Abstract—Recently, a parallel decoding algorithm of $G_N$-coset codes was proposed. The algorithm exploits two equivalent decoding graphs. For each graph, the inner code part, which consists of independent component codes, is decoded in parallel. The extrinsic information of the code bits is obtained and iteratively exchanged between the graphs until convergence. This algorithm enjoys a higher decoding parallelism than the previous successive cancellation algorithms, due to the avoidance of serial outer code processing. In this work, we present a hardware implementation of the parallel decoding algorithm, it can support maximum $N = 16384$. We complete the decoder’s physical layout in TSMC 16nm process and the size is 999.936µm × 999.936µm, ≈ 1.00mm². The decoder’s area efficiency and power consumption are evaluated for the cases of $N = 16384$, $K = 13225$ and $N = 16384$, $K = 14161$. Scaled to 7nm process, the decoder’s throughput is higher than 477Gbps/mm² and 533Gbps/mm² with five iterations.

I. INTRODUCTION

A. Background

High throughput is one of the primary targets for the evolution of mobile communications. The next generation of mobile communication, i.e., 6G, is expected to supply 1 Tbps/mm² throughput [1], which requires roughly a 100x increase in throughput over the 5G standards.

$G_N$-coset codes, defined by Arıkan in [2], are a class of linear block codes with the generator matrix $G_N$. $G_N$ is an $N \times N$ binary matrix defined as $G_N \triangleq \mathcal{F}_N^\otimes n$, in which $N = 2^n$ and $\mathcal{F}_N^\otimes n$ denotes the $n$-th Kronecker power of $\mathcal{F} = [\begin{smallmatrix}1 & 0 \\ 0 & 1 \end{smallmatrix}]$.

Polar codes is a specific type of $G_N$-coset codes, adopted for the 5G control channel, respectively. The throughput of polar codes is limited by the successive cancellation (SC) decoders, since they are serial in nature.

Recently, a parallel decoding framework of $G_N$-coset codes is proposed in [3]. It is alternately decoded on two factor graphs $G$ and $G_\pi$, as shown in Fig. 1. The permuted graph is generated by swapping the inner codes and outer codes. The decoder only decodes the inner codes of each graph $\Lambda \in \{ G, G_\pi \}$. In each $\Lambda$, the inner codes are $\sqrt{N}$ independent sub-codes that can be decoded in parallel. The code construction under the parallel decoding algorithm is different from polar/RM codes, and is studied separately in [3].

B. Motivations and Contributions

This paper introduce an ASIC implementation based on the parallel decoding framework (PDF). We set up a decoder which can support $N = 16384$ $G_N$-coset codes. It deploys $\sqrt{N} = 128$ sub-decoder to decode the 128 independent sub-codes in parallel. The target of high throughput and area efficiency is decomposed into the reduction of sub-codes decoding latency, worst-case iteration time, and chip area, and optimized respectively.

We adopt the proposal in [4] which employs successive cancellation (SC) decoders as the component decoder. It can support soft-in-hard-out decoding which results in low decoding complexity and reduced interconnection among the component decoders. In this work, we propose hardware-oriented optimizations on LLR generation and quantization. We implemented the whole decoder in hardware and present the ASIC layout to evaluate multiple key metrics. The hardware-specific data is obtained from the cells flip ratio from the circuit simulation results and the parasitic capacitance extracted from the layout result. With 16nm process, the area efficiency is 120Gbps/mm², the power consumption is 100mW and energy efficiency is around 1pJ/bit. Scaled to 7nm, the area efficiency can reach 533Gbps/mm² with five iterations.

II. PARALLEL DECODING

A parallel decoding framework is introduced in [3], where three types of component decoders (i) soft-output SC list (SCL), (ii) soft-output SC permutation list and (iii) soft cancellation (SCAN) are employed to decode the sub-codes (inner codes). To achieve even higher area efficiency, this work adopts SC [4], i.e., without list decoding, as the sub-decoder. In this section, we describe the parallel decoding framework-successive cancellation (PDF-SC) algorithm from the implementation point of view.

A. Parallel decoding framework (PDF)

We use $(i, j)$ to denote the $j$-th code bit position of $i$-th inner sub-code, and $k$ to denote the code bit position in the $G_N$-coset code. They have the following relationship

$$k = \begin{cases} j \times \sqrt{N} + i, & \text{for } \Lambda = G, \\ i \times \sqrt{N} + j, & \text{for } \Lambda = G_\pi. \end{cases}$$

(1)
The aforementioned parallel turbo-like decoding framework is described in Algorithm 1. In every two iterations, the algorithm **alternately** decodes the two graphs $G$ and $G_{\pi}$ (line 4 in Algorithm 1) with $\sqrt{N}$ inner component decoders. The $i$-th component decoder, denoted by $SubDecoder()$, is a SC decoder assisted by the error detector.

Note that each component decoder takes a soft input vector $L_{t}^{i}$, but outputs hard code bit estimates $\hat{c}$ and error detecting results $e$. The mismatch between soft input and hard output poses a challenge for iterative decoding, since the SC component decoders in the next iteration cannot directly take the hard output from the previous iteration as input. To solve this problem, [4] proposes to generate soft values from the hard outputs. Specifically, the log likelihood ratio (LLR) of the $(i, j)$-th code bit in the $t$-th iteration, denoted by $L_{t}^{i,j}$, is calculated from the hard decoder outputs from the alternate graph (line 5).

For the $i$-th component decoder, the hard output vector and soft input vector have length $\sqrt{N}$

$$
L_{t}^{i} \triangleq \{ L_{t}^{i,j}, j = 0 \cdots \sqrt{N} - 1 \},$

$$
e_{t}^{i} \triangleq \{ e_{t}^{i,j}, j = 0 \cdots \sqrt{N} - 1 \},$

and the error detection indicator $e_{t}^{i}$ is a binary value.

The $\sqrt{N}$ independent inner sub-codes allow us to instantiate $\sqrt{N}$ component decoders for maximum degree of parallelism. After $t_{\text{max}}$ iterations, the algorithm outputs all $N$ hard bits.

### B. Component decoder: $SubDecoder()$

The component decoder $SubDecoder()$ [4] is described in Algorithm 2. Before each SC decoding, error detection is performed. This can be achieved by applying a syndrome check based on hard decisions (line 2~5 in Algorithm 2). The cases with detected errors are denoted by Type-1 and otherwise Type-2.

The error detector brings two-fold advantages. On the one hand, since Type-2 component codes require no further SC decoding, this approach reduces power consumption. If all $\sqrt{N}$ sub-codes pass error detection, decoding can be early terminated for further power saving. On the other hand, the error detection result provides us with additional information that the input LLRs of Type-2 component codes are more reliable than those of Type-1. Such information can be used to improve the overall performance by estimating the input LLRs from the hard outputs.

### Algorithm 1 Parallel decoding framework.

**Require:**

The received signal $y = \{ y_{k}, k = 0 \cdots N - 1 \}$;

**Ensure:**

The recovered codeword: $\hat{x} = \{ \hat{x}_{k}, k = 0 \cdots N - 1 \}$;

1: Initialize: $L_{ch,k} \triangleq 2 \frac{y_{k}}{\sigma^{2}}$ for $k = 0 \cdots N - 1$;
2: for iterations: $t = 1 \cdots t_{\text{max}}$ do
3: Select decoding graph: $\Lambda = \{ t\%2 \}? G : G_{\pi}$;
4: for inner component decoders: $i = 0,1 \cdots \sqrt{N} - 1$ (in parallel) do
5: $L_{t}^{i,j} = \text{Lgen}(L_{ch,k}, e_{t-1}^{i,j}, e_{t-1}^{i-2,j}, \gamma_{t,j})$, $\forall j$;
6: $e_{t}^{i,j} = \text{SubDecoder}(L_{t}^{i,j}, F_{A,i})$;
7: end for
8: end for
9: for $\forall (i,j)$ do
10: $\hat{x}_{k} = e_{t}^{i} \parallel j$; $i \notin k$ is described in [1].
11: end for

### Algorithm 2 The $i$-th $SubDecoder()$

**Require:**

The input LLRs: $L_{t}^{i}$;

Frozen set: $F_{A,i}$.

**Ensure:**

Binary output $e_{t}^{i,j}$, error detected indicator $e_{t}^{i}$;

1: $e_{t}^{i,j} = False$;

2: Hard decisions:

3: $e_{t}^{i,j} = 1$, $j \in \{ 0,1, \cdots , \sqrt{N} - 1 \}$;

4: Vector $u_{t}^{i} = e_{t}^{i} \times G_{\pi}$;

5: Syndrome check:

6: if $u_{t}^{i} \neq 0$, $\forall j \in F_{A,i}$ then
7: $e_{t}^{i} = True$;
8: SC decoding: $\hat{c}_{t}^{i} = \text{SCDecoder}(L_{t}^{i}, F_{A,i})$;
9: end if

### C. Input LLR generator: $\text{Lgen()}$

In each iteration, the input LLRs are calculated by

- **Type-1:** $e_{t-1}^{i,j} = 1$

$$
L_{t}^{i,j} = L_{ch,k} + 2\frac{e_{t-1}^{i,j}}{\sigma^{2}}(1 - 2e_{t-2}^{i,j}) - 2\frac{e_{t-1}^{i,j}}{\sigma^{2}}(1 - 2e_{t}^{i,j}),
$$

- **Type-2:** $e_{t-1}^{i,j} = 0$

$$
L_{t}^{i,j} = L_{ch,k} + 2\frac{e_{t,j}}{\sigma^{2}}(1 - 2e_{t}^{i,j}),
$$

where a set of damping factors ($\alpha^{t}$, $\beta^{t}$, $\gamma^{t}$) are defined for each iteration, and $k$ is calculated from $(i,j)$ according to [1].

The specific values of damping factors are optimized in [4].
Since SC decoder is invariant to input LLR scaling, we can cancel noise variance $\sigma^2$ during LLR initialization. By multiplying both sides of (2) and (3) by $\frac{\sigma^2}{2}$, the equations are simplified as follows

$$\left\{ \begin{array}{l}
\hat{L}_{i,j} = y_k + \alpha^t (1 - 2^{l-1}_{(j,i)}) - \beta^t (1 - 2^{l-2}_{(j,i)}) \ ,
\hat{e}^t_{j} = 1, \\
\hat{L}_{i,j} = y_k + \gamma^t (1 - 2^{l-1}_{(j,i)}),
\hat{e}^{t-1}_{j} = 0,
\end{array} \right.$$  \hspace{1cm} (4)

where $\hat{L} \triangleq \frac{L \times \sigma^2}{2}$ and $y$ is the received signal.

We use a pair of new coefficients to replace $\alpha$ and $\beta$:

$$\left\{ \begin{array}{l}
\delta^t = \alpha^t + \beta^t, \\
\theta^t = \alpha^t - \beta^t.
\end{array} \right.$$  \hspace{1cm} (5)

Hence (4) can be replace by (5).

$$\hat{L}_{i,j} = y_k + \Delta^t_{i,j} (1 - 2^{l-1}_{(j,i)}),$$

in which $\Delta^t_{i,j}$ is determined by the binary $e^{t-1}_{j}$ and the hard outputs of the previous two iterations $c^{t-1}_{j,i}, c^{t-2}_{j,i}$ as in (6). The input signal calculation reduces to only one addition operation.

$$\Delta^t_{i,j} = \left\{ \begin{array}{l}
\delta^t, \text{where } e^{t-1}_{j} = 1 \text{ and } e^{t-1}_{j} \neq e^{t-2}_{j}, \\
\theta^t, \text{where } e^{t-1}_{j} = 1 \text{ and } e^{t-1}_{j} = e^{t-2}_{j}, \\
\gamma^t, \text{where } e^{t-1}_{j} = 0.
\end{array} \right.$$  \hspace{1cm} (6)

III. AN ASIC IMPLEMENTATION

In this section, we present the ASIC implementation of a PDF-SC decoder in TSMC 16nm process for $N = 16384$. The hardware optimization addresses both the SC decoders for component codes and the overall parallel decoding framework for $G_N$-coset codes.

A. Bit Quantization

Lower precision quantization is the key to higher throughput, thanks to its reduced implementation area and increased clock frequency. As a tradeoff, performance loss is expected. To maximize throughput while retaining performance, we must determine an appropriate quantization width. Specifically, we use simulation to find out the smallest quantization width of a fixed-point decoder within 0.1dB performance loss from a floating decoder.

First, we compare the performance of component codes under Algorithm2. We test two cases $N_{\text{sub}} = 128, K_{\text{sub}} = 115$ and $N_{\text{sub}} = 128, K_{\text{sub}} = 119$ with different quantization widths. According to Fig 2, 6-bit quantization achieves the same performance as floating-point, 5-bit quantization incurs < 0.1dB loss, and 4-bits quantization yields significant loss. Therefore, we set the quantization width to 5 or 6 bits.

We then simulate the BLER performance of the long codes $N = 128^2, K = 115^2$ and $N = 128^2, K = 119^2$ under different quantizations, as shown in Fig. 3. Similarly, 6-bit quantization has no performance loss, and 5-bit quantization only incurs < 0.1dB loss for both code rates. Again, 4-bit quantization suffers from performance degradation.

| TABLE I | SUB-DECODER DECODING LATENCY FOR N=128 POLAR CODES |
|----------|------------------------------------------|
| Information Bits(K) | 111 | 115 | 119 | 122 |
| Latency (Cycle) | 24 | 19 | 18 | 13 |
| (ns) | 22.8 | 18.05 | 17.1 | 12.35 |

Fig. 2. Sub-Decoder Performance Comparison between Floating Point and Fixed Point.

Fig. 3. Decoder Performance Comparison between Floating Point and Fixed Point.

B. SC Core Optimization

A component SC decoder is optimized via Rate-0 nodes, Rate-1 nodes [5], single parity check (SPC) nodes and repetition nodes (REP) [6]. The decoder skips all Rate-0 nodes, parallelizes Rate-1, SPC and REP nodes for code blocks shorter than 32. If neither applies, maximum-likelihood (ML) multi-bit decision [7] is employed for 4-bit blocks.

The architecture of SC decoders used here is described in [8], with supported code length reduced from 32768 to 128 to save area. With TSMC 16nm technology, an SC core synthesizes to 4,100$\mu$m$^2$ area. Under 1.05GHz clock frequency, its latency is shown in Table I.

C. SC Core Sharing

A unique design that significantly reduces area is called “SC core sharing”. In particular, we bind four SC cores as a subdecoder group. The four SC cores share input/output pins, LLR updating circuits and error detector related components. The sharing reduces a lot of local computation resources and global
TABLE II
DECORDER AREA EFFICIENCY

| Info Size | Iteration | Es/N0 (dB) | Latency (ns) | Area Eff. (Gbps/mm²) | Convert to 10nm | Convert to 7nm |
|-----------|-----------|------------|--------------|----------------------|----------------|----------------|
| 13225     | 4         | 7.14       | 91.2         | 135.07               | 310.66         | 596.47         |
|           | 5         | 6.82       | 114          | 108.06               | 248.53         | 477.17         |
|           | 6         | 6.55       | 136.8        | 90.05                | 207.11         | 397.64         |
|           | 7         | 6.36       | 159.6        | 77.18                | 177.52         | 340.84         |
|           | 8         | 6.20       | 182.4        | 67.53                | 155.33         | 298.23         |
| 14161     | 4         | 7.79       | 87.4         | 150.92               | 347.11         | 666.45         |
|           | 5         | 7.48       | 109.25       | 120.73               | 277.69         | 533.16         |
|           | 6         | 7.22       | 131.1        | 100.61               | 231.41         | 444.30         |
|           | 7         | 7.06       | 152.95       | 86.24                | 198.35         | 380.83         |
|           | 8         | 6.97       | 174.8        | 75.46                | 173.55         | 322.22         |

D. Global Layout

Combining all algorithmic and hardware optimizations, the synthesized decoder ASIC requires 999.936µm × 999.936µm ≈ 1.00mm² area. The global layout is presented in Fig. 5. In the center of the layout is the top logic, including the input channel LLR storage, finite state machine (FSM) controller, interleaved connection routing and output buffer. The aforementioned 32 sub-decoder groups (SG in the figure) are placed around the top logic, highlighted by different colors.

Fig. 4. The Sub-Decoders Group architecture.

Fig. 5. The global layout of ASIC for parallel decoding.

IV. KEY PERFORMANCE INDICATORS

The key performance indicators (KPIs) are examined. First of all, we evaluate the area efficiency using equation $Area\ Efficiency = \frac{Latency (ns)}{Area (mm^2)}$. The proposed decoder can reach up to hundreds of gigabits per second within one square millimeter. The evaluated throughput in given in Table I[16]. With TSMC 16nm process, the area efficiency for code rate 13225/16384 and 14161/16384 are 67Gbps/mm² and 75Gbps/mm² when the maximum number of iterations is eight. If we reduce to five iterations by allowing 0.5db performance loss, the area efficiency can reach 108Gbps/mm² and 120Gbps/mm². The estimated throughput under 10nm and 7nm technologies can be converted from 16nm[16]. With the more advanced 7nm process, the throughput is as high as 477Gbps/mm² and 533Gbps/mm², which are much higher than the 100Gbps/mm²@7nm target given in the EPIC project [12]. Note that the KPI is achieved at code length 16384, which exhibits significant coding gain over codes with length 1024 [4]. With future technologies of 5nm and below, it is promising to achieve an extremely challenging target of 1Tbps/mm².

The area efficiency is also compared with a highly-optimized and fabricated ASIC polar decoder in [8]. For both code rates, the throughput of the proposed decoder (with five iterations) is nine times as fast as a polar fast-SC decoder, and 53 times that of a CA-SC-List-8 decoder. Detailed comparison with other current state-of-the-art decoders is presented in Table I[16].

1. The error detector can early terminate the decoding, but its worst-case latency is guaranteed by maximum iteration times, as required by most practical communication systems.

2. The third column “Es/N0” is chosen such that BLER = 10^{-4}.

3. The converting ratios including cell density ratio and speed improvement ratio are obtained from the TSMC process introductions [10], [11].

4. The polar codes are constructed by Gaussian approximation at Es/N0=5.3dB, 6.3dB, 7.3dB and 8.0dB for (N,R)=(16384,0.807), (16384,0.864), (32768,0.807) and (32768,0.864) respectively. N is code length and R is code rate.

5. In [8], the fabricated ASIC SC decoder supports code length 32768.
is around 1 much lower power, which reduced the averaged power level. During the first five (1-5) iterations. The last three (5-8) iterations consume with five iterations. This is due to the lower error detection successful rate can reach 100 achieved within approximately 400. Scaled to 16nm process, the area efficiency is around 50 Gbps/mm², again the power consumption is smaller than 100mW/mm². The energy efficiency is around 1 pJ/bit, again meeting the target proposed in [12]. These results are plotted in Fig. 6. The power consumption and energy efficiency are evaluated with TSMC 16nm process.

V. CONCLUSIONS

In this paper, we present an ASIC implementation of high-throughput G_N-coset codes. The parallel decoding framework leads to a hardware with high area efficiency and low decoding power consumption. An area efficiency of 120Gbps/mm² is achieved within approximately 1mm²@16nm process. The power consumption is as low as 100mW and energy efficiency is around 1pJ/bit. Scaled to 7nm process, the area efficiency can reach 533Gbps/mm². It confirms that G_N-coset codes can meet the high-throughput demand in next-generation wireless communication systems.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, “A vision of 6G wireless systems: applications, trends, technologies, and open research problems,” IEEE Network, 2019.
[2] E. Arkan, “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, Jul. 2009.
[3] X. Wang et al., “On the construction of G_N-coset codes for parallel decoding,” Accepted by IEEE Wireless communications and Networking Conference 2019.
[4] X. Wang et al., “Toward Terahits-per-second Communications: Low-Complexity Parallel Decoding of G_N-coset Codes,” Available on ArXiv.
[5] A. Alamdar-Yazdi and F. R. Kschischang, “A Simplified Successive-Cancellation Decoder for Polar Codes,” in IEEE Communications Letters, vol. 15, no. 12, pp. 1378-1380, December 2011.
[6] G. Sarkis, P. Giard, A. Vardy, C. Thibeault and W. J. Gross, “Fast and flexible successive-cancellation list decoders for polar codes,” IEEE Transactions on Signal Processing, vol. 65, no. 21, pp. 5756–5769, Nov 2017.
[7] A. Balatsoukas-Stimoung, M. B. Parizi and A. Burg, “LLR-Based Successive Cancellation List Decoding of Polar Codes,” in IEEE Transactions on Signal Processing, vol. 63, no. 19, pp. 5165-5179, Oct. 2015.
[8] X. Liu et al., “A 5.16Gbps decoder ASIC for Polar Code in 16nm FinFET,” 2018 15th International Symposium on Wireless Communication Systems (ISWCS), Lisbon, 2018, pp. 1-5.
[9] X. Liu et al., “A 5.16Gbps decoder ASIC for Polar Code in 16nm FinFET,” Accepted by IEEE Transactions on Signal Processing, vol. 63, no. 19, pp. 5165-5179, Oct. 2015.
[10] Taiwan Semiconductor Manufacturing Company Limited, “TSMC 10nm Technology,” [Online]. Available: https://www.tsmc.com/english/dedicatedFoundry/technology/10nm.htm.
[11] Taiwan Semiconductor Manufacturing Company Limited, “TSMC 7nm Technology,” [Online]. Available: https://www.tsmc.com/english/dedicatedFoundry/technology/7nm.htm.
[12] “EPIC - Enabling practical wireless Tb/s communications with next generation channel coding,” [Online]. Available: https://epic-h2020.eu/results.