Abstract—LDPC codes are usually decoded by iterative belief propagation. However especially for small block lengths conventional belief propagation exhibits significant losses in signal-to-noise ratio compared to maximum likelihood decoding. In this paper we propose the combination of a conventional min-sum decoder enhanced by an advanced decoding scheme, that acts as a kind of “afterburner” to improve the frame error rate. We present hardware architectures and implementation results for a 28 nm ASIC technology. The new decoder has a slightly higher complexity, but provides a gain of up to 1.6 dB signal-to-noise ratio over conventional belief propagation decoding for short block length. In addition, we show, that the new decoder implementation can decrease the amount of dark silicon.

I. INTRODUCTION

Low-density parity-check (LDPC) codes are powerful channel codes, which are widely used in practical applications. For efficient hardware implementations of LDPC decoders usually belief propagation (BP) heuristics are used. A drawback of these heuristics is, that their frame error rate (FER) is considerably worse than optimum maximum likelihood (ML) decoding. This effect becomes more evident if the block length of the code is short, due to loops in the Tanner graph. It has been shown in literature, that gains of more than 1 dB are possible with Binary Phase Shift Keying (BPSK) modulation [1], [2] and even more when higher order modulation is used for ML decoding.

It is therefore worth to investigate new decoding methods and their hardware implementations, to improve the FER of currently used systems. Several algorithms have been published, e.g. using information set decoding [3], augmented BP [2] or saturated min-sum (SMS) decoding [4].

In practical applications, decoding algorithms must be implemented as dedicated hardware to meet throughput and power requirements. In [4] an SMS decoding algorithm and hardware implementation have been presented to improve the FER of LDPC decoders by up to 1.2 dB. However, the drawback of this advanced algorithm is its complexity, which leads to a poor area efficiency, i.e. large chip area or low decoder throughput.

In this paper, we propose a novel hybrid approach, which features a conventional LDPC decoder and an improved SMS decoder. It combines the superior FER of the SMS decoder with the throughput of a conventional decoder. SMS decoding serves as a kind of “afterburner”, which is only activated if the conventional decoder fails. The main objective of this approach is to increase the communications performance. We present a serial and parallel hardware architecture for this new approach. These architectures have been implemented using a 28 nm Fully Depleted Silicon On Insulator (FD-SOI) ASIC technology. Detailed implementation results for area, throughput and power consumption are given. Compared to sole SMS decoding of [4] the new approach decreases the average complexity largely. This results in more efficient hardware designs or a further improvement in FER. Gains of up to 1.6 dB have been achieved over conventional min-sum decoding with only little impact on area efficiency and power consumption.

Moreover, it is shown in this paper, that the proposed decoding scheme lends itself very good to the dark silicon phenomenon. The term “dark silicon” stems from the fact that in emerging technologies large parts of a chip cannot be activated simultaneously due to thermal design power constraints. This can be explained by the discontinuation of Dennard scaling [5]. It states that as transistors get smaller, their power density stays constant. The discontinuation of Dennard scaling in future technologies leads to a rapidly increasing on-chip power density, that disallows to run the complete chip with full power, eventually leading to so-called dark silicon. Recent studies show that for 8 nm technologies the amount of dark silicon, i.e. the amount of transistors which cannot be turned on at the same time can be as much as 80% [6], [7], [8]. As illustrated in [9] an intelligent temperature management limiting the peak temperature and avoiding hot spots on the chip, so called dark silicon patterning [10] can reduce the amount of dark silicon. In this paper, we show how dark silicon can be exploited in LDPC decoding by activating costly SMS decoding only on demand and using dark silicon patterning for conventional decoding as will be described in Section V.

The structure of the paper is as follows: Section II introduces the system setup followed by the SMS decoding algorithm in Section III and the new approach in Section IV. Hardware architectures are presented in Section V and detailed results on FER and the ASIC implementation are given in Section VI.

II. PRELIMINARIES

We consider the transmission of $N$ code bits $x = (x_0, x_1, \ldots, x_{N-1})$, $x_i \in GF(2)$ over an additive white Gaussian noise (AWGN) channel using BPSK and 64-QAM modulation. On the receiver side the received bits are described by a vector of log likelihood ratios (LLRs) $y = (y_0, y_1, \ldots, y_{N-1})$, $y_i \in \mathbb{R}$, with $y_i = \ln \left( \frac{p(x_i=1)}{p(x_i=0)} \right)$. $|y_i|$ can be seen as a measure of reliability of the received bits. A small $|y_i|$ corresponds to
an unreliable bit that is more likely in error than others. The below described SMS algorithm will utilize this information.

Since the FER of BP in general gets worse for decreasing block length we focus on short LDPC codes. For this kind of codes used in today’s or future applications the need for sophisticated algorithms is large. For analysis of the algorithm and the hardware implementation we consider a (96,48) quasi cyclic LDPC code and a (128,64) multi-edge type (MET) LDPC code [11] (public available from [12]).

MET codes have an excellent FER performance and are therefore interesting for systems with strict FER requirements. For the same reason, we have set the maximum number of decoding iterations for the LDPC decoders to a fairly large number of 30, above which only very little improvements in FER can be achieved.

In this paper we use hardware efficient min-sum LDPC decoding, instead of the optimal sum-product algorithm (SPA). However Figure 1 shows, that for the considered codes the min-sum decoder with fixed point values does not compromise the FER in comparison to the SPA. Besides that it should be mentioned, that the ideas in this paper are not restricted to the min-sum check node, but apply to other check node types as well.

III. THE SMS DECODING ALGORITHM

The SMS decoding algorithm has been introduced in [4] and is briefly reviewed in the following. The algorithm is inspired by augmented BP [2]. This kind of algorithms perform multiple LDPC decoding runs. Each run gets different input LLRs, where some of the unreliable bits are saturated. Finally the best codeword among all runs is output according to the soft decision distance metric commonly used for list decoding [13].

In the SMS algorithm S bit positions are selected for saturation prior to starting the decoder. In a first step the least reliable bits are determined by partial sorting of $|y_i|$. After that, $2^S$ input LLR vectors $y^l, l = 0, \ldots, 2^S - 1$ are generated by setting the least reliable positions to the minimum or maximum LLR value, i.e. they are saturated. In total $2^S$ vectors are generated and for each vector a decoding run has to be executed.

The main difference between augmented BP in [2] and SMS decoding is in its potential of an efficient hardware implementation without compromising FER. Advantages over augmented BP are:

- simplified node selection, i.e. reliability based only
- simplified scheduling of decoding runs
- advanced stopping criterion, which does not impact FER
- simplified check node with min-sum decoding

Since the SMS decoding process is time consuming for large values of $S$ a stopping criterion has been introduced in [4]. This stopping criterion is based on the observation, that in $2^S$ decoding runs only a small number of decoders converge to a codeword (often only two or three even for $2^S = 16$ or $2^S = 64$). Instead of evaluating the results of all decoding runs, the decoder stops after it has found a previously specified number of codewords (regardless if they are different codewords or not). This number is reflected by the threshold given in Algorithm 1 Line 8. The number of codewords that have to be considered is dependent on the used code and the parallelism of the hardware architecture. In [4] is has been shown, that typical values range between 2 to 4 for $2^S = 16$ or 2 to 5 for $2^S = 64$.

IV. SMS DECODING AS AFTERBURNER

SMS as a standalone decoder is quite complex since many decoding runs have to be performed. Thus its hardware implementation has a poor area efficiency. The main idea of this paper is to overcome this problem by using a hybrid approach of a min-sum decoder operating with the original LLR channel values, in the following termed “conventional decoding”, and a sophisticated SMS decoder.
Decoding is first tentatively performed with the conventional min-sum decoder. If it fails to find a valid codeword, and only then, SMS decoding is executed. In this configuration, SMS decoding acts just like an afterburner in a jet plane: most of the time “normal” and economic operation of the engine is sufficient. But if the pilot requires maximum thrust in critical situations, he can activate the afterburner. It injects additional fuel burned after the turbine of the engine and provides maximum performance, however at a very high cost of fuel. Following this principle: in most cases conventional (and economic) min-sum decoding is sufficient. But if it fails, and only then, a large effort is spent by activating SMS decoding as an afterburner. It “burns” down the intractable LLR vector and thus often eventually outputs the correct code word, achieving very good FER. For planes and decoding the same holds true: since the afterburner is only activated, when required, the overall system operates very economically. Algorithm 2 and Figure 2 outline the procedure.

Algorithm 2 Decoding Algorithm with an SMS Afterburner

Input: LLR vector $y$, number of saturated bits $S$
Output: decoded codeword $x_{best}$

1: Perform 30 iterations of standard min-sum decoding
2: if Codeword is found (and optional: CRC is fulfilled) then
3:       Output this codeword
4: else
5:       Perform SMS decoding (Algorithm 1)
6:       Output SMS result
7: end if

Figure 3 shows how often the afterburner SMS is activated in comparison to the total number of decoded frames. Note, that different values of $S$ lead to different plots, because the Signal-to-Noise Ratio (SNR) to achieve a specific FER is dependent on the FER for the afterburner decoding and therefore on $S$. Even for an FER of $10^{-3}$ as it is typically used in communication systems, the SMS afterburner is activated only rarely in approximately 2% of the cases and less.

Optional CRC aided Decoding: It is worth to mention, that Algorithm 2 fails, if the conventional decoder finds an incorrect, but valid, codeword. The occurrence of such cases may be dependent on the operating SNR region and the used channel code. Although this problem does not occur with the considered codes for this paper, we propose a solution to this potential problem: the adoption of CRC checks. The CRC can determine with a very high reliability, if the found codeword is the correct one. Checking the correctness of the conventional decoder output avoids accepting a (wrong) codeword. If the CRC fails SMS is activated to raise the probability to find the correct codeword. For many communication standards, that utilize CRC checks, this technique can be easily adopted.

V. Architecture

In the following section the hardware architectures for the decoder are presented. Figure 2 shows the basic decoder setup with SMS afterburning. Since the min-sum decoder and the SMS decoder both require an architecture for min-sum decoding, they can share the same min-sum decoder kernel.

Two approaches are possible: A parallel or a serial architecture, see Figure 4 and Figure 5. In both cases, the min-sum decoder in Figure 2 is moved into the SMS decoder architecture. General min-sum decoding is performed by the “Inner Min-Sum Decoder”. It can be implemented by any state-of-the-art architecture used for LDPC decoding today. For our case we used a well-known partially parallel LDPC decoder architecture implementing min-sum check nodes. It is based on the slot-layered concept [14] processing three submatrices of the parity check matrix in parallel and applying layered decoding. The architecture has a good trade-off between implementation complexity and throughput. However, the presented concept of an afterburner for min-sum decoding can be adapted for any possible inner decoder architecture.

Choosing a serial or parallel architecture for the SMS processing has no impact on the communications performance of the final decoder, but affects area efficiency and latency.

A. Serial Decoding Architecture

In the serial architecture only one inner min-sum decoder is instantiated. Embedded in the serial decoding architecture, see Figure 4, this inner decoder first performs conventional decoding using the channel values as input. If conventional decoding fails, the additional $2^S$ decoding runs of SMS decoding are successively carried out by the same inner decoder instance. In this case, it uses saturated input LLR vectors.

The number of iterations for a block, and therefore latency depends largely on the success of the first (conventional) min-sum decoding run. Very often it will converge after only a few
As already stated, in most cases only one of the $2^S$ inner decoders of the parallel architecture is active at a time, performing conventional decoding. We can switch off the other decoders by power gating, resulting in a low average power consumption of the decoder. Furthermore, to avoid hotspots, we do not always use the same inner decoder for conventional decoding. Instead we switch after every decoded frame to the next inner decoder in a round robin fashion. Then a single inner decoder is active only for approximately $1/2^S$th of the time. Thus the inner decoders can cool down before the next activation and the temperature distribution is averaged. This "heat dissipation aware" decoding largely reduces hot spots on the chip, a major concern in future technologies. This can be seen as a dark silicon patterning [10] for LDPC decoding. SMS decoding is applied only in the rare event of a min-sum decoding failure. In this case all inner decoders are activated for a short time. However, the largest part of the decoder is switched on only rarely (for SMS decoding) and keeps the average power consumption and thus heat dissipation low.

VI. Results

In the following section, we first introduce the setup used for the hardware implementation, then present the resulting FER performance and ASIC synthesis results.

A. Implementation Setup

All designs are implemented on a 28 nm low power FD-SOI CMOS library. We considered the following PVT parameters: Worst Case (WC, 0.9 V, 125$^\circ$C), Nominal Case (NOM, 1.0 V, 25$^\circ$C) and Best Case (BC, 1.1 V, -40$^\circ$C). Synthesis as well as place & route (P&R) were performed with the Worst Case PVT settings of the 28 nm library. The designs where synthesized for a target frequency of 500 MHz which is achieved under Worst Case conditions after P&R. The power consumption was calculated using Synopsys PrimeTime-PX. The average power consumption at a clock frequency of 500 MHz is given for the Nominal Case in Table I and Table II. We show results on area, throughput and energy. Area efficiency normalizes throughput to area and allows for a fair comparison of different architectures with different parallelisms. Results are shown for a FER of $10^{-5}$ which is realistic for systems with high demands on the error correction performance. The decoders operate on fixed-point values, which were optimized to not impacting the FER. The chosen quantisation is 7 bits for the (96,48) LDPC code and 8 bits for the MET code.

B. Simulation Results for FER

Figures 6 to 8 show the FER for the implemented decoders for the (96,48) LDPC and the (128,64) multi-edge type code. All provided SNR gains in dB of the proposed designs are relative to the FER of a conventional min-sum decoder also shown in the figures. Under BPSK modulation the gain ranges between 0.5 and 1.0 dB and under 64-QAM between 0.7 and 1.6 dB for four, six or ten saturated bits. The ML decoding plots show the maximum theoretically achievable gain for the code. This plots have been obtained by the methods described in [1]. With ten saturated bits the FER can be pushed close to ML decoding performance.
TABLE I. COMPARISON OF CONVENTIONAL MIN-SUM AND SMS AFTERBURNER DECODER ARCHITECTURES FOR THE (96,48) CODE

<table>
<thead>
<tr>
<th>Architecture</th>
<th>min-sum</th>
<th>min-sum + serial SMS</th>
<th>min-sum + parallel SMS</th>
<th>min-sum + serial SMS</th>
<th>min-sum + parallel SMS</th>
<th>min-sum + serial SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code (96,48)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># saturated bits S</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Post P&amp;R Area [mm²]</td>
<td>0.06</td>
<td>0.07</td>
<td>0.89</td>
<td>0.07</td>
<td>3.52</td>
<td>0.07</td>
</tr>
<tr>
<td>TP @ FER=10⁻⁵ [MHz/s]</td>
<td>797</td>
<td>713</td>
<td>737</td>
<td>570</td>
<td>712</td>
<td>86</td>
</tr>
<tr>
<td>Area Eff. [Gbit/s/mm²]</td>
<td>14.6</td>
<td>10.4</td>
<td>0.8</td>
<td>8.3</td>
<td>0.2</td>
<td>1.3</td>
</tr>
<tr>
<td>Worst Case Latency [iterations]</td>
<td>30</td>
<td>510</td>
<td>60</td>
<td>1950</td>
<td>60</td>
<td>30750</td>
</tr>
<tr>
<td>Energy Efficiency [pJ/Codeword]</td>
<td>394</td>
<td>551</td>
<td>533</td>
<td>680</td>
<td>552</td>
<td>4574</td>
</tr>
</tbody>
</table>

SNR gain BPSK [dB] 0.0 0.5 0.5 0.7 0.7 1.0
SNR gain 64-QAM [dB] 0.0 0.7 0.7 1.2 1.2 1.6

TABLE II. COMPARISON OF CONVENTIONAL MIN-SUM AND SMS AFTERBURNER DECODER ARCHITECTURES FOR THE (128,64) MULTI-EDGE TYPE LDPC

<table>
<thead>
<tr>
<th>Architecture</th>
<th>min-sum</th>
<th>min-sum + serial SMS</th>
<th>min-sum + parallel SMS</th>
<th>min-sum + serial SMS</th>
<th>min-sum + parallel SMS</th>
<th>min-sum + serial SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code (128,64)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># saturated bits S</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Post P&amp;R Area [mm²]</td>
<td>0.11</td>
<td>0.13</td>
<td>1.82</td>
<td>0.13</td>
<td>7.24</td>
<td></td>
</tr>
<tr>
<td>TP @ FER=10⁻⁵ [MHz/s]</td>
<td>420</td>
<td>388</td>
<td>391</td>
<td>364</td>
<td>382</td>
<td></td>
</tr>
<tr>
<td>Area Eff. [Gbit/s/mm²]</td>
<td>3.7</td>
<td>3.1</td>
<td>0.2</td>
<td>2.9</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Worst Case Latency [iterations]</td>
<td>30</td>
<td>510</td>
<td>60</td>
<td>1950</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

SNR gain BPSK [dB] 0.0 0.4 0.4 0.5 0.5
SNR gain 64-QAM [dB] 0.0 0.7 0.7 1.0 1.0

C. Synthesis Results

Synthesis results for different SMS decoder architectures and the two considered codes are shown in Tables I and II.

For the serial architectures three scenarios are considered: flipping four, six or ten bits. One can observe that the decoder throughput and area efficiency is very good. Only when flipping ten bits, there is a significant impact on these parameters. This shows that the serial architecture is an excellent candidate to replace a conventional min-sum decoder. The only drawback of this architecture is the worst case latency which is significantly increased. In average the latency is however similar to the one of the min-sum decoder.

Parallel architectures solve this issue and exhibit only a small increase in worst case latency. Due to the fully parallel processing of the SMS decoding, the worst latency is doubled compared to min-sum decoding. However they suffer from reduced area efficiency.

Summarizing, serial architectures perform best unless latency requirements are strict. In this case, parallel architectures should be chosen. If very good FER is required and therefore a large number of saturated bits is used, a fully parallel architecture is no more feasible because of the huge area requirements, while a serial architecture may be prohibitive due to the very high latency. Thus a partial parallel architecture...
In this paper we have introduced a decoder hardware, that uses SMS decoding as an afterburner, if conventional min-sum decoding fails. The approach has been shown to decrease the average energy and area efficiency for practical FER only slightly compared to a standard min-sum decoder, while providing a large gain in SNR of up to 1.6 dB. Furthermore, it has been shown how the proposed architecture can use dark silicon in future ASIC technologies as an advantage by sophisticated scheduling and power gating.

ACKNOWLEDGMENT

We gratefully acknowledge financial support by the DFG (project-ID: KI 1754/1-1 and WE 2442/8-1).

REFERENCES


