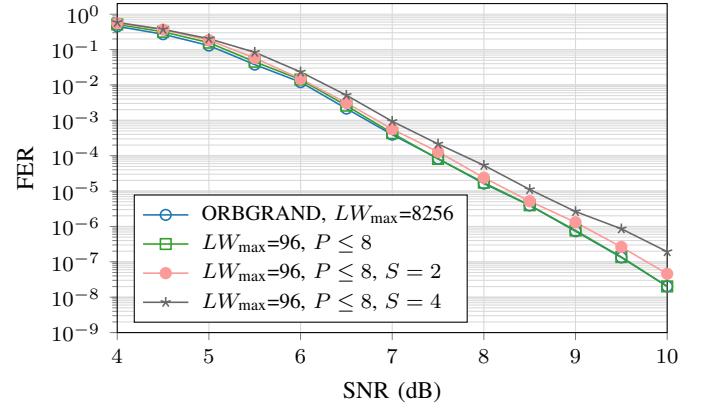


Fig. 13. Comparison of decoding performance of ORBGRAND decoding with different parameters (LW_{max} , P , S) for Polar Code(128,105+11)

S) can be tuned to strike a balance between decoding performance requirements, area overhead, and worst-case decoding latency/throughput budget (Fig. 12) for a target application.

V. CONCLUSION

We proposed the first hardware architecture for the ORBGRAND algorithm in this paper. ORBGRAND is a soft input variant that generates test error patterns in a fixed logistic weight order that is independent of the vector of channel observation values, rendering it suitable for parallel hardware implementation. Due to the code-agnostic nature of the GRAND and its variants, the proposed ORBGRAND architecture can decode any code as long as the length and

rate constraints are met. We suggest modifications in the ORBGRAND algorithm to simplify the hardware implementation and reduce the decoding complexity. Furthermore, the proposed ORBGRAND VLSI architecture uses parameters that can be tweaked to meet the optimal decoding performance as well as the decoding latency for a specific application. According to ASIC synthesis results, an average decoding throughput of 42.5 Gbps can be achieved for a code length of 128 and a target FER of 10^{-7} . The proposed VLSI architecture improves decoding performance by at least 2 dB over the GRANDAB, a hard-input variant of GRAND. In comparison to the state-of-the-art DSCF decoder for polar codes, the proposed VLSI implementation achieves $49\times$ higher decoding throughput, $32\times$ higher energy efficiency, and $5\times$ higher area efficiency for a 5G CA (128,105) polar code. Furthermore, the proposed architecture is the first step toward the hardware implementation of GRAND family soft-input decoders.

APPENDIX A PROOF OF LEMMA 1

Proof. It is sufficient to show that for all i ($i \in [2, P]$)

$$\lambda_i^{\max} < \frac{2 \times m - (i \times (i-1)) + 2 - 2 \times \sum_{j=i+1}^P \lambda_j}{2 \times i}. \text{ We use induction on } i.$$

Base case ($i = 2$):

$$\lambda_2^{\max} < \frac{m - \sum_{j=3}^P \lambda_j}{2}$$

Since the λ_i are ordered $\therefore \lambda_2 < \lambda_1$

$$\begin{aligned} \Rightarrow \lambda_2^{\max} &< m - \sum_{j=2}^P \lambda_j \quad (\because \lambda_1 = m - \sum_{j=2}^P \lambda_j) \\ \Rightarrow \lambda_2^{\max} &< m - \sum_{j=3}^P \lambda_j - \lambda_2^{\max} \\ \therefore \lambda_2^{\max} &< \frac{m - \sum_{j=3}^P \lambda_j}{2} \end{aligned}$$

Inductive hypothesis ($i = k$):

$$\lambda_k^{\max} < \frac{2 \times m - (k \times (k-1)) + 2 - 2 \times \sum_{j=k+1}^P \lambda_j}{2 \times k}$$

Inductive step ($i = k+1$):

$$\lambda_{k+1}^{\max} < \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+2}^P \lambda_j}{2 \times (k+1)}$$

\therefore Inductive hypothesis

$$\begin{aligned} \lambda_k^{\max} &< \frac{2 \times m - (k \times (k-1)) + 2 - 2 \times \sum_{j=k+1}^P \lambda_j}{2 \times k} \\ \Rightarrow \lambda_k^{\max} &< \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+1}^P \lambda_j}{2 \times k} + 1 \end{aligned}$$

$$\begin{aligned} &(\because k \times (k-1) = k \times (k+1) + 2 \times (k)) \\ \Rightarrow \lambda_k^{\max} &< \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+2}^P \lambda_j}{2 \times k} \\ &+ 1 - \frac{\lambda_{k+1}^{\max}}{k} \\ &(\because \sum_{j=k+1}^P \lambda_j = \lambda_{k+1}^{\max} + \sum_{j=k+2}^P \lambda_j) \\ \Rightarrow \frac{\lambda_{k+1}^{\max} + \lambda_k^{\max} - k}{k} &< \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+2}^P \lambda_j}{2 \times k} \\ &\Rightarrow \frac{\lambda_{k+1}^{\max} + \lambda_k^{\max} - k}{k+1} \\ &< \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+2}^P \lambda_j}{2 \times (k+1)} \quad (7) \end{aligned}$$

Since λ_i are ordered $\therefore \lambda_{k+1} < \lambda_k$

$$\Rightarrow \lambda_{k+1}^{\max} + 1 \leq \lambda_k^{\max} \quad (\because \lambda_{k+1}^{\max} + 1 \leq \lambda_k \leq \lambda_k^{\max})$$

$$\Rightarrow k \times \lambda_{k+1}^{\max} + k + \lambda_{k+1}^{\max} \leq k \times \lambda_k^{\max} + \lambda_{k+1}^{\max}$$

$$\Rightarrow \lambda_{k+1}^{\max} \leq \frac{\lambda_{k+1}^{\max} + \lambda_k^{\max} - k}{k+1}$$

$$\text{Using (7): } \lambda_{k+1}^{\max} \leq \frac{\lambda_{k+1}^{\max} + \lambda_k^{\max} - k}{k+1}$$

$$< \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+2}^P \lambda_j}{2 \times (k+1)}$$

$$\Rightarrow \lambda_{k+1}^{\max} < \frac{2 \times m - (k \times (k+1)) + 2 - 2 \times \sum_{j=k+2}^P \lambda_j}{2 \times (k+1)} \quad (\because a \leq b < c \Rightarrow a < c)$$

□

REFERENCES

- [1] Claude Elwood Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, 1948.
- [2] Richard Wesley Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, pp. 147–160, 1950.
- [3] Alexis Hocquenghem, "Codes correcteurs d'erreurs," *Chiffres*, 1959.
- [4] Raj Chandra Bose and Dwijendra K Ray-Chaudhuri, "On a class of error correcting binary group codes," *Information and control*, vol. 3, no. 1, pp. 68–79, 1960.
- [5] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding: Turbo-codes. 1," in *Proceedings of ICC '93 - IEEE International Conference on Communications*, 1993, vol. 2, pp. 1064–1070 vol.2.

- [6] R. Gallager, "Low-density parity-check codes," *IRE Transactions on information theory*, vol. 8, no. 1, pp. 21–28, 1962.
- [7] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.
- [8] Eren Şaşoğlu, Emre Telatar, and Erdal Arikan, "Polarization for arbitrary discrete memoryless channels," in *2009 IEEE Information Theory Workshop*, 2009, pp. 144–148.
- [9] M. C. Coşkun, G. Durisi, T. Jerkovits, G. Liva, W. Ryan, B. Stein, and F. Steiner, "Efficient error-correcting codes in the short blocklength regime," *Physical Communication*, vol. 34, pp. 66–79, 2019.
- [10] K. R. Duffy, J. Li, and M. Médard, "Capacity-achieving guessing random additive noise decoding," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4023–4040, 2019.
- [11] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.
- [12] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5g: Ran, core network and caching solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.
- [13] Zheng Ma, Ming Xiao, Yue Xiao, Zhibo Pang, H. Vincent Poor, and Branka Vucetic, "High-reliability and low-latency wireless communication for internet of things: Challenges, fundamentals, and enabling technologies," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7946–7970, 2019.
- [14] Ming Zhan, Zhibo Pang, Dacfeey Dzung, and Ming Xiao, "Channel coding for high performance wireless control in critical applications: Survey and analysis," *IEEE Access*, vol. 6, pp. 29648–29664, 2018.
- [15] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, 2018.
- [16] M. P. C. Fossorier and S. Lin, "Soft-decision decoding of linear block codes based on ordered statistics," vol. 41, no. 5, pp. 1379–1396, 1995.
- [17] Thibaud Tonnellier, Marzieh Hashemipour, Nghia Doan, Warren J. Gross, and Alexios Balatsoukas-Stimming, "Towards practical near-maximum-likelihood decoding of error-correcting codes: An overview," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8283–8287.
- [18] Wei An, Muriel Médard, and Ken R. Duffy, "Keep the bursts and ditch the interleavers," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.
- [19] K. R. Duffy and M. Médard, "Guessing random additive noise decoding with soft detection symbol reliability information - SGRAND," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 480–484.
- [20] Ken R. Duffy, "Ordered reliability bits guessing random additive noise decoding," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8268–8272.
- [21] A. Solomon, K. R. Duffy, and M. Médard, "Soft maximum likelihood decoding using grand," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.
- [22] A. Valembois and M. Fossorier, "An improved method to compute lists of binary vectors that optimize a given weight function with application to soft-decision decoding," vol. 5, no. 11, pp. 456–458, 2001.
- [23] S. M. Abbas, T. Tonnellier, F. Ercan, M. Jalaleddine, and W. J. Gross, "High-Throughput VLSI Architecture for Soft-Decision Decoding with ORBGRAND," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8288–8292.
- [24] W. W. Peterson and D. T. Brown, "Cyclic codes for error detection," *Proceedings of the IRE*, vol. 49, no. 1, pp. 228–235, 1961.
- [25] Robert G. Gallager, "Information theory and reliable communication," 1968.
- [26] J.T. Coffey and R.M. Goodman, "Any code of which we cannot think is good," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1453–1461, 1990.
- [27] S. M. Abbas, T. Tonnellier, F. Ercan, and W. J. Gross, "High-throughput VLSI architecture for GRAND," in *2020 IEEE Workshop on Signal Processing Systems (SiPS)*, 2020, pp. 1–6.
- [28] Wei An, Muriel Médard, and Ken R. Duffy, "CRC codes as error correcting codes," *arXiv preprint arXiv:2104.13663*, 2021.
- [29] J. Butler and T. Sasao, "High-speed hardware partition generation," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 7, pp. 1–17, 01 2014.
- [30] K. R. Duffy, A. Solomon, K. M. Konwar, and M. Médard, "5G NR CA-Polar maximum likelihood decoding by GRAND," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2020, pp. 1–5.
- [31] 3GPP, "NR: Multiplexing and Channel Coding," Tech. Rep. TS 38.212, April 2020, Rel. 16.1.
- [32] L. Chandresris, V. Savin, and D. Declercq, "Dynamic-scflip decoding of polar codes," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2333–2345, 2018.
- [33] E. Berlekamp, "Nonbinary BCH decoding (abstr)," *IEEE Transactions on Information Theory*, vol. 14, no. 2, pp. 242–242, 1968.
- [34] J. Massey, "Shift-register synthesis and BCH decoding," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 122–127, 1969.
- [35] Thomas H. Cormen, *Algorithms for Sorting and Searching*, pp. 25–59, 2013.
- [36] K. E. Batcher, "Sorting networks and their applications," in *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference*, New York, NY, USA, 1968, AFIPS '68 (Spring), p. 307–314, Association for Computing Machinery.
- [37] F. Ercan, T. Tonnellier, N. Doan, and W. J. Gross, "Practical dynamic SC-flip polar decoders: Algorithm and implementation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5441–5456, 2020.