Polar Code decoder exploration framework

Claus Kestel, Stefan Weithoffer, and Norbert Wehn
Microelectronic Systems Design Research Group, Technische Universität Kaiserslautern, Kaiserslautern, Germany

Correspondence: Claus Kestel (kestel@eit.uni-kl.de)

Received: 25 January 2018 – Revised: 19 March 2018 – Accepted: 6 May 2018 – Published: 4 September 2018

Abstract. The increasing demand for fast wireless communications requires sophisticated baseband signal processing. One of the computational intense tasks here is advanced Forward Error Correction (FEC), especially the decoding. Finding efficient hardware implementations for sophisticated FEC decoding algorithms that fulfill throughput demands under strict implementation constraints is an active research topic due to increasing throughput, low latency, and high energy efficiency requirements.

This paper focuses on the interesting class of Polar Codes that are currently a hot topic. We present a modular framework to automatically generate and evaluate a wide range of Polar Code decoders, with emphasis on design space exploration for efficient hardware architectures. To demonstrate the efficiency of our framework a very high throughput Soft Cancellation (SCAN) Polar Code decoder is shown that was automatically generated. This decoder is, to the best of our knowledge, the fastest SCAN Polar Code decoder published so far.

1 Introduction

Over the last decades, digital communication systems became an essential part of our lives. We consume, produce, and store an ever increasing amount of data. Their transport requires fast networks and advanced baseband signal processing to handle the growing amount of wireless traffic. Here, the demand for high energy efficiency, low latency, and high throughput is increasing. One crucial and challenging part of communication systems is channel coding, or FEC, that copes with transmission errors in noisy environments (Weithoffer et al., 2017).

A class of codes, that gained much interest in the last years for achieving the channel capacity (Shannon, 1948) for specific channels, are Polar Codes. This property and their subsequent adoption for the next generation mobile networks (5G) has put Polar Codes in the focus of many investigations. In this paper we only focus on the decoding because it’s the most computationally intense part.

Various hardware implementations for Polar Code decoders have been proposed in recent years. They cover pipelined high-throughput architectures (Giard et al., 2017), flexible (Dizdar and Arikan, 2016), and iterative decoder implementations (Park et al., 2014) to name but a few of them. These decoders are generally optimized only for one specific target, e.g., throughput, area or code rate flexibility, and especially for only one specific decoding algorithm. Moreover, a thorough analysis and comparison of the various architectures with respect to trade-offs is missing from literature.

2 Contribution and outline

In this paper, we present a framework for fast exploration of the design space for Polar Code decoder hardware implementations. The framework also allows the fast adaption to changing requirements, i.e., the code structure during the standardization of 5G.

The contributions are as follows: First, we propose Polar Factor Trees as a generalized model of Polar Code decoding. Based on the Polar Factor Tree, we present an architectural template for high-throughput Polar Code decoders. The Polar Factor Tree and the architectural template are the basis of our framework. Second, we present a framework that allows fast exploration of the Polar Code decoder design space. Output are detailed design metrics (throughput, area, etc.). And third, as a proof-of-concept of our framework, we present the first high-throughput SCAN Polar Code decoder that achieves more than 100 Gbps on a 28 nm technology.

The paper is structured as follows: In Sect. 3 a short introduction to Polar Codes, their encoding and decoding algorithms is given. Section 4 focuses on high-throughput de-
3 Polar Codes

Polar Codes, invented 2009 by Erdal Arikan, are the first codes proven to achieve channel capacity (Arikan, 2009) for Binary Symmetric Memoryless Channels (BSMC). Polar Codes are block codes and belong to the class of multi-level concatenated codes, but, in contrast to the similar Reed-Muller-Codes, they use the phenomenon of channel polarization to maximize coding efficiency. As a candidate for the FEC of the next generation mobile networks (5G) they play an important role in the ongoing standardization process.

3.1 Channel polarization

Channel polarization is a transformation of $N$ independent copies of a channel $W$ into a new set of $N$ channels $\{W^{(i)}_N\}$, where, for large $N$, the symmetric channel capacity $I(W^{(i)}_N)$ approaches either 0 (completely noisy) or 1 (completely error-free). Figure 1 shows an example for a set of two channels before and after the polarization. The channel with decreased capacity is called $W^-$ (red), and the other with the improved capacity is called $W^+$ (blue).

3.2 Code construction

The butterfly structure in Fig. 1b represents the smallest possible Polar Code with $N = 2$. Larger codes can be built by recursively combining two codes of length $N/2$ to a new code of length $N$. Therefore, to create a code of length $N, 2^{\log_2 N}$ steps of combining are necessary. The code rate $N/K$ is then selected by setting the least reliable channels to fixed known values, e.g., 0 (called frozen bits), and using the more reliable channels to transmit the information bits. Figure 2 shows the structure of a Polar Code of length $N = 8$ and $K = 4$ information bits $u_{1..4}$. The frozen bits are set to 0.

Polar Codes are defined with a generator matrix $G_N$ that is built by using the Kronecker power $F_2^\otimes \log_2 N$ of $F_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. Note that $F_2$ represents the butterfly structure of Fig. 1b, and the Kronecker power equals the combination of smaller codes into bigger ones.

3.3 Encoding

The encoding is shown in Fig. 2a for a $N = 8/K = 4$ Polar Code. A sequence of information bits $u_{1..4}$ and frozen bits is passed through the code structure from left to right to get a codeword $x_{1..8}$ of length $N = 8$. Formally, the encoding is $x = uG_N$ with $G_N$ being the generator matrix and $u$ the information vector that includes the frozen bits.

3.4 Decoding

The decoding of Polar Codes can be seen as the reversal of the encoding. Thus, decoding passes the received channel values $y$ from left to right as shown in Fig. 2b.

But while the encoding is simply a linear transformation, the decoding is the estimation of the originally transmitted information bits $\hat{u}$ from the received channel values $y$, commonly represented as Log Likelihood Ratios (LLRs). LLRs are a measure of the probability of a received value corresponding to either 0 or 1 and are defined...
as $\text{LLR}(y_i) = \log \left( \frac{P(y_i|1)}{P(y_i|0)} \right)$. This estimation can be done by different algorithms.

The most popular decoding algorithm for Polar Codes is the Successive Cancellation (SC) algorithm that was presented by Arikan (2009) in his original paper. SC passes the received channel values from left to right over the Polar Factor Graph in Fig. 2b. Four types of operations have to be performed for every butterfly structure (see Sect. 4.1). The decoding of the individual bits is performed in a recursive manner on the Polar Factor Graph.

An extension of SC is the Successive Cancellation List (SCL) algorithm. In this algorithm a list of $L$ most probable decisions are considered, i.e., $L$ codewords are decoded. The final codeword is selected out of the list by, e.g., the highest overall probability or by other checks like a Cyclic Redundancy Check (CRC). The SC and SCL are sequential algorithms.

An inherently parallel decoding algorithm is the Belief Propagation (BP) algorithm for Polar Codes. It iterates from left to right and reverse over the Polar Code Factor Graph from Fig. 2b. Thus, the bits are decoded in an iterative manner similar to the BP for Low Density Parity Check (LDPC) codes. However, Abbas et al. (2017) showed that up to 100 iterations are necessary to match the error correcting performance of the SC algorithm.

The SCAN algorithm (Fayyaz and Barry, 2014) uses the general processing order of the SC algorithm in combination with the exchange of soft values from the BP algorithm. The SCAN decoding algorithm is highly parameterizable in terms of processing order of the nodes and allows various trade-offs.

Figure 3 shows the communications performance of the SC, SCL, and SCAN algorithms, respectively. While SC and SCAN provide similar communications performance and mostly differ in the hard or soft output, the SCL algorithm performs significantly better, but at the cost of increased decoding complexity. The SCL was simulated with list size $L = 4$ and the SCAN algorithm with one iteration $I = 1$ for a code of $N = 1024$, $K = 512$ over an AWGN channel and floating-point arithmetic. The Polar Code was constructed by using the gaussian approximation method (Vangala et al., 2015) for a design Signal-to-Noise Ratio (SNR) of 0 dB.

4 Modeling of high-throughput architectures

In Sect. 3 we used the Polar Factor Graph to explain decoding. A compressed version of this graph is the Polar Factor Tree (see Fig. 4a, b). For that, groups of concurrent calculation steps are combined and represented as a single node in the tree. A row of operations is mapped on the same tree level. Leaf nodes, which are frozen, are marked white (in Fig. 4b), leaf nodes containing information bits blue. Non-leaf nodes containing only frozen or information bits as child nodes are marked white and blue, respectively. Nodes with mixed leaf nodes are colored half white and half blue.

SC, SCL, and SCAN decoding can be modeled by a depth-first search on this tree. Whenever during search a node is visited, a corresponding operation has to be performed. BP corresponds to a breadth-first search on the same tree.

The factor tree is a better representation than the graph representation and used in many approaches: Alamdar-Yazdi and Kschischang (2011), Sarkis et al. (2014), Sarkis et al. (2016), and Lin et al. (2017). We use the Polar Factor Tree as central data structure for our framework.

4.1 Node operations

As said, all decoding algorithms can be mapped onto a depth-first (SC, SCL, SCAN) or breadth-first search (BP) on this Polar Factor Tree. Here we focus on the first class of algorithms. The differences in the three decoding algorithm is in the node operations while the tree is traversed.

Figure 4c and d show the performed operations of one butterfly unit and one node, respectively. $\alpha_t$, $\beta_t$, and $\beta_v$ denote vectors that are input to a node (marked in gray). $\alpha_t$, $\alpha_r$, and $\beta_u$ are corresponding vectors calculated by the node (marked red). The notation $\alpha_{t,i}$ with $i = 0, \ldots, N$ is used to address one element of the corresponding vector. The dimensions of the vectors depend on the tree level $t$, starting with $t = 0$ for the root node. For the vectors $\alpha_e$ and $\beta_v$ the length is $N/2^t$ and for $\alpha_t$, $\beta_t$, $\alpha_r$, and $\beta_v$ the length is $N/2^{t+1}$.

Every node (Fig. 4d) that is visited during depth-first search first calculates a new vector of LLRs $\alpha_t$ and sends it to the left child. In the backtracking phase of the search, the left child provides $\beta_t$ to this node. Next, $\alpha_e$ is calculated and sent to the right child, awaiting the result $\beta_v$ of the right child. Finally $\beta_v$ is calculated and sent to its parent node. Table 1 shows the corresponding equations for the calculation of...
Figure 4. Polar Code Representations.

Table 1. Node operations of different decoding algorithms

<table>
<thead>
<tr>
<th>$i = 0, \ldots, N/2^{l+1}$</th>
<th>SC</th>
<th>SCL</th>
<th>SCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{l_i}$</td>
<td>$f(\alpha_{v_i}, \alpha_{v_i+N/2})$</td>
<td>$L \times f(\alpha_{v_i}, \alpha_{v_i+N/2})$</td>
<td>$f(\alpha_{v_i}, \alpha_{v_i+N/2})$</td>
</tr>
<tr>
<td>$\alpha_{r_i}$</td>
<td>$f(\alpha_{v_i}, \beta_{l_i} \oplus \alpha_{v_i+N/2})$</td>
<td>$L \times f(\alpha_{v_i}, \beta_{l_i} \oplus \alpha_{v_i+N/2})$</td>
<td>$f(\alpha_{v_i}, \beta_{l_i} + \alpha_{v_i+N/2})$</td>
</tr>
<tr>
<td>$\beta_{v_i}$</td>
<td>$\beta_{l_i} \oplus \beta_{r_i}$</td>
<td>$L \times \beta_{l_i} \oplus \beta_{r_i}$</td>
<td>$f(\beta_{l_i}, \beta_{r_i} + \alpha_{v_i+N/2})$</td>
</tr>
<tr>
<td>$\beta_{v_i+N/2}$</td>
<td>$\beta_{r_i}$</td>
<td>$L \times \beta_{r_i}$</td>
<td>$f(\beta_{l_i}, \alpha_{v_i+N/2}) + \beta_{r_i}$</td>
</tr>
</tbody>
</table>

Please note that for SCL decoding algorithm, additional effort has to be spent within the nodes for the list management, in particular the sorting and pruning of the list according to path metrics which are not shown here for simplicity. Also, the leaf operations differ slightly between SC, SCL, and SCAN. While SC and SCL decide for either 1 or 0 in the leaf node, the SCAN decoding algorithm follows the BP approach for Polar Codes and propagates constants, i.e., infinity for frozen bits and zero for information bits.

Table 2. Optimized pipeline stages for different codes

<table>
<thead>
<tr>
<th>Code $N/K$</th>
<th>Node Visits/ Optimized Pipeline Stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>32/16</td>
<td>125 17</td>
</tr>
<tr>
<td>128/64</td>
<td>253 65</td>
</tr>
<tr>
<td>1024/512</td>
<td>4093 385</td>
</tr>
<tr>
<td>1024/768</td>
<td>4093 273</td>
</tr>
</tbody>
</table>

Figure 5. Architectural template for unrolled tree traversals.

4.2 Unrolling the tree traversal

To build high-throughput decoders, pipelining is the most efficient way to gain maximal parallelism and at the same time maximized data locality. For iterative decoding algorithms like the BP, pipelining can be achieved by unrolling the iterations and pipelining the different stages. The number of pipelining stages corresponds to the number of iterations. Such an architecture was for the first time demonstrated for LDPC decoding by Schläfer et al. (2013). In their paper, a min-sum LDPC decoding algorithm was unrolled over 9 iterations for an $N = 672$, Rate = $13/16$ LDPC code, achieving an outstanding throughput and energy efficiency, but at the cost of flexibility.

Here we focus on SC, SCL, and SCAN that perform depth-first traversal of the Polar Factor Tree during decoding. Comparable to unrolled LDPC decoding, we can similarly unroll the tree traversal: Whenever a tree node is visited during the
Pipeline optimization

Since the number of traversed nodes’ visits directly maps into the amount of pipeline stages, minimizing the tree corresponds to minimizing the number of pipeline stages. A smaller tree therefore leads to less pipeline stages, thus lower latency, improved area, and energy efficiency. Sarkis et al. (2014) have proposed some techniques to reduce the size of the Polar Factor Tree without a loss in error correction performance.

A reduction is achieved by first replacing complete subtrees with a single node that can decode the subtree in one step. Two possible subtree optimizations are shown in Fig. 6:

A 4 bit subtree with three frozen and one information bit as leaves, equaling a repetition code, is replaced by a new specialized node (marked red). Similar a 4 bit subtree with only one frozen bit and three information bits as leaves, acting like a simple parity check code, is likewise replaced by a new specialized node (marked green).

The tree can be further reduced by optimizing homogeneous subtrees (composed of either completely frozen or completely information nodes) and merging them into their parent node. This is valid because a subtree with only frozen bit leaves (marked white in Fig. 6) equals a Rate-0 code, that doesn’t carry any information but only noise. Accordingly a subtree with only information bits as leaves (marked blue) equals a Rate-1 code that consists of only information and no redundancy, therefore has no error correction capabilities.

Table 2 shows the number of node visits, or pipeline stages, for selected Polar Codes with different code lengths and code rates, respectively, before and after the optimization for the SCAN decoding algorithm. The numbers for the optimized pipeline stages are created by traversing the reduced tree described in this section. While the improvement in the number of pipeline stages and therefore also latency and area are significant, it also becomes obvious, that every change in the code length or code rate results in a different tree and, thus, a different architecture.

5 Framework for design space exploration

Every change in either the decoding algorithm, the code (the position of frozen bits), its length or rate affects the optimal architecture for a Polar Code decoder and spans a de-
sign space (Fig. 7) of possible architectures. This space is too large for manual investigation and can only be explored by automating the hardware decoder generation.

We developed a framework to explore this design space. Input are a Polar Code, a specific decoding algorithm, and further parameters, e.g., approximations, quantization, etc. Output are simulation results to evaluate the error correction performance and VHDL code of decoder architectures that can be synthesized, placed&routed. This information is fed back into the framework for further optimization. The framework consists of a number of different components that are modular and interact as shown in Fig. 8. Their functionality is as follows:

Node Operations: In this block, the node functionality is selected according to the decoding algorithm. The node functionality is defined on C++ level for simulation and on VHDL level for architecture generation (Table 1).

The Simulation Model generates the error correction performance for the specified decoding algorithms, approximations, quantization etc., and generates FER/Bit Error Rate plots as shown in Fig. 3.

The Polar Factor Tree from Sect. 4 represents the main data structure of the framework and uses a Polar Code as input. The tree is then traversed to generate a Pipelined Architecture (see Sect. 4.2). This pipelined architecture uses the sequence of operations from the tree traversal and the VHDL blocks to assemble a fully synthesizable Polar Code decoder in VHDL.

Both the Polar Factor Tree and the pipelined architecture are subject to Optimization based on results from the simulation model and synthesis/place&route. For the Polar Factor Tree this can be, e.g., optimizations on the tree (see Sect. 4.3). For pipeline optimizations we refer to the next section.

5.1 Case study

To show the performance of our framework we present as a case study, an unrolled SCAN Polar Code decoder with one iteration. We further show various design trade-offs. There are two extreme cases. We can remove all registers from the pipeline stages. This results in a pure combinatorial circuit (Fig. 9a). Advantages are low area, low power, and low latency. Main disadvantage is low throughput. The other extreme is to have a pipeline register between each stage (Fig. 9b). Advantage is high-throughput, disadvantages are high area, high power, and a high latency. Our framework allows to balance the register count between the two extrema using its internal timing engine. So, for a given throughput constraint, the number and optimum position of registers can be automatically calculated (Fig. 9c). Table 3 shows the various trade-offs. Results are after place&rout and worst case Process-Voltage-Temperature (PVT) on a 28 nm FD-SOI technology. Throughput refers to coded bits.

The first two columns show the pure combinatorial and fully pipelined architecture, respectively. Column 3 shows the architecture for a throughput constraint of 100 Gbps. All

<table>
<thead>
<tr>
<th>Technology</th>
<th>Combinatorial Logic</th>
<th>Fully Pipelined</th>
<th>Optimized for 100Gbps</th>
<th>Lin et al. (2017)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>128/64</td>
<td>128/64</td>
<td>128/64</td>
<td>1024/512</td>
</tr>
<tr>
<td>Architecture</td>
<td>Unrolled</td>
<td>Unrolled</td>
<td>Unrolled</td>
<td>Sequential</td>
</tr>
<tr>
<td>Throughput (Gbps)</td>
<td>6.39</td>
<td>121.9</td>
<td>108.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Max. Frequency (MHz)</td>
<td>49.9</td>
<td>952.4</td>
<td>847.5</td>
<td>1440</td>
</tr>
<tr>
<td>Nr. of Reg. Stages</td>
<td>0</td>
<td>64</td>
<td>24</td>
<td>–</td>
</tr>
<tr>
<td>Total Size of Reg. (KB)</td>
<td>0</td>
<td>13.9</td>
<td>5.23</td>
<td>–</td>
</tr>
<tr>
<td>Area (mm²)</td>
<td>0.36</td>
<td>0.76</td>
<td>0.39</td>
<td>0.12</td>
</tr>
<tr>
<td>Area Efficiency (Gbps mm⁻²)</td>
<td>17.7</td>
<td>159.6</td>
<td>276.0</td>
<td>20.5</td>
</tr>
</tbody>
</table>
three architectures were automatically generated with our framework. The last column shows the fastest SCAN decoder implementation published (Lin et al., 2017). However, their decoder uses a sequential processing, a different code and a different technology. To give at least a first estimate we scaled the decoder down to 28 nm. Our architecture is more than 40 times faster and shows a much better area efficiency.

To the best of our knowledge, this fully pipelined architecture is the fastest SCAN Polar Code decoder, while the partially pipelined architecture is the most area efficient published so far.

6 Conclusion and outlook

Based on the works of Sarkis et al. (2014), Giard et al. (2015), and Lin et al. (2017) we proposed a generalized model for the decoding of Polar Codes with the SC, SCL, and SCAN algorithms. This model is used to provide a generic architectural template for unrolled high-throughput decoders. We further presented a framework to evaluate the error correction performance of Polar Codes, automate the unrolling and generate a fully functional and synthesizable VHDL implementation of a Polar Code decoder.

Further, we presented an optimization of pipeline stages to improve the area and energy efficiency. To demonstrate the performance of our framework we presented a high throughput SCAN decoder. With a throughput of over 100 Gbps it is to our knowledge the fastest published SCAN decoder.

Data availability. All data are represented in the text and functions. The raw data is not meaningfully interpretable without the use of special simulation/synthesis tools which are not publicly available.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue “Kleinhübcher Berichte 2017”. It is a result of the Kleinhübcher Tagung 2017, Miltenberg, Germany, 25–27 September 2017.

Acknowledgements. We gratefully acknowledge financial support by the DFG (project-ID: 2442/8-1) and the EU (project-ID: 760150-EPIC).

Edited by: Jens Anders
Reviewed by: two anonymous referees

References


