A 47 Gb/s LDPC Decoder with Improved Low Error Rate Performance

Zhengya Zhang, Venkat Anantharam, Martin J. Wainwright, Borivoje Nikolić
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

Abstract

A parallel low-density parity-check (LDPC) decoder is designed for the (2048,1723) Reed-Solomon-based LDPC (RS-LDPC) code suitable for 10GBASE-T Ethernet. A two-step decoding scheme lowers the error floor to a 10^{-14} BER. The decoder architecture is optimized for area, power, and high throughput. The resulting 5.35 mm^2, 65nm CMOS chip achieves a decoding throughput of 47.7 Gb/s. With scaled frequency and voltage, the chip delivers a 6.67 Gb/s throughput while dissipating 144 mW of power.

Introduction

LDPC codes have been demonstrated to perform very close to the Shannon limit when decoded iteratively. However, difficulties persist in the deployment of LDPC codes in high-throughput applications due to the high cost of parallel decoder architectures [1] and the error floors at moderate to low error rates [2]. We address the challenges by two approaches: 1) an improved decoding algorithm to resolve a class of combinatorial structures causing error floors even with a very short wordlength of 4 bits; and 2) an efficient grouped parallel architecture that balances the throughput and the area tradeoff.

Decoding Algorithm

The parity-check matrix **H** of the (6,32)-regular (2048,1723) RS-LDPC code is composed of permutation submatrices of size 64×64. This code has been adopted in the IEEE 802.3an standard which supports 10 Gb/s Ethernet over 100 meters of CAT-6a UTP cable [3]. The RS-LDPC code is selected to provide sufficient coding gain to allow for an error-free operation down to the BER level of 10^{-12}.

A high-throughput FPGA-based hardware emulation has been used to initially investigate the low error rate performance of this code, and it has been discovered that a class of (8,8) absorbing-set errors dominate the error floors. The conventional sum-product (SP) decoder implementation suffers from numerical saturation and worsens the effects of absorbing sets, while an offset min-sum (MS) implementation eliminates numerical saturation and achieves a 0.5 dB SNR gain (Fig. 1). To further lower the error floor, a message biasing scheme introduces perturbations to the absorbing sets [4]. The scheme is realized in a two-step decoder composed of a regular decoder to correct the majority of the errors and a post-processor to apply message biasing and resolve the absorbing-set errors (Fig. 2). The post-processor proves to be highly effective: a 4-bit decoder, aided by the post-processor, surpasses the performance of a 6-bit decoder at high SNR (Fig. 1). The post-processor can be conveniently implemented in a small logic block added to each variable processing node.

Architectural Optimization

The intrinsically-parallel message-passing decoding algorithm relies on the message exchange between variable processing nodes (VN) and check processing nodes (CN) in the graph defined by the **H** matrix. A direct mapping of the interconnection graph causes large wiring overhead and low area utilization. Architectures with lower degrees of parallelism can be attractive, especially for very structured codes.

Each permutation submatrix of a structured **H** matrix can be viewed as a router (Fig. 3 (a) and (b)). To reduce the level of parallelism, the routers are combined and the routing operations are time-multiplexed. In the (6,32)-regular **H** matrix under consideration, the 6 routers in every column can be combined and time-multiplexed, resulting in a 1-dimensional, 32-way parallel architecture (1d32x). The resulting local units, shown as the VNG blocks in Fig. 3(c), encapsulate local wiring irregularities so that wires outside of the local units are structured. We define the area expansion factor (AEF) as the area of the complete decoder over the simple concatenation of the areas of all stand-alone processing nodes. For example, the 1d32x architecture uses 2,048 VNs and 64 CNs, with an AEF of 1.46. A selected set of architectures has been explored: from 1d8x, 1d16x, to 1d32x-2d2x (a 64-way parallel architecture spanning two dimensions) with increasing degrees of parallelism. A less parallel design results in more complex local units, while a highly parallel design incurs more global wiring overhead. The optimal level of parallelism lies in the flat region of the AEF versus throughput curve, which corresponds to the balance of throughput and area. A design in this optimal region incurs the lowest incremental wiring overhead per additional processing node (Fig. 4). As such, the 1d32x architecture (Fig. 5) is selected for implementation.

The two-step decoding scheme allows the decoder wordlength to be reduced to 4 bits. First, the 6-bit wordlength that satisfies the BER requirement is reduced to 4 bits to cut the wiring by 41%, shrink the area by 38%, and increase the maximum clock frequency to 400 MHz (Fig. 6). Second, the post-processor is added to the 4-bit MS decoder to lower the error floor by orders of magnitude costing only 13.7% additional area and merely 1.7% additional wiring (Fig. 6).

Implementation and Measured Results

The decoder supports automated functional testing by incorporating AWGN noise generators and an error collection. The decoder is implemented in 65nm 7M low-leakage CMOS technology. An initial density of 80% is used in place and route to produce the final density of 84.5% in a 5.35 mm^2 core. The waterfall curve shows an excellent BER performance down to 10^{-14} with the post-processor enabled (Fig. 7), which matches hardware emulation (Fig. 1) with extended BER by two orders of magnitude at high SNR. The post-processor suppresses the error floor by eliminating the absorbing-set errors. In fact, five of the seven unresolved errors at the highest SNR point measured are due to undetected errors (codewords), indicating the near maximum-likelihood decoding performance.

The power consumption of the chip is reduced largely by wordlength reduction and architectural optimization, which also permit a higher clock frequency and thus a higher decoding throughput. An early termination scheme is implemented on-chip to improve throughput by detecting early convergence and immediately proceeding to the next input frame. Additional power reduction is achieved by frequency and voltage scaling. The effect of each step of power reduction is shown in Fig. 8. The chip operates with a maximum clock frequency of 700 MHz using the nominal 1.2 V supply, which delivers a throughput of 47.7 Gb/s (Fig. 9) measured at an SNR level of 5.5 dB with early termination enabled on-chip. To achieve the required 6.67 Gb/s of throughput for 10GBASE-T Ethernet, the chip can be frequency and voltage scaled to operate at 100 MHz in a 0.7 V supply, resulting in a maximum decoding latency of 960 ns (assume an 8-iteration decoding limit) and a power dissipation of 144 mW (Fig. 9). The techniques used in designing this decoder are also suitable for applications in data storage, high-speed wireless, and optical communications.

Acknowledgements

The authors acknowledge chip fabrication donation of STMicroelectronics, NSF CCF grant no. 0635372, CNS RI grant no. 0403427, Marvell, Intel Corp., Infineon, and the UC MICRO program.

References

[1] A.J. Blanksby, C.J. Howland, *IEEE JSSC*, Mar. 2002.
[2] T. Richardson, *Proc. Allerton Conf.*, Oct. 2003.
[3] *IEEE Std 802.3an-2006*, Sep. 2006.
[4] Z. Zhang, et al., *Proc. IEEE GLOBECOM*, Nov. 2008.

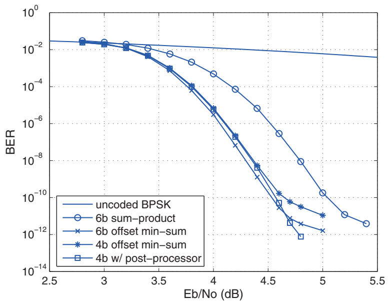


Fig. 1 BER performance comparisons: sum-product, min-sum, wordlength, and post-processing (results are obtained using hardware emulation).

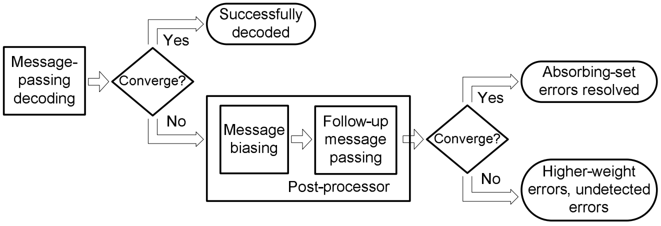
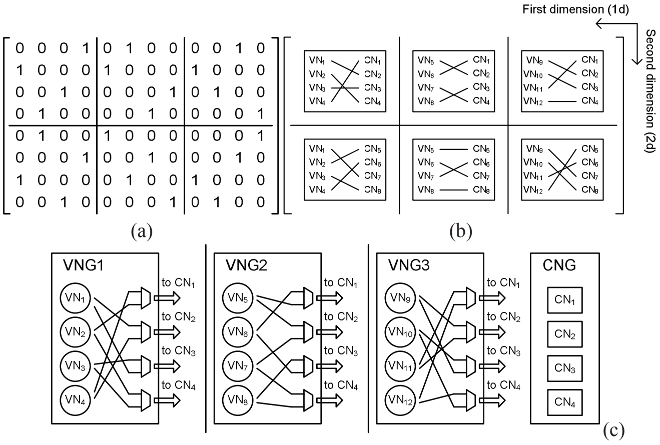


Fig. 2 A two-step decoder composed of a regular decoder and a post-processor.



(a) (b) (c)

Fig. 3 Illustrations of (a) a simple structured **H** matrix, and (b) the fully parallel architecture, and (c) a 1d3x parallel architecture.

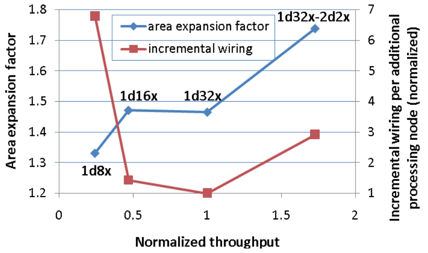


Fig. 4 Architectural optimization by the area expansion metric.

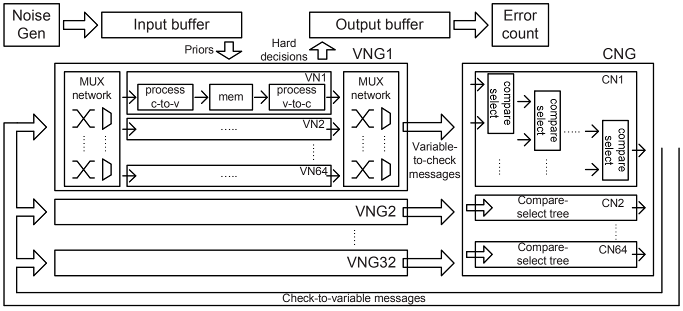


Fig. 5 The decoder implementation using the 1d32x architecture.

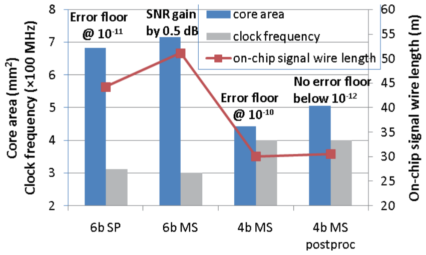


Fig. 6 Area and performance improvement evaluated on the 1d32x architecture using synthesis, place and route results in the worst-case corner.

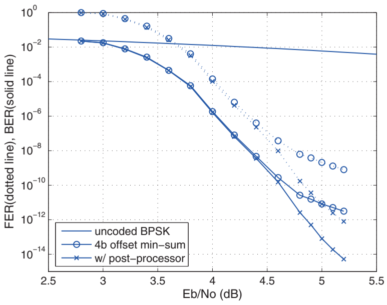


Table I. Error Statistics (Number of Frame Errors before and after Post-processing)

SNR (dB)	Frame errors	Post proc
4.8	3396	95
4.9	4229	40
5.0	4553	18
5.1	5714	11
5.2	7038	7

Fig. 7 Measured FER (dotted lines) and BER (solid lines) performance of the decoder chip and error statistics.

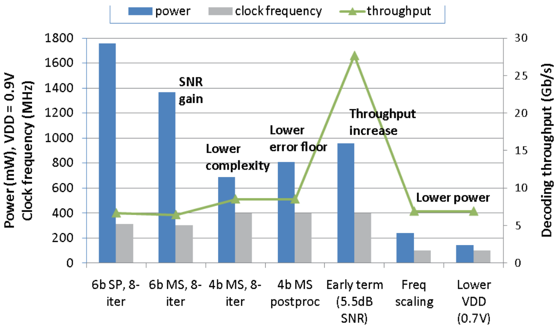


Fig. 8 Power reduction steps with results from synthesis, place and route in the worst-case corner.

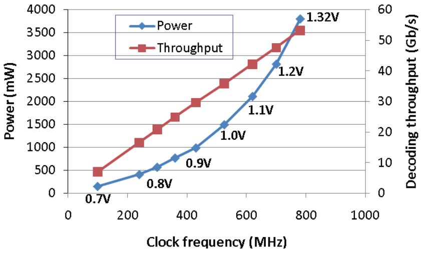


Fig. 9 Frequency and power measurement results of the decoder chip.

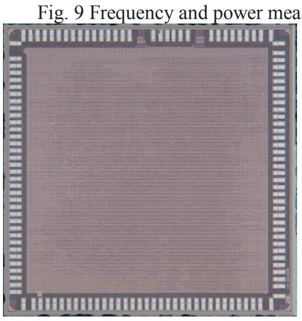


Table II. Chip Features

Technology	65nm low-leakage CMOS	
Core area	2.316 × 2.311 mm	
Chip area	2.556 × 2.608 mm	
Density	80% initial, 84.5% final	
Frequency	100 MHz	700 MHz
Supply	0.7 V	1.2 V
Power	144 mW	2.80 W
Throughput	6.67 Gb/s	47.7 Gb/s
Latency	960 ns	137 ns
Energy	21.5 pJ/bit	58.7 pJ/bit

Fig. 10 Chip microphotograph and feature summary. Throughput is measured at 5.5 dB SNR with early termination enabled. Maximum latency assumes an 8-iteration decoding limit.