A 1.6-mm^2 38-mW 1.5-Gb/s LDPC Decoder Enabled by Refresh-Free Embedded DRAM

Youn Sung Park, David Blaauw, Dennis Sylvester, Zhengya Zhang
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor

Abstract

Memory dominates the power consumption of high-throughput LDPC decoders. A 700 MHz refresh-free embedded DRAM (eDRAM) is designed as a low-power memory to retain data for the required access window. 32 1-kb eDRAM arrays are integrated in a 1.6 mm^2, 65nm LDPC decoder suitable for IEEE 802.11ad. The LDPC decoder consumes 38 mW for a 1.5 Gb/s throughput at 90 MHz and 10 decoding iterations, and it achieves up to 9 Gb/s at 540 MHz.

Introduction

High-throughput LDPC decoders have been built using highly parallel architectures supported by high-bandwidth memory [1], [2]. In such designs, registers are the preferred choice for high speed and wide access, at the expense of high power and area. We address the memory challenge by proposing a new refresh-free eDRAM to replace registers, in conjunction with sequential addressing, full-bandwidth access, and optimal memory timing for low power and high integration density.

Efficient Memory for High-Throughput LDPC Decoder

A conventional eDRAM achieves higher integration density than SRAM, but with a slower access speed [3]-[5]. Periodic refresh is necessary to maintain continuous data retention. Refresh-free eDRAM is motivated by the observation that an LDPC decoder follows stream processing where data is produced and consumed in a deterministic pattern. The time between the first write and the last read of a piece of data, known as the access window, can be short relative to eDRAM's retention time, thus refresh can be eliminated to reduce power and complexity. A recent logic-compatible eDRAM retains data for up to 1.2 ms, relying on the innate capacitances of MOS devices [5]. The excess retention time beyond the access window can be traded off for higher access speed, or the supply voltage can be scaled for reduced power.

A performance-optimized, 672-parallel LDPC decoder (Fig. 1(a)) meets the required throughput of 1.5 Gb/s, 3 Gb/s, and 6 Gb/s for IEEE 802.11ad [6], [7] at 90 MHz, 180 MHz, and 360 MHz respectively at 10 decoding iterations. The decoder consists of 16 groups of 42 variable processing nodes (VN) and 1 group of 42 check processing nodes (CN), along with shifters to shuffle data between processing nodes. Post-layout synthesis of the decoder in 65nm CMOS reveals that 57% of the power is attributed to the internal memory (Fig. 2) that supports the 2 Tb/s data access, in addition to a significant portion of the clock tree power. Memory power can be further broken down based on the type of the data stored: 1) prior memory for channel inputs, 2) extrinsic memory for data exchanged between processing nodes, 3) posterior memory for decoding results, and 4) buffers for data alignment. Altogether, extrinsic memory and buffers consume 70% of the memory power (Fig. 2). The access window of these power-hungry memory and buffers are short: 5 clock cycles for extrinsic data and 5 cycles for buffer data based on the pipeline structure in Fig. 1(c). The short access window makes refresh-free eDRAM a good choice, offering a small footprint, low power, and competitive access speed.

The deterministic memory access enables further optimization of the eDRAM to use sequential addressing rather than random addressing to reduce overhead. To support pipelining, the eDRAM provides one write and one read every clock cycle to two different memory locations. Considering extrinsic and buffer data access patterns (Fig. 1(c)), read from eDRAM initiates a pipeline stage and write to eDRAM completes a pipeline stage. The eDRAM timing is designed accordingly to allow for reading on the leading edge of the clock with write completion by the trailing edge.

Design and Optimization of Refresh-Free Embedded DRAM

The short access window permits the use of a high-speed 3T LVT NMOS cell in Fig. 3(a). Separate read word line (RWL) and write word line (WWL) provide dual-port access for one write and one read per cycle. Due to the lack of a dedicated storage capacitor, the cell storage node is vulnerable to capacitive coupling. By connecting RWL to the storage transistor T2, coupling from WWL and RWL can be balanced (Fig. 3(b)): the storage node voltage drops with the falling WWL upon completing a write; this voltage drop is then compensated by the rising RWL at the beginning of a read. By properly sizing T1 and T2, a degraded "1" is compensated before each read, and an inflated "0" is kept low. A compact, single-ended thyristor-based sense amplifier (Fig. 3(a)) offers fast response to the rising read bit line (RBL) (Fig. 3(c)) [8]. To avoid leakage-induced misfires when reading "0", the sense amplifier is implemented in HVT devices.

A 210-column, 5-row (slightly above 1 kb) dual-port eDRAM array is presented in Fig. 4 as one extrinsic (or buffer) memory bank for the LDPC decoder. A compact layout is illustrated in Fig. 5 for an example 4-column, 2-row array where the 4 columns of each row are optimally oriented for sharing WWL, RWL, and convenient connection to the 8 bit lines. This layout is optimized with all M2 tracks fully utilized for the 8 bit lines and the routing of 2 word lines. The array provides access to all its columns at once for parallel processing by a group of 42 VNs. Post-layout extraction of the eDRAM in Fig. 6(a) indicates an operating voltage between 0.7V and 1.2V and a nominal frequency of 700 MHz at 1.0V and 25ºC.

The use of LVT NMOS (for high speed) cuts the data retention time to 20 ns at 125ºC due to heightened leakage. Extra margin can be enabled by under-driving WWL to suppress leakage (Fig. 6(b)): a 0.3V under-drive improves the nominal retention time to 1.6 µs. The retention time is verified to be at least 170 ns for reliable operations by post-layout Monte Carlo simulation at 125ºC (Fig. 7).

Chip Implementation and Measurement Results

An LDPC decoder chip suitable for IEEE 802.11ad is implemented in TSMC 65nm CMOS incorporating 32 1-kb refresh-free eDRAM arrays as extrinsic memory banks and buffers. As a proof of concept, the decoder supports the 1/2-rate (672, 336) code (the worst-case H matrix) [6], but the architecture accommodates the three higher-rate codes with low reconfiguration overhead. The core logic and memory are connected to separate power domains for the optimal supply voltage of each. The decoder is tested using on-chip noise generators and input vectors provided through scan chains. The waterfall curve reaches a BER of 10^{-7} (Fig. 8), which is sufficient for the application. The decoding throughput (at 10 iterations) reaches 9 Gb/s at 540 MHz and down to 0.5 Gb/s at 30 MHz (Fig. 9). The decoder consumes 38 mW, 106 mW, and 374 mW for a throughput of 1.5 Gb/s, 3.0 Gb/s, and 6.0 Gb/s respectively at the optimal core and memory supply voltages indicated in Fig. 9. Power consumption over the SNR range of interest is shown in Fig. 10. Early termination is built-in to increase throughput at high SNR if needed. The chip microphotograph and feature summary are shown in Fig. 11.

The power consumption of the refresh-free eDRAM increases almost linearly with frequency compared to the quadratic increase in core logic power as in Fig. 9, demonstrating the advantage of the eDRAM at high frequency. At 6 Gb/s, the eDRAM consumes only 23% of the total power, and the proportion is further reduced to 21% at 9 Gb/s. The LDPC decoder chip compares favorably to the state-of-the-art implementations in both energy and area efficiency [1], [2], [9]-[12] (Fig. 12). The proposed refresh-free eDRAM and the techniques used in designing this decoder are suitable for other high-throughput DSP applications.

Acknowledgements

The work is supported by NSF CCF-1054270. The authors acknowledge the advice by Yoonmyung Lee.

References

[1] Zhang *et al*., *JSSC*, Apr. 2010.
[2] Cevrero *et al*., *ASSCC*, Nov. 2010.
[3] Somasekhar *et al*., *JSSC*, Jan. 2009.
[4] Barth *et al*., *JSSC*, Jan. 2011.
[5] Chun *et al*., *JSSC*, Jun. 2011.
[6] *IEEE Draft Std 802.11ad*, Nov. 2011.
[7] Weiner *et al*., *ISCAS*, May 2011.
[8] Satpathy *et al*., *VLSI*, Jun 2010.
[9] Hung *et al*., *ASSCC*, Nov. 2010.
[10] Chen *et al*., *ESSCIRC*, Sep. 2009.
[11] Roth *et al*., *ASSCC*, Nov 2010.
[12] Peng *et al*., *ASSCC*, Nov. 2011.

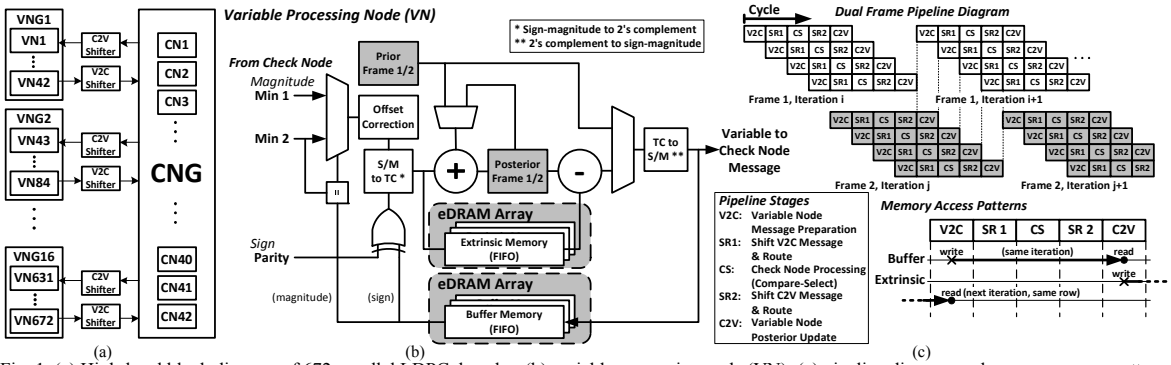

Fig. 1. (a) High-level block diagram of 672-parallel LDPC decoder; (b) variable processing node (VN); (c) pipeline diagram and memory access pattern.

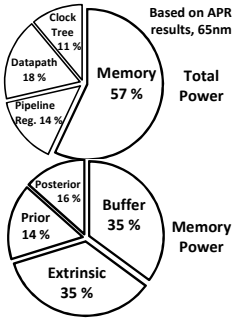

Fig. 2. Power breakdown of register-based IEEE 802.11ad LDPC decoder.

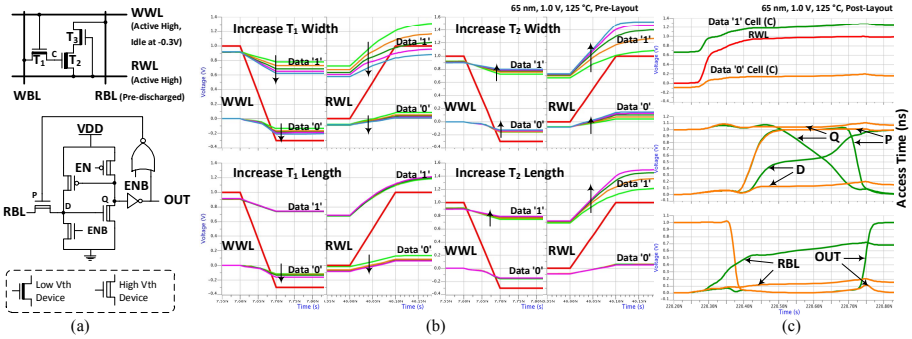

Fig. 3. (a) 3T refresh-free embedded DRAM cell and single-ended thyristor-based sense amplifier; (b) coupling effects with respect to transistor sizing; (c) simulated waveform for a read operation.

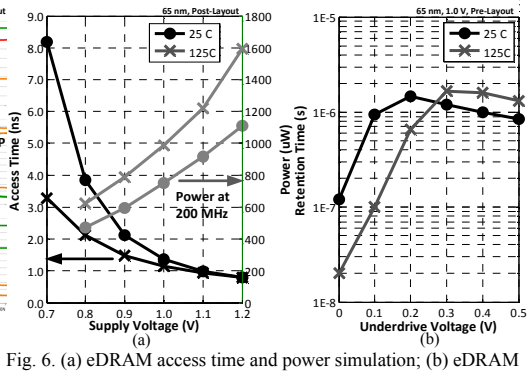

Fig. 6. (a) eDRAM access time and power simulation; (b) eDRAM retention time with word line under-drive.

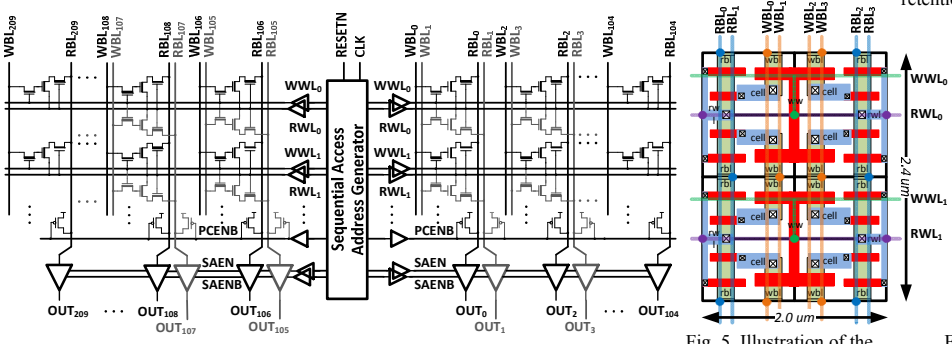

Fig. 4. An eDRAM array consisting of 210 columns and 5 rows of eDRAM cells, together with 210 sense amplifiers and sequential access address generator.

Fig. 5. Illustration of the layout of a small eDRAM array of 4 columns and 2 rows.

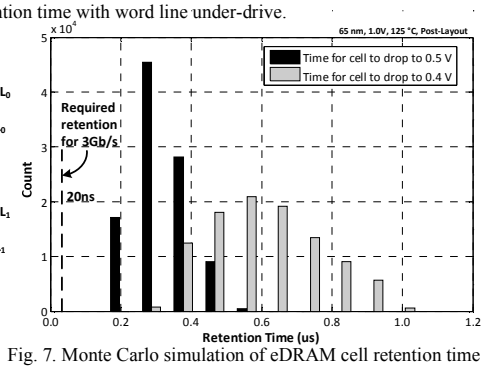

Fig. 7. Monte Carlo simulation of eDRAM cell retention time (measured as time to reach a cell voltage of 0.4V and 0.5V, 0.4V is the minimum needed to guarantee a correct read).

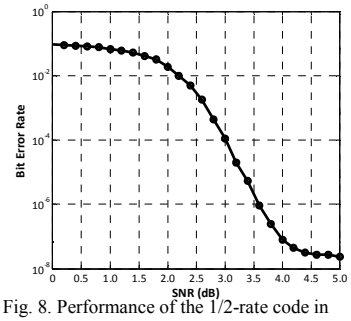

Fig. 8. Performance of the 1/2-rate code in IEEE 802.11ad using 5-bit quantization and 10 decoding iterations.

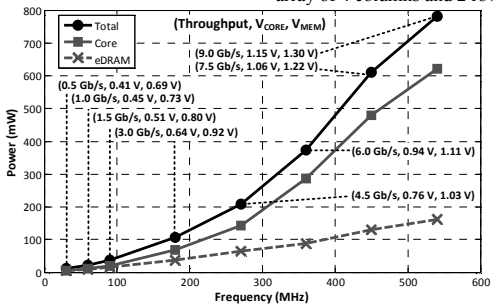

Fig. 9. Core logic and eDRAM power with frequency (measured at 5.0dB SNR, 10 decoding iterations and optimal core and eDRAM voltage).

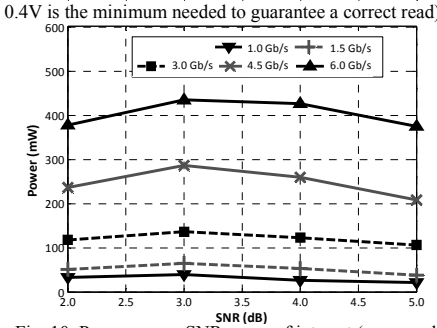

Fig. 10. Power across SNR range of interest (measured at 10 decoding iterations and optimal core and eDRAM voltage).

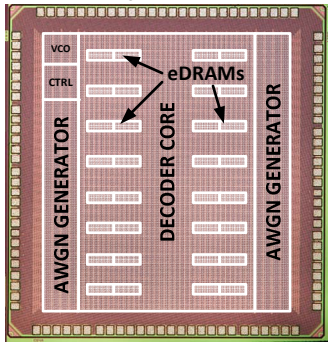

Fig. 11. Chip microphotograph and feature summary.

Technology	TSMC 65nm GP CMOS				
Core Area	1.6 × 1.0 mm (1.6 mm^2)				
Chip Area	1.94 × 1.84 mm (3.57 mm^2)				
Frequency (MHz)	30	90	180	360	540
Core Supply (V)	0.41	0.51	0.64	0.94	1.15
eDRAM Supply (V)	0.69	0.80	0.92	1.11	1.3
Power (mW)*	11.75	37.7	105.8	373.6	782.9
Throughput (Gb/s)*	0.5	1.5	3.0	6.0	9.0
Energy Efficiency (pJ/bit)	21.0	25.6	34.5	61.7	89.5
Area Efficiency (Gb/s/mm^2)	0.31	0.94	1.88	3.75	5.63

* Measured at 5.0dB SNR and 10 iterations for the 1/2-rate code.

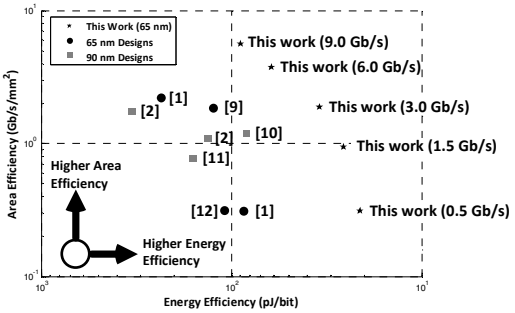

Fig. 12. Energy and area efficiency comparison with state-of-the-art 65nm and 90nm LDPC decoder chip implementations (throughput is normalized at 10 decoding iterations and area normalized to 65nm).