HARDWARE IMPLEMENTATION OF ITERATIVE PROJECTION-AGGREGATION DECODING OF REED-MULLER CODES

Marzieh Hashemipour-Nazari, Kees Goossens, Alexios Balatsoukas-Stimming

Eindhoven University of Technology

ABSTRACT
In this work, we present a simplification and a corresponding hardware architecture for hard-decision recursive projection-aggregation (RPA) decoding of Reed-Muller (RM) codes. In particular, we transform the recursive structure of RPA decoding into a simpler and iterative structure with minimal error-correction degradation. Our simulation results for RM(7, 3) show that the proposed simplification has a small error-correcting performance degradation (0.005 in terms of channel crossover probability) while reducing the average number of computations by up to 40%. In addition, we describe the first fully parallel hardware architecture for simplified RPA decoding. We present FPGA implementation results for an RM(6, 3) code on a Xilinx Virtex-7 FPGA showing that our proposed architecture achieves a throughput of 171 Mbps at a frequency of 80 MHz.

1. INTRODUCTION
Reed-Muller (RM) codes were first proposed in 1954 [1]. Recently, there has been a renewed interest in RM codes because, in some cases, they were shown to be capable of achieving the Shannon capacity of the binary erasure channel (BEC) [2] and the binary symmetric channel (BSC) [3]. The oldest decoding algorithm for RM codes is based on majority voting [1], and it guarantees correction of the error patterns with a weight less than half of the minimum distance. A wide variety of algorithms has been proposed afterward to improve decoding capacity. For example, the Sidel’nikov-Pershakov algorithm [4] corrects most of the corrupted codewords with a number of errors less than \((1 - \varepsilon)n/2\), where \(n\) indicates blocklength of the RM codes and \(\varepsilon \geq n^{-1/3}\). Some hardware architectures are also available for the aforementioned methods. A parallel decoding architecture for majority-logic decoding algorithms was provided in [5], and a low-area decoder for the Reed decoding method was introduced in [6].

Successive-cancellation (SC) decoding [7] and SC list (SCL) decoding [8], make use of the decomposable structure of RM codes to provide recursive decoding methods with reasonable complexity. The work of [9] improved the performance of SC and SCL decoding methods by exploring several carefully selected permutations of the factor graph of RM codes. The work of [10] exploited the symmetric structure of RM codes and applies an iterative decoding method to provide near maximum likelihood (ML) performance. Other works focused on special cases of RM codes. For example, [11] is a modified version of the Sidel’nikov-Pershakov algorithm that improves error-correcting performance for second-order RM codes. Moreover, the work of [12] provided a new ML decoder with a lower complexity for RM codes of order \(m = 3\), where \(m = \log_2 n\).

The main drawback of the aforementioned algorithms is that they have poor error-correcting performance for short blocklength RM codes. For this reason, the authors of [13] proposed a new algorithm called recursive projection-aggregation (RPA) decoding that improves the error-correcting performance of RM codes in the regimes of interest of ultra-reliable low-latency communications (URLLC) and of the Internet of Things (IoT), i.e., low rate and short blocklength RM codes. The RPA algorithm is highly parallelizable. However, it has a high complexity and its recursive structure is not particularly amenable to hardware implementations. The authors of [14] proposed a collapsed projection-aggregation (CPA) decoding algorithm, which merges multiple recursion levels into a single step. However, this comes at the cost of an increased complexity for the projection step, since more than two bits (or LLRs) are combined at each step. The authors of [14] also introduce recursive puncturing-aggregation (RXA), which is more suitable for high-rate RM codes.

 Contributions: In this paper, we present a simplified version of the RPA algorithm to make a trade-off between the error-correcting performance and computations. We simplify the RPA algorithm by carefully removing computations in the recursion levels to make the structure suitable for hardware implementations. Moreover, we propose the first fully parallel hardware architecture for RPA decoding.

2. REED-MULLER CODES
The focus of this paper is on the BSC, so all operations and vectors are in \(\mathbb{F}_2\). RM codes are denoted by \(RM(m, r)\), where \(m\) indicates the code length \(n = 2^m\) and \(r\) is the order. RM codes are linear block codes with rate \(R = \frac{1}{n}, k = \sum_{r=0}^{m} \binom{m}{r}\), and with the following recursively defined generator matrix:

\[
G_{(m,r)} = \begin{bmatrix} G_{(m-1,r)} & G_{(m-1,r-1)} \\ 0 & G_{(m-1,r-1)} \end{bmatrix}, \quad G_{(1,1)} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.
\]
Algorithm 1 The RPA decoding of RM codes RM(m, r)

Input: The noisy codeword y, m, r, N_{max}
Output: The decoded codeword c
1: if r = 1 do
2: c ← order-1-decoding (y, m)
3: else
4: for j = 1 : N_{max} do
5: for i = 1 : 2^{m} - 1 do
6: \hat{y}_i ← \text{Proj}(y, i, m)
7: end for
8: end for
9: \hat{y} ← \text{Agg}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{2^m - 1}, \hat{y}_3, \hat{y}_4, \ldots, \hat{y}_{N_{max}})
10: if y = \hat{y} do
11: break -- RPA converges to a fixed point
12: end if
13: y ← \hat{y}
14: end if
15: end if

Algorithm 2 The projection function Proj

Input: \text{yn}(0 to n-1), i, m
Output: \text{yout}(0 to n/2 - 1)
1: n ← 2^m
2: if i < n/2 do
3: y_{out}(0 to n/4 - 1) ← \text{Proj}(y_{in}(0 to n/2 - 1), m, \text{Ind})
4: y_{out}(n/4 to n/2 - 1) ← \text{Proj}(y_{in}(n/2 to n - 1), m, \text{Ind})
5: else
6: for j = 1 : n/2 - 1 do
7: y_{mp}(2j) ← y_{in}(j)
8: y_{mp}(2j+1) ← y_{in}(j + i)
9: end for
10: y_{mp}(0) ← y_{in}(0)
11: y_{mp}(1) ← y_{in}(i)
12: for t = 0 : n/2 - 1 do
13: y_{out}(t) ← y_{mp}(2t) \oplus y_{mp}(2t + 1)
14: end for
15: end if

Algorithm 3 The aggregation function Agg

Input: m, y_m, y_1, y_2, \ldots, y_{n-1}, y_1, y_2, \ldots, y_{n-1}
Output: y_{out}
1: for z = 0 : 2^m - 1 do
2: vote(z) ← 0
3: for i = 1 : 2^m - 1 do
4: \text{Ind} ← \text{FindIndex}(z, i, m)
5: vote(z) ← vote(z) \oplus z_{(\text{Ind})} \oplus \hat{y}_{(\text{Ind})} + vote(z)
6: end for
7: y_{out}(z) ← y_{in}(z) \oplus 1 \left[\text{vote}(z) > 2^{m-1}\right]
8: end for

Algorithm 4 The function FindIndex

Input: Index z, branch number i, m
Output: \text{Ind}
1: if i \geq 2^{m-1} do
2: if z < 2^{m-1} do
3: \text{Ind} ← z
4: else
5: \text{Ind} ← b2d2e(\text{dc2b}(z) \oplus \text{dc2b}(i))
6: end if
7: else
8: if z < 2^{m-1} do
9: \text{Ind} ← \text{FindIndex}(z, i, m - 1)
10: else
11: \text{Ind} ← \text{FindIndex}(z - (2^{m-1}), i, m - 1) + 2^{m-2}
12: end if
13: end if

2.1. Recursive Projection Aggregation Decoding

As Algorithm 1 shows, the RPA algorithm has three main steps: projection (line 6), recursive decoding (line 7), and aggregation (line 9). Let us consider a noisy received vector y of the transmitted codeword c of length n.

In the projection step, y is transformed into n - 1 distinct vectors of length n/2. For hard-decision decoding, each transformed vector y_i, i \in \{1, 2, \ldots, n-1\}, is obtained by taking the modulo-2 sum over specific coordinates of the input vector corresponding to the i-th projection, as shown on lines 2-11 of Algorithm 2. Next, on line 13, a binary XOR operation sums every two adjacent bits to convert each n-bit input vector y to an n/2-bit vector y_i.

In the recursive decoding step, each vector y_i, produced in the projection step, is recursively decoded by RPA for RM(m-1, r-1) until first-order RM codes are reached, which can be decoded efficiently using the fast Hadamard transform (FHT) [15]. Each y_i is a decoded vector of y_i.

In the aggregation step, for each coordinate \hat{y}(z), Algorithm 4 finds the corresponding coordinates in y_i that were originally created with y(z). These coordinates together with their decoded value in y_i represent n - 1 estimations for each coordinate of vector \hat{y} (see line 3-6 in Algorithm 3). Next, in line 7 of Algorithm 3, per-coordinate majority voting is performed to produce an estimate \hat{y} of the transmitted codeword.

This procedure is repeated for multiple iterations. In [13], the maximum number of iterations is set to N_{max} = \lfloor m/2 \rfloor.

3. ITERATIVE PROJECTION-AGGREGATION DECODING

As can be seen on line 4 of Algorithm 1, RPA decoding performs multiple iterations at each level of the recursion. After each aggregation, if \hat{y} \neq y, y will be updated by \hat{y}, and the whole procedure from projection to aggregation will iterate again. Unfortunately, having iterations on each recursion level makes RPA unsuitable for hardware implementations, as it requires complicated control circuitry and memory structures.

In the case of a noisy received vector y with only one error, for all projected vectors at each recursion level, it can be verified from Algorithm 2 that there exists exactly one error for every level of the recursion and for all projections, which is corrected in level r = 1 because FHT decoding guarantees the correction of one error. However, the condition for skipping the remaining iterations is not satisfied (see line 10 in Algorithm 1), and as a result, RPA runs another iteration at this level. This additional iteration is unnecessary because it performs projection, first-order decoding, and aggregation on the already corrected codewords. More generally, and motivated by the above example, if the iteration loops for a recursion level run more than once but stop before reaching
Algorithm 5: The IPA decoding of \( \text{RM}(m, r) \) codes

**Input:** The noisy codeword \( y \), \( m \), \( r \), \( N_{\text{max}} \)

**Output:** The decoded codeword \( e \)

```
1: \( y(1,0) \leftarrow y; \quad n_{\text{tmp}} \leftarrow 1 \\
2: \text{for } j = 1 : N_{\text{max}} \text{ do } \\
3: \quad \text{for } l = 1 : r - 1 \text{ do } --\text{Projection loop} \\
4: \quad \quad n_{\text{tmp}} \leftarrow n_{\text{tmp}} \times (2^m - 1) \\
5: \quad \quad \text{for } i = 1 : n_{\text{tmp}} \text{ do } \\
6: \quad \quad \quad y_{\text{tmp}} \leftarrow y_{\lfloor i/(2^m-1) \rfloor} \\
7: \quad \quad \text{end for} \\
8: \quad \text{end for} \\
9: \quad m \leftarrow m - 1 \\
10: \text{end for} \\
11: \text{for } l : 1 : n_{\text{tmp}} \text{ do } --\text{First-order decoding} \\
12: \quad \hat{y}_{(l,r-1)} \leftarrow \text{order-1-decoding}(y_{(l,r-1)}, m) \\
13: \text{end for} \\
14: \text{for } l = r - 2 : 0 \text{ do } --\text{Aggregation loop} \\
15: \quad m \leftarrow m + 1; \quad t \leftarrow 2^m - 1; \quad n_{\text{tmp}} \leftarrow n_{\text{tmp}}/t \\
16: \quad \text{for } i = 1 : n_{\text{tmp}} \text{ do } \\
17: \quad \quad d \leftarrow (i - 1) \times t \\
18: \quad \quad y_{(i,l)} \leftarrow \text{Agg}(y_{(i,l)}, y_{(d+1,l+1)}, \ldots, y_{(d+t,l+1)}) \\
19: \quad \text{end for} \\
20: \text{end for} \\
21: \text{if } y_{(1,0)} = \hat{y}_{(1,0)} \text{ break } --\text{IPA converges to a fixed point} \\
22: \text{end if} \\
23: y_{(1,0)} \leftarrow \hat{y}_{(1,0)} \\
24: \text{end for} \\
25: e \leftarrow \hat{y}_{(0,1)}
```

\( N_{\text{max}} \), the last iteration always runs only to check the stop condition. We call these iterations ineffective.

Based on our simulations of various RM codes, at low channel crossover probabilities, more than 50% of internal iterations are ineffective. Motivated by this observation, we present a simplification of RPA by removing iterations on the internal levels of the RPA recursion. Effectively, our proposed iterative projection-aggregation (IPA) algorithm sets \( N_{\text{max}} = 1 \) for all recursive decoding steps of the RPA algorithm except for the first one. This can be concluded by comparing Algorithm 5, in which the iterative structure of IPA is shown, with Algorithm 1. The difference is that there exists only one iteration loop around the entire function in Algorithm 5. Moreover, the iterative structure of Algorithm 5 is more convenient for a hardware implementation. As we show in Section 5, this reduces complexity and hardware implementation significantly, with a small error-correcting penalty.

It can be shown that the complexity of RPA decoding with internal iterations is \( O(n^r(\log_2 n)^{r+1}) \). For the IPA algorithm, the complexity is \( O(n^r(\log_2 n)^2) \) as we remove the internal iterations. We also show in the Section 5 that the overall calls to the first-order decoder, which is a more practical complexity measure, are decreased significantly.

4. PROPOSED HARDWARE ARCHITECTURE

Our proposed fully parallel IPA architecture, which is shown in Fig. 1, consists of three main components and a control unit.

The first component is the projection, including \( r - 1 \) levels of the projection for \( \text{RM}(m, r) \) codes (line 3 of Algorithm 5).

The second component, which we call the first-order decoder, has parallel decoders for all \( \text{RM}(m - r + 1, 1) \) codes generated in the innermost level of the RPA (line 11 of Algorithm 5).

The third component is the aggregation unit performing \( r - 1 \) levels of aggregation (line 14 of Algorithm 5).

The projection component performs \( r - 1 \) levels of projection, as described in Section 2.1. Each projection level has parallel projection units, each consisting of a re-ordering unit (ROU) and an XOR unit. The re-ordering unit \( \text{ROU}(m,i) \) finds the coordinates for \( i \)-th projection of the input vector \( y \) with length of \( 2^m \) based on lines 2-11 of Algorithm 2. Then, an XOR unit is assigned to each projection branch for performing the sum operations as described in lines 12-13 of Algorithm 2.

The first-order decoder component provides first-order decoders (FODs) for all \( \text{RM}(m - r + 1, 1) \) codes, obtained in the innermost level of projection, in parallel. Each FOD was designed based on the decoding method proposed in [15], and consists of three sub-units: FHT, Argmax, and Generator matrix. The first unit gives the vector \( 1 \), which is the result of the FHT on a binary input vector \( y \):

\[
i = (1 - 2y)H_{2^m},
\]

where the Hadamard matrix \( H_{2^m} \) is

\[
H_{2^m} = \begin{bmatrix} H_{2^{m-1}} & H_{2^{m-1}} \\ H_{2^{m-1}} & -H_{2^{m-1}} \end{bmatrix} \text{ and } H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.
\]

The architecture of FHT unit is derived from [16]. The Argmax unit finds the index \( z \) of the maximum value of \( 1 \). The output of the FOD unit is:

\[
\hat{y} = \bar{x}G_{(m,1)},
\]

where \( \bar{x} = \left[ \frac{1 - \text{sign}(1(z))}{2} \right] z_{\text{bin}} \) with \( z_{\text{bin}} \) being the binary representation of \( z \), and where \( G_{(m,1)} \) is the generator matrix of \( \text{RM}(m, 1) \).

The aggregation component provides \( r - 1 \) levels of aggregation, each of which has \( \prod_{i=0}^{j} \binom{2^m - i - 1}{m - r + 1 + i} \) AGG units in parallel (line 16 of Algorithm 5), where \( j \) denotes the current level of aggregation. As Fig. 1 shows, each AGG unit consists of \( 2^m - 1 \) RRUMs (the hardware implementation of Algorithm 4) and one majority voter to aggregate into a \( n \)-bit codeword \( \hat{y} \) calculated in line 7 of Algorithm 3. Finally, XOR gates are used to flip the desired bits of input vector \( y \) in \( \hat{y} \) as described in line 7 of Algorithm 3.

The throughput of the decoder is calculated by:

\[
\text{Throughput} = \frac{\text{Frequency}}{N_{\text{iter}}N_{\text{cycles/iter}}} \times n,
\]

where \( N_{\text{iter}} = N_{\text{max}} \) for the minimum throughput and \( N_{\text{iter}} = N_{\text{avg}} \) is the average number of iterations for the average throughput. The data path is pipelined to \( N_{\text{cycles/iter}} \) stages.
In particular, \( r - 1 \) and \( 2(r - 1) \) registers are inserted between the projection and aggregation levels, respectively. Additionally, three registers are inserted between the components of the FODs, and one register is used to check the termination condition. As such, we have \( N_{\text{cycles/iter}} = 3(r - 1) + 4 \).

5. RESULTS

Simulation results for IPA decoding and RPA decoding for the RM(6, 3) and RM(7, 3) codes over the BSC channel are shown in Fig. 2. We observe that IPA decoding has exactly the same frame error rate (FER) as RPA decoding for RM(6, 3), while there is a minimal error-correcting performance degradation up to 0.005 in terms of channel cross-over probability for RM(7, 3). We also increased \( N_{\text{max}} \) to see if this compensates the performance degradation of IPA, but we observed that it unfortunately does not help.

Fig. 3 shows the average number of first-order decodings for IPA and RPA decoding for RM(6, 3) and RM(7, 3) codes over the BSC channel. We observe that the number of the first-order decodings is decreased by up to 40% for RM(6, 3) and up to 50% for RM(7, 3).

We provide post-PAR results of our IPA decoder architecture for RM(6, 3) on a Xilinx Virtex-7 FPGA with a frequency of 80 MHz in Table 1. The resource utilization is high due to the fully parallel nature of the decoder, but the achieved decoding throughput is also relatively high. As there are no other implementations of RPA in the literature, we cannot perform a direct comparison.

![Figure 1: An example of the proposed hardware architecture for IPA decoding for RM\((m, 3)\) codes.](image)

![Figure 2: Frame error rate Comparison between RPA and IPA for RM(6, 3) and RM(7, 3) codes over BSC channels.](image)

![Figure 3: Comparison of the number of the first-order decoding between RPA and IPA for RM(6, 3) and RM(7, 3) codes.](image)

**Table 1.** Post-PAR results for an RM(6, 3) code on a Xilinx Virtex-7 FPGA (xc7vx1140T).

<table>
<thead>
<tr>
<th>Resource</th>
<th>RM(6, 3)</th>
<th>RM(7, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUTs</td>
<td>602,111/712,000(84.57%)</td>
<td>65,699/1,424,000(4.6%)</td>
</tr>
<tr>
<td>Flip-flops</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clock frequency</td>
<td>80 MHz</td>
<td></td>
</tr>
<tr>
<td>Min. throughput ((N_{\text{max}} = 3))</td>
<td>171 Mbps</td>
<td>333 Mbps</td>
</tr>
<tr>
<td>Avg. throughput @ FER=(10^{-3})</td>
<td>284 Mbps</td>
<td></td>
</tr>
</tbody>
</table>
6. REFERENCES


