Low-Power High-Throughput LDPC Decoder Using Non-Refresh Embedded DRAM

Youn Sung Park, *Student Member, IEEE*, David Blaauw, *Fellow, IEEE*, Dennis Sylvester, *Fellow, IEEE*, and Zhengya Zhang, *Member, IEEE*

***Abstract*—The majority of the power consumption of a high-throughput LDPC decoder is spent on memory. Unlike in a general-purpose processor, the memory access in an LDPC decoder is deterministic and the access window is short. We take advantage of the unique memory access characteristic to design a non-refresh eDRAM that holds data for the necessary access window, and further improve its access time by trading off the excess retention time. The resulting 3T eDRAM cell is designed to balance wordline coupling to reliably retain data for a fast access. We integrate 32 5x210 non-refresh eDRAM arrays in a row-parallel LDPC decoder suitable for the IEEE 802.11ad standard. Memory refresh is eliminated and random access is replaced with a simple sequential addressing. With row merging and dual-frame processing, the 1.6 mm² 65 nm LDPC decoder chip achieves a peak throughput of 9 Gb/s at 89.5 pJ/b, of which only 21% is spent on eDRAMs. With voltage and frequency scaling, the power consumption of the LDPC decoder is reduced to 37.7 mW for a 1.5 Gb/s throughput at 35.6 pJ/b.**

***Index Terms*—Embedded DRAM, LDPC code, LDPC decoder architecture, low-power DSP design.**

I. Introduction

FOLLOWING the rediscovery of low-density parity-check (LDPC) code [1], [2] and the demonstration of its near-capacity error correcting performance [3], LDPC codes have found widespread applications including WiFi (IEEE 802.11n) [4], WiMAX (IEEE 802.16e) [5], digital satellite broadcast (DVB-S2) [6], 10-gigabit Ethernet (IEEE 802.3an) [7], magnetic [8] and solid-state storage [9] to support higher data rates and better noise immunity. The excellent error correcting performance of LDPC codes comes at the cost of encoding and decoding, and the cost escalates with increasing throughput.

A 4.84 mm^2 0.13 μm LDPC decoder for WiMAX consumes more than 340 mW for a throughput up to 955 Mb/s [10]. With technology scaling, the area and power consumption of LDPC decoders continue to improve. A 1.56 mm^2 65 nm LDPC decoder for the high-speed wireless standard IEEE 802.15.3c consumes 360 mW for a throughput of 5.79 Gb/s [11]. For a higher throughput, the decoder architecture can be further parallelized, but the power and area increase accordingly. A 5.35 mm^2 65 nm 10-gigabit Ethernet LDPC decoder consumes 2.8 W for up to 47 Gb/s [12].

Parallelizing LDPC decoder for a high throughput increases the interconnect complexity [13]–[17] and memory bandwidth [18]. Though the interconnect challenge has largely been addressed through the use of structured codes and row-parallel [11], [12], [16] or block-parallel architectures [10], [18]–[25], memory bandwidth still remains a major challenge. To support highly parallel architectures, SRAM array needs to be partitioned into smaller banks, resulting in very low area efficiency. Gb/s LDPC decoders use registers for high-speed and wide access, at the expense of high power and area. As a result, memory dominates the power consumption and area of LDPC decoders [26].

We propose logic-compatible embedded DRAM (eDRAM) [27]–[30] as a promising alternative to register-based memory that has been used in building high-throughput LDPC decoders. Logic-compatible eDRAM does not require a special DRAM process and it is both area efficient and low power – an eDRAM cell can be implemented in three transistors [27] and it supports one read and one write port, at half the size of a dual-port SRAM cell and its energy consumption is substantially lower than a register. A conventional eDRAM is however slow. A periodic refresh is also necessary to maintain continuous data retention. Interestingly, we find that when eDRAM is used in high-speed LDPC decoding, refresh can be completely eliminated to save power and access speed can be improved by trading off the excess retention time.

In this work, we co-design a non-refresh eDRAM with the LDPC decoder architecture to optimize its read and write timing and simplify its addressing. An analysis of the LDPC decoder's data access shows that the access window of the majority of the data ranges from only a few to tens of clock cycles. The non-refresh eDRAM is designed to meet the access window with a sufficient margin and the excess retention time is cut short to increase the speed. The resulting 3T eDRAM cell balances wordline coupling to mitigate the effects on its storage. We integrate 32 5$\times$210 non-refresh eDRAM arrays in the design of a 65 nm LDPC decoder to support the (672, 336) LDPC code for the high-speed wireless standard IEEE 802.11ad [31]. All columns of the eDRAM arrays can be accessed in parallel to provide the highest bandwidth. The decoder throughput is further improved using row merging and dual-frame processing to increase hardware utilization and remove pipeline stalls. The resulting decoder achieves a throughput up to 9 Gb/s and consumes only 37.7 mW at 1.5 Gb/s.

Manuscript received August 06, 2013; revised October 28, 2013; accepted December 22, 2013. Date of current version March 05, 2014. This paper was approved by Associate Editor Stefan Rusu. This work was supported in part by NSF CCF-1054270.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 USA (e-mail: parkyoun@umich.edu; blaauw@umich.edu; dennis@eecs.umich.edu; zhengya@eecs.umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2014.2300417

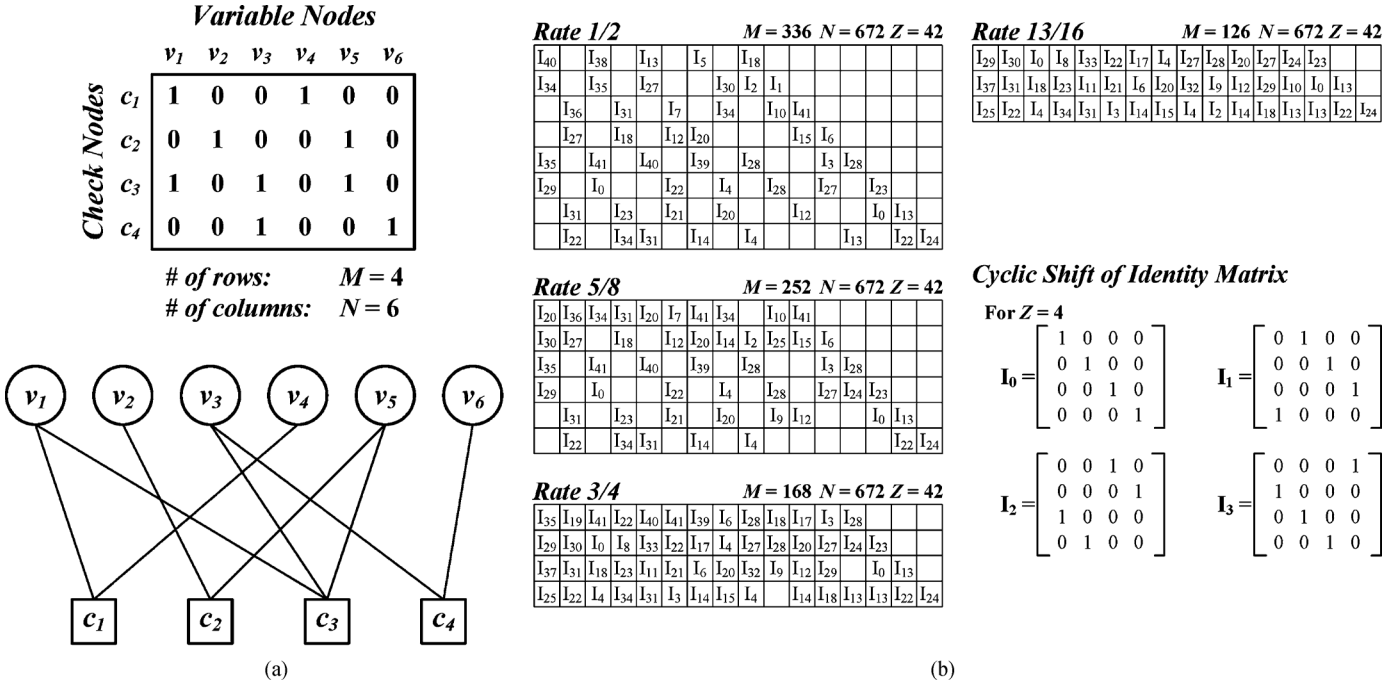


Fig. 1. (a) An example LDPC code's H matrix and bipartite graph representation; and (b) H matrices of the rate-1/2, rate-5/8, rate-3/4 and rate-13/16 LDPC code for the IEEE 802.11ad standard [31].

The remainder of the paper is organized as follows. Section II provides a review of LDPC decoding and the decoder architecture. Two architectural techniques, row merging and dual-frame processing, are used to enhance the decoder throughput, which are described in Section III. Section IV explains how to take advantage of the memory access pattern in LDPC decoding to design a non-refresh eDRAM, and introduces approaches to overcome coupling noise and retention time challenges in the cell design. In Section V, we present a compact eDRAM array for a seamless integration with the LDPC decoder. The implementation of the LDPC decoder test chip and its measurement results are presented in Section VI. Section VII concludes this work.

II. Decoder Architecture

An LDPC code is defined by a $M \times N$ parity-check matrix H [1], [2], where N is the block length (number of bits in the codeword) and M is the number of parity checks. The elements of the matrix $\mathrm{H}(i, j)$ are either 0 or 1 to represent whether bit j of the codeword is part of parity check i. An H matrix can be represented using a bipartite graph composed of two sets of nodes: a variable node (VN) for each column of the H matrix and a check node (CN) for each row. An edge is drawn between $\mathrm{VN}(j)$ and $\mathrm{CN}(i)$ if $\mathrm{H}(i, j) = 1$. An example H matrix with its corresponding bipartite graph is shown in Fig. 1(a).

Almost all of the latest applications have adopted LDPC codes whose H matrix is constructed using submatrices that are cyclic shifts of an identity matrix or a zero matrix. For example, the newest high-speed wireless standard IEEE 802.11ad [31] specifies a family of four LDPC codes whose H matrices are constructed using cyclic shifts of the $Z \times Z$ identity matrix or zero matrix where $Z = 42$. The structured H matrix can be partitioned along submatrix boundaries, e.g., the H matrix of the rate-1/2 (672, 336) code can be partitioned to 8 rows and 16 columns of 42×42 submatrices as shown in Fig. 1(b). The IEEE 802.11ad standard requires the throughput of the LDPC decoding to be between 1.5 Gb/s and 6 Gb/s.

A practical decoder design follows either the sum-product [2] or the min-sum algorithm [32], which are two popular implementations of the belief propagation algorithm. Using the sum-product algorithm in the log-domain, the VNs perform sum operations and the CNs perform log-tanh, sum and inverse log-tanh operations. Min-sum simplifies the CN operation to the minimum function. The min-sum algorithm usually performs worse than the sum-product algorithm, and techniques including offset correction and scaling [33] are frequently applied to improve the performance. We use the min-sum algorithm with offset correction in our implementation.

A. Row-Parallel Architecture

Common LDPC decoder architectures belong to one of three classes: full-parallel, row-parallel and block-parallel [34]. The row-parallel architecture [11], [12], [16] provides a high throughput of up to tens of Gb/s, while its routing complexity can still be kept low, permitting a high energy and area efficiency. To meet the 6 Gb/s that is required by the IEEE 802.11ad standard, we choose the row-parallel decoder architecture. The IEEE 802.11ad standard [31] specifies four codes of rate-1/2, rate-5/8, rate-3/4 and rate-13/16, whose H matrices are made up of 16 columns $\times$ 8 rows, 6 rows, 4 rows and 3 rows of cyclic shifts of the 42×42 identity matrix or zero matrix, as illustrated in Fig. 1(b). The four matrices are compatible, sharing the same block length and component submatrix size.

A row-parallel decoder using flooding schedule is designed using 672 VNs and 42 CNs. The 672 VNs process the soft inputs of 672 bits in parallel by computing VN-to-CN (V2C) messages

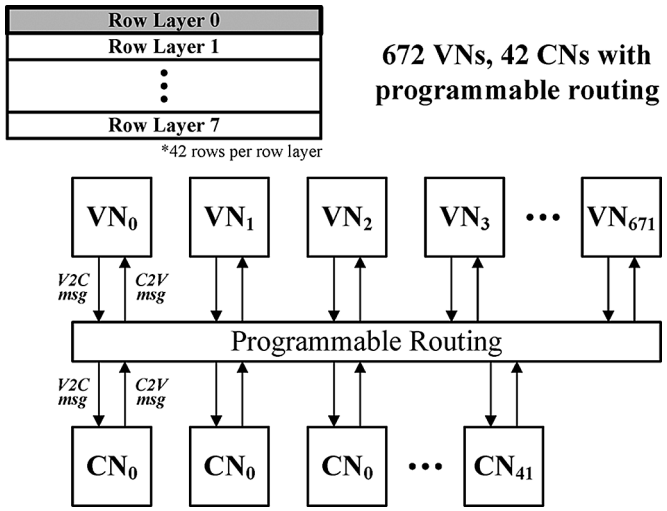


Fig. 2. Illustration of row-parallel LDPC decoder architecture. The shaded part represents the section of the H matrix that is processed simultaneously.

and send them to the 42 CNs following the H matrix shown in Fig. 1(b). The 42 CNs compute the parity checks and send CN-to-VN (C2V) messages back to the VNs. The C2V messages are post-processed by the VNs and stored in their local memories. The row-parallel architecture operates on one block row of submatrices in the H matrix at a time, as highlighted in Fig. 2.

The VN and CN designs in detail are shown in Fig. 3. A VN computes a V2C message by subtracting the C2V message stored in the C2V memory from the posterior log-likelihood ratio (LLR). The V2C message is then sent to the CN while a copy is stored in the V2C memory for post-processing the C2V message later in the iteration. A CN receives up to 16 V2C inputs from the VNs and computes the XOR of the signs of the inputs to check if the even parity is satisfied. The CN also computes the minimum and the second minimum magnitude among the inputs by compare-select for an estimate of the reliability of the parity check. Both the XOR and the compare-select are done using a tree structure. The CN prepares the C2V message as a packet composed of the parity, the minimum and the second minimum magnitude.

After the C2V message is received by the VN, it compares the V2C message stored in memory with the minimum and the second minimum magnitude to decide whether the minimum or the second minimum is a better estimate of the reliability of the bit decision. The sign and the magnitude are then merged and an offset is applied as an algorithmic correction. The post-processed C2V message is stored in the C2V memory. The C2V message is accumulated and summed with the prior LLR to compute the updated posterior LLR. A hard decoding decision is made based on the sign of the posterior LLR at the completion of each iteration. The messages and computations are quantized for an efficient implementation. We determine based on extensive simulations that a 5-bit fixed-point quantization offers a satisfactory performance.

B. Pipelining and Throughput

In the LDPC decoding described above, the messages flow in the following order: (1) each of the 672 VNs computes a

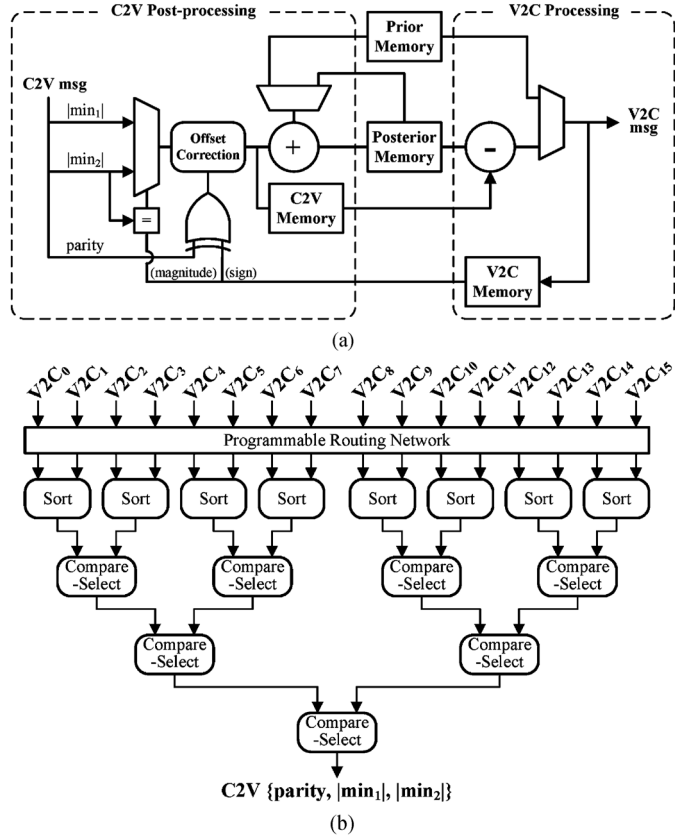


Fig. 3. (a) Variable node, and (b) check node design. (An XOR gate is incorporated in the sort and compare-select logic of the CN to perform the parity check.)

V2C message, which is routed to one of the 42 CNs through point-to-point links; (2) each CN receives up to 16 V2C messages, and computes a C2V message to be routed back to the VNs through a broadcast link; and (3) each VN post-processes the C2V message and accumulates it to compute the posterior LLR. These steps complete the processing of one block row of submatrices. The decoder then moves to the next block row and the V2C routing is reconfigured using shifters or multiplexers. Based on these steps, we can design a 5-stage pipeline: (1) VN computing V2C message, (2) routing from VN to CN, (3) CN computing C2V message, (4) routing from CN to VN, and (5) VN post-processesing C2V messages and computing posterior. For simplicity, the five stages are named VC, R1, CS, R2, and PS, as illustrated in Fig. 4(a). The throughput of a row-parallel architecture is determined by the number of block rows m_b and the number pipeline stages, n_p. The H matrix of the rate-1/2, 5/8, 3/4, and 13/16 code has $m_b = 8$, 6, 4, and 3, respectively. Based on the pipeline chart in Fig. 4(a), the number of clock cycles per decoding iteration is $m_b + n_p - 1$. Suppose the number of decoding iteration is n_{it}, then the decoding throughput is given by

$$\mathrm{TP} = \frac{f_{clk} N}{(m_b + n_p - 1) n_{it}} \tag{1}$$

where f_{clk} is the clock frequency and N is the block length of the LDPC code. $N = 672$ for the target LDPC code. The 1/2-rate LDPC code has the most number of block rows, $m_b = 8$. $n_p = 5$ for the 5-stage pipeline. To meet the 6 Gb/s throughput

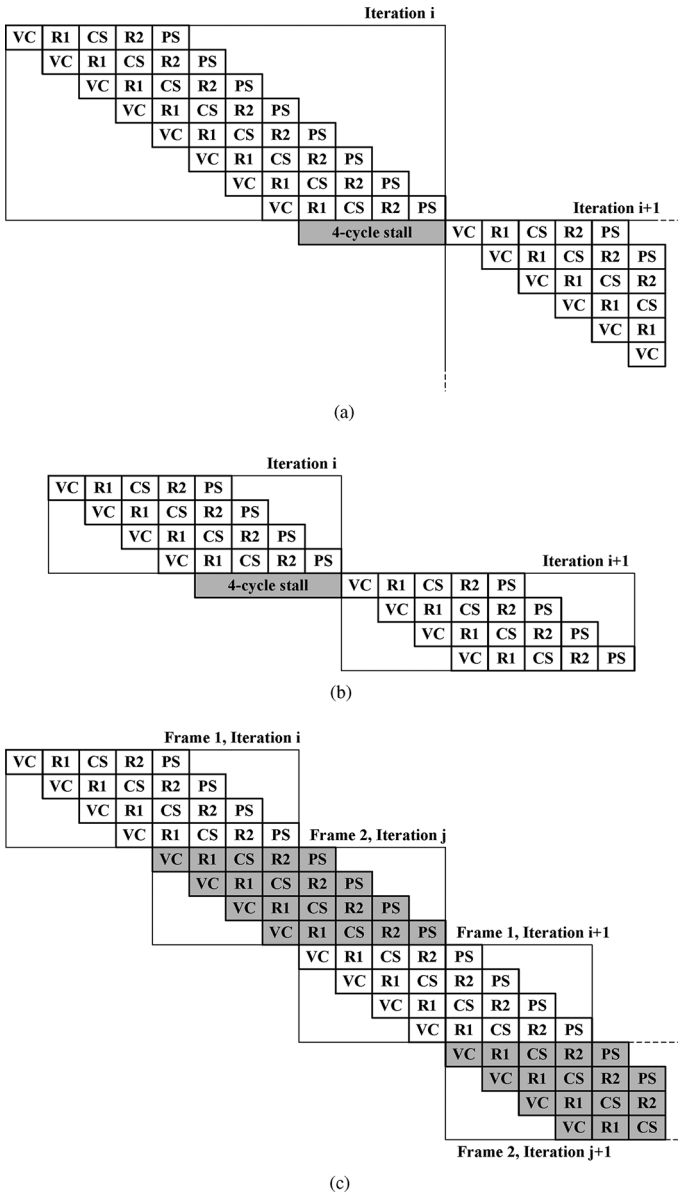


Fig. 4. Pipeline schedule of (a) a conventional single-frame decoder without row-merging, (b) a conventional single-frame decoder with row-merging, and (c) proposed dual-frame decoder with row-merging. Note that (a) and (b) require stalls in-between frames due to data dependency between the PS and VC stages.

with 10 decoding iterations ($n_{it} = 10$), the minimum clock frequency is 1.07 GHz, which is challenging and entails high power consumption.

Each VN in this design includes two message memories, V2C memory and C2V memory. CN does not retain local memory. Each memory contains $m_b = 8$ words to support the row-parallel architecture for the 1/2-rate LDPC code. Each word is 5-bit wide, determined based on simulation. In each clock cycle, one message is written to the V2C memory and one is read from the V2C memory. The same is true for the C2V memory.

For a scalable design and a higher efficiency, the 672 VNs in the row-parallel LDPC decoder are grouped to 16 VN groups (VNG), each of which consists of 42 VNs. The V2C memories of the 42 VNs in a VNG are combined in one V2C memory that contains $m_b = 8$ words and each word is 5 bits $\times\, 42 = 210$ bits wide. Similarly, the C2V memories of the 42 VNs in a VNG are combined in one C2V memory of 8×210 bits. In each clock cycle, one 210 bit word is written to the V2C memory and one 210 bit word is read from the memory. The same is true for the C2V memory. Each memory's read and write access latency have to be shorter than 0.933 ns to meet the 1.07 GHz clock frequency.

III. Throughput Enhancement

The throughput of the LDPC decoder depends on the number of block rows. To enhance the throughput, we reduce the number of effective block rows to process using row merging and apply dual frame processing to improve efficiency [26].

A. Row Merging

The H matrix of the rate-1/2 code has the most number of block rows among the four codes, but note that the H matrix of the rate-1/2 code is sparse with many zero submatrices. We take advantage of the sparseness by merging two sparse rows to a full row so that they can be processed at the same time (e.g., merge row 0 and row 2, row 1 and row 3, etc.), as illustrated in Fig. 5(a). To support row merging, each 16-input CN is split to two 8-input CNs, as in Fig. 5(b), when decoding the rate-1/2 code with minimal hardware additions.

The same technique can be applied to decoding the rate-5/8 code by merging row 2 and row 4, and row 3 and row 5. Row merging reduces the effective number of rows to process to 4, 4, 4, and 3 for the rate-1/2, 5/8, 3/4, and 13/16 codes, respectively. Row merging improves the worst-case throughput to

$$\mathrm{TP} = \frac{f_{clk} N}{(n_p + 3) n_{it}}. \tag{2}$$

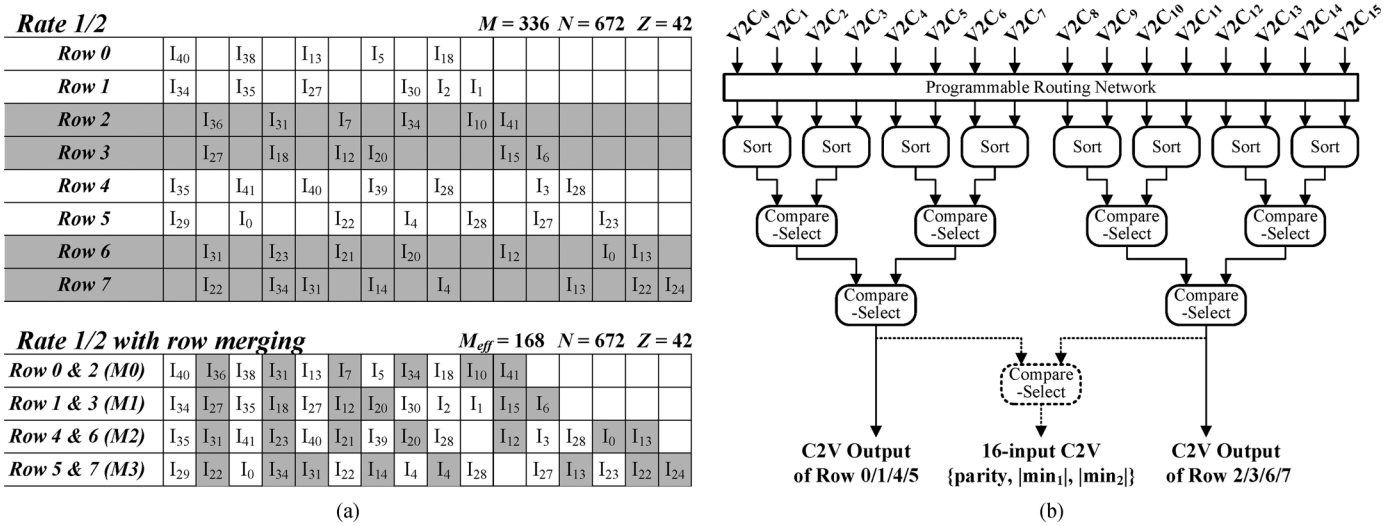
To meet the 6 Gb/s throughput with 10 decoding iterations, the minimum clock frequency is reduced to 720 MHz. Row merging reduces the V2C memory and C2V memory in each VNG to 4×210 bits. Each memory's read and write access latency is relaxed, but it has to be below 1.4 ns to meet the required clock frequency.

B. Dual-Frame Processing

The 5-stage pipeline introduces a 4 clock cycle pipeline stall between iterations, as shown in Fig. 4(a) and (b), because the following iteration requires the most up-to-date posterior LLRs from the previous iteration (i.e., the result of the PS stage) to calculate the new V2C messages. The stall reduces the hardware utilization to as low as 50%.

Instead of idling the hardware during stalls, we use it to accept the next input frame as shown in Fig. 4(c). The ping-pong between the two frames improves the utilization, while requiring only the prior and posterior memory to double in size. The message memories can be shared between the two frames and the computing logic and routing remain the same, keeping the additional cost low. With dual-frame processing, the worst-case throughput is increased to

$$\mathrm{TP} = \frac{f_{clk} N}{4 n_{it}}. \tag{3}$$



Rate 1/2 $M = 336$ $N = 672$ $Z = 42$

Row 0	I_{40}		I_{38}		I_{13}		I_5		I_{18}							
Row 1	I_{34}		I_{35}		I_{27}			I_{30}	I_2	I_1						
Row 2		I_{36}		I_{31}		I_7		I_{34}		I_{10}	I_{41}					
Row 3		I_{27}		I_{18}		I_{12}	I_{20}				I_{15}	I_6				
Row 4	I_{35}		I_{41}		I_{40}		I_{39}		I_{28}			I_3	I_{28}			
Row 5	I_{29}		I_0			I_{22}		I_4		I_{28}		I_{27}		I_{23}		
Row 6		I_{31}		I_{23}		I_{21}		I_{20}			I_{12}			I_0	I_{13}	
Row 7		I_{22}		I_{34}	I_{31}		I_{14}		I_4				I_{13}		I_{22}	I_{24}

Rate 1/2 with row merging $M_{eff} = 168$ $N = 672$ $Z = 42$

Row 0 & 2 (M0)	I_{40}	I_{36}	I_{38}	I_{31}	I_{13}	I_7	I_5	I_{34}	I_{18}	I_{10}	I_{41}					
Row 1 & 3 (M1)	I_{34}	I_{27}	I_{35}	I_{18}	I_{27}	I_{12}	I_{20}	I_{30}	I_2	I_1	I_{15}	I_6				
Row 4 & 6 (M2)	I_{35}	I_{31}	I_{41}	I_{23}	I_{40}	I_{21}	I_{39}	I_{20}	I_{28}		I_{12}	I_3	I_{28}	I_0	I_{13}	
Row 5 & 7 (M3)	I_{29}	I_{22}	I_0	I_{34}	I_{31}	I_{22}	I_{14}	I_4	I_4	I_{28}		I_{27}	I_{13}	I_{23}	I_{22}	I_{24}

(a)

(b)

Fig. 5. (a) Illustration of row merging applied to the H matrix of the rate-1/2 LDPC code of IEEE 802.11ad. The merged matrix has only 4 rows, shortening the decoding iteration latency; and (b) modified check node design to support row merging.

To meet the 6 Gb/s throughput with 10 decoding iterations, the minimum clock frequency is reduced to 360 MHz. To avoid the read after write data hazard due to dual-frame processing, an extra word is added to the V2C and C2V memory. The size of each memory in a VNG is 5×210 bits. Each memory's read and write access latency is further relaxed, but it has to be below 2.8 ns to meet the required clock frequency.

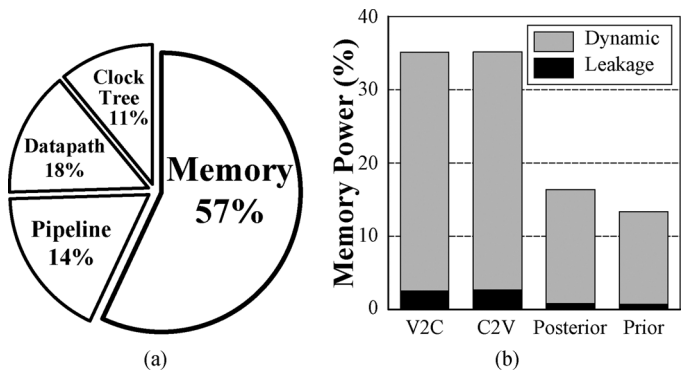
IV. Low-Power Memory Design

The memory in sub-Gb/s LDPC decoder chips is commonly implemented in SRAM arrays, while registers dominate the designs of Gb/s or above LDPC decoder chips. SRAM arrays are the most efficient in large sizes, but the access bandwidth of an SRAM array is very low compared to its size. Therefore SRAM arrays are only found in block-parallel architectures. A full-parallel or row-parallel architecture uses registers as memory for high bandwidth and flexible placement to meet timing.

To estimate the memory power consumption in a high-throughput LDPC decoder, we synthesized and physically placed and routed a register-based row-parallel LDPC decoder that is suitable for the the IEEE 802.11ad standard in a TSMC 65 nm CMOS technology. The decoder follows a 5-stage pipeline and incorporates both row merging and dual-frame processing. In the worst-case corner of 0.9 V supply and 125 °C, the post-layout design is reported to achieve a maximum clock frequency of 200 MHz, lower than the required 360 MHz for a 6 Gb/s throughput.

The power breakdown of this decoder at 200 MHz is shown in Fig. 6. The memory power is the dominant portion, claiming 57% of the total power. In addition to memory, pipeline registers consume 14% of the total power. On the other hand, the datapaths, which include all the combinational logic, consume only 18% of the total power. The clock tree consumes 11% of the total power, the majority of which is spent on clocking the registers. Therefore, reducing the memory power consumption is the key to reducing the chip's total power consumption.

The memory power consumption can be further broken down based on the type of data stored. 35% of the memory power is spent on V2C memory; 35% for C2V memory; 16% for storing posterior LLRs (posterior memory) and 14% for storing prior LLRs (prior memory). The V2C memory and C2V memory account for 70% of the memory power consumption, so they will be the focus for power reduction.



(a) (b)

Fig. 6. (a) Power breakdown of a 65 nm synthesized 200 MHz row-parallel register-based LDPC decoder for the IEEE 802.11ad standard, and (b) memory power breakdown. Results are based on post-layout simulation.

A. Memory Access Pattern

The V2C memory and C2V memory access patterns are illustrated in Fig. 7. When a VN sends a V2C message to a CN, it also writes the V2C message to the V2C memory. The V2C message is finally read when the C2V message is returned to the VN for post-processing the C2V message. From this point on, the V2C message is no longer needed and can be overwritten.

A VN writes every C2V message to the C2V memory, and the C2V message is finally read when the VN computes the V2C message in the next iteration, when the C2V message is subtracted from the posterior LLR to compute the V2C message. From this point on, the C2V message is no longer needed and can be overwritten.

The V2C and C2V memory are continuously being written and read in the FIFO order. The data access window, defined as the duration between when the data is written to memory to the

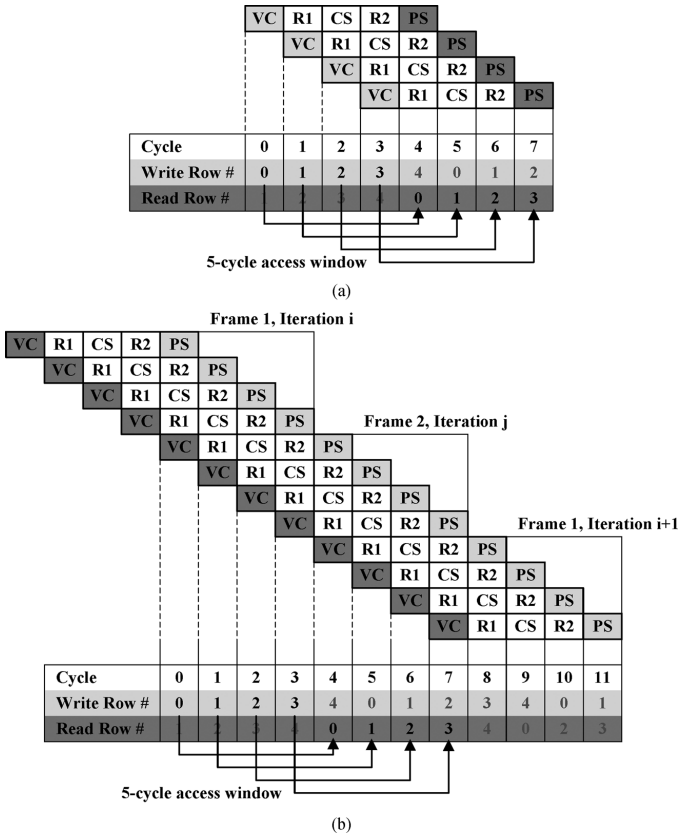


Fig. 7. (a) V2C memory access pattern, and (b) C2V memory access pattern.

last time it is read, is only 5 clock cycles. The IEEE 802.11ad standard specifies throughputs between 1.5 Gb/s and 6 Gb/s, which require clock frequencies between 90 MHz and 360 MHz using the proposed throughput-enhanced row-parallel architecture. The data access window for both the V2C memory and C2V memory is 5 clock cycles, which translates to 14 ns at 360 MHz (6 Gb/s) or 56 ns at 90 MHz (1.5 Gb/s). Therefore, the data retention time has to be at least 56 ns.

The short data access window, deterministic access order, and shallow and wide memory array structure motivate the design of a completely new low-power memory for the LDPC decoder. In the following we describe the low-power memory design to take advantage of the short data access window. The memory allows dual-port one read and one write in the same cycle to support pipelining and full-bandwidth access required by the decoder architecture.

B. Non-Refresh Embedded DRAM

Register memory found in highly parallel LDPC decoders consumes high power and occupies a large footprint. Embedded dynamic random access memory (eDRAM) [28]–[30], [35]–[37] is much smaller in size. A 3T eDRAM cell does not require a special process option. It supports nondestructive read, so it is not necessary to follow each read with write, resulting in a faster performance. The 3T eDRAM cell also supports dual-port access that is required for our application. However, eDRAM is slower than register. A periodic refresh is also necessary to compensate the leakage and maintain continuous data retention. The refresh power is a significant part of eDRAM's total power consumption.

As discussed previously, the memory for LDPC decoder has a short data access window. As long as the access window is shorter than the eDRAM data retention time, refresh can be eliminated for a significant reduction in eDRAM's power consumption, making it attractive from both area and power standpoint. A faster cell often leaks more and its data retention time has to be sacrificed. In the LDPC decoder design, the memory access pattern is well defined and the V2C and C2V memory access window is only 5 clock cycles, therefore we can consider a low-threshold-voltage (LVT) NMOS 3T eDRAM cell to provide only enough retention time, but a much higher access speed.

C. Coupling Noise Mitigation

Consider the classic 3T eDRAM cell in Fig. 8(a) for an illustration of the coupling problem. To write a 1 to the cell, the write wordline (WWL) is raised to turn on T_1 and write bitline (WBL) is driven high and the storage node will be charged up. Upon completion, WWL drops and the falling transition is coupled to the storage node through the T_1 gate-to-source capacitance, causing the storage node voltage to drop. The voltage drop results in a weak 1, reducing the data retention time and the read current. On the other hand, the coupling results in a strong 0 as the storage node will be pulled lower than ground after a write. A possible remedy is to change T_1 to a PMOS and WWL to active low to help write a strong 1, but it results in a weak 0 instead.

To mitigate the capacitive coupling and the compromise between 1 and 0, we redesign the 3T cell as in Fig. 8(b) to create capacitive coupling from two opposing directions based on [29]. Similar ideas have also been discussed in [38], [39]. Compared to [29], we use LVT NMOS transistors to improve the access speed by trading off the excess retention time. In this new design, T_2 is connected to the read wordline (RWL), which is grounded when not reading. To write to the cell, WWL is raised. WWL coupling still pulls the storage node lower after write, resulting in a weak 1 and strong 0. At the start of reading, the read bitline (RBL) is discharged to ground and RWL is raised. The rising transition of RWL is coupled to the storage node through the T_2 gate-to-drain capacitance, causing the storage node voltage to rise. The design goal is to have the positive RWL coupling cancel the negative WWL coupling. The sizing of T_1 and T_2 can be tuned to balance the coupling. Note that the focus here is on the falling WWL and rising RWL because they determine the critical read speed. Rising WWL in the beginning of write does not matter because the effect is only transient. Falling RWL in the end of read causes storage node voltage to drop, but it will be recovered when RWL rises in the beginning of the next read.

D. Retention Time Enhancement

After the cell design is finalized, we need to ensure that its data retention time is still sufficient to meet the access window required without refreshing. The data retention time of the 3T eDRAM cell is determined by the storage capacitance and the leakage currents: mainly the subthreshold leakage through the

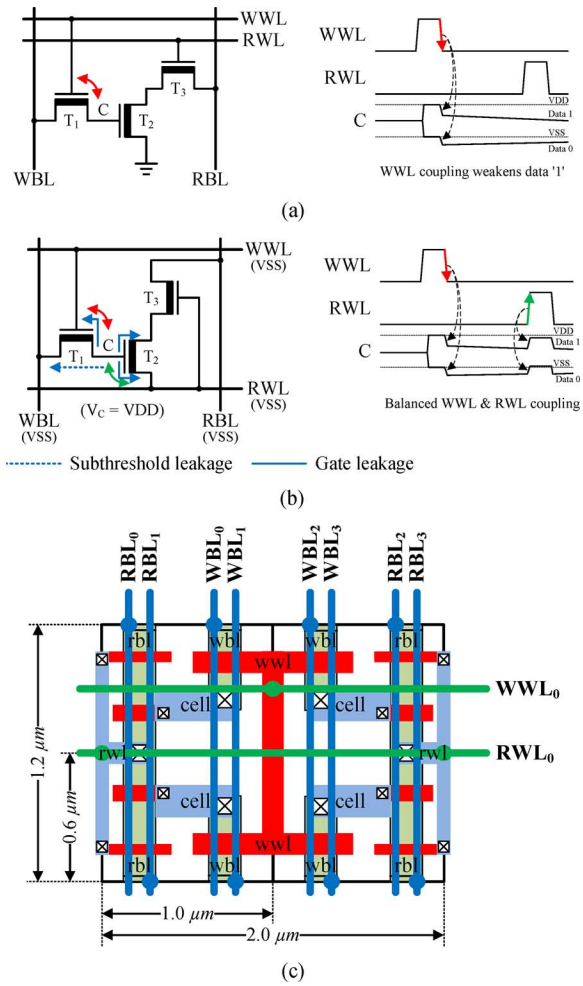


Fig. 8. Schematic and capacitive coupling illustration of the (a) classic 3T cell [27], and (b) proposed 3T cell and (c) its 4-cell macro layout.

write access transistor T_1, and the gate-oxide leakage of T_1 and the storage transistor T_2. Fig. 8(b) illustrates the leakage currents for data 1. Data 1 is more critical than data 0 as it incurs more leakage and its read is critical.

Both subthreshold and gate-oxide leakage are highly dependent on the technology and temperature. For the 65 nm CMOS process used in this design, the subthreshold leakage is dominant over gate-oxide leakage. To reduce the subthreshold leakage current, we use negative WWL voltage [35] to super cut-off T_1 after write. Fig. 9 shows the effect of negative WWL voltage on data 1 retention time at 25 °C and 125 °C. At 25 °C, the retention time improves from 100 ns to over 1 μs with a −200 mV WWL. At 125 °C, the retention time worsens to 20 ns, but it can be improved to over 1 μs with a −300 mV WWL. A 100k-point Monte-Carlo simulation is used to confirm that a −300 mV WWL is still sufficient even after considering process variation. Note that as a proof-of-concept design, the negative WWL voltage is provided from an off-chip supply. However, based on [29], charge pumps can be included to generate the negative voltage on-chip with relatively small impact on the area and power.

The proposed eDRAM design is scalable to a lower technology node. However, managing the cell leakage will be important with the continued reduction of storage capacitance. In a future process technology where leakage becomes more significant, an LVT NMOS eDRAM may not be able to provide the necessary retention time. Regular or high threshold voltage devices and a low-power process may be necessary to ensure a reliable data retention.

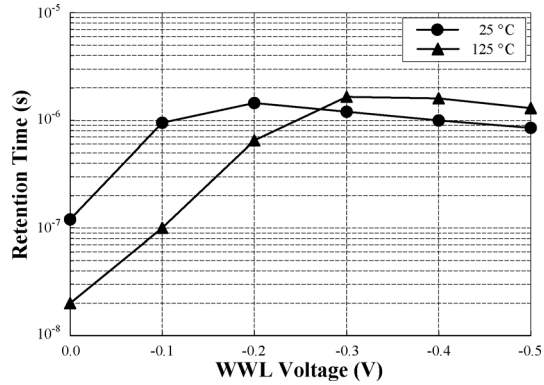


Fig. 9. Cell retention time with negative WWL voltage.

V. Efficient Memory Integration

A compact 1.0 mm × 0.6 mm layout of the 3T eDRAM cell in a 65 nm CMOS technology using standard logic design rules is shown in Fig. 8(c). The length of T_1 and T_2 are increased slightly beyond the minimum length to keep good voltage levels for storing data 0 and 1. The increased T_1 length also reduces the subthreshold leakage. The width of both T_2 and T_3 are increased slightly to improve the read speed. The two bitlines WBL and RBL are routed vertically on metal 2 and the two wordlines WWL and RWL are routed horizontally on metal 3.

An area-efficient 4-cell macro can be created in a 2 × 2 block using a bit cell, its horizontal and vertical reflections, and its 180 ° rotation, as shown in Fig. 8(c). This layout allows poly WWL and diffusion RWL to be shared between neighboring cells to reduce area. Four RBLs and four WBLs run vertically on metal 2. The 8 bitlines have fully occupied the metal 2 tracks.

A larger memory can be designed by instantiating the 4-cell macro. An illustration of a 5 row × 210 column eDRAM array for the V2C memory or C2V memory in a VNG is illustrated in Fig. 10. The array is broken to two parts to shorten the wordlines. 210 single-ended sense amplifiers [40] are attached to RBLs to provide 210 bits/cycle full-bandwidth access. The sense amplifier includes a self-reset function to save power and accommodate process variation.

The cell efficiency for the eDRAM IP is relatively low at 15% due to the shallow memory and full-bandwidth access without column multiplexing. The array efficiency can be improved for a deeper memory. Even at this array efficiency, the effective area per bit is 4.0 μm^2, much smaller than a register. The structured placement of the eDRAM cells improves the overall area utilization.

A. Sequential Address Generation

Memory address decoder is part of all standard random-access memories, but it is not necessary for the memory designed for LDPC decoder as it only requires sequential access. The

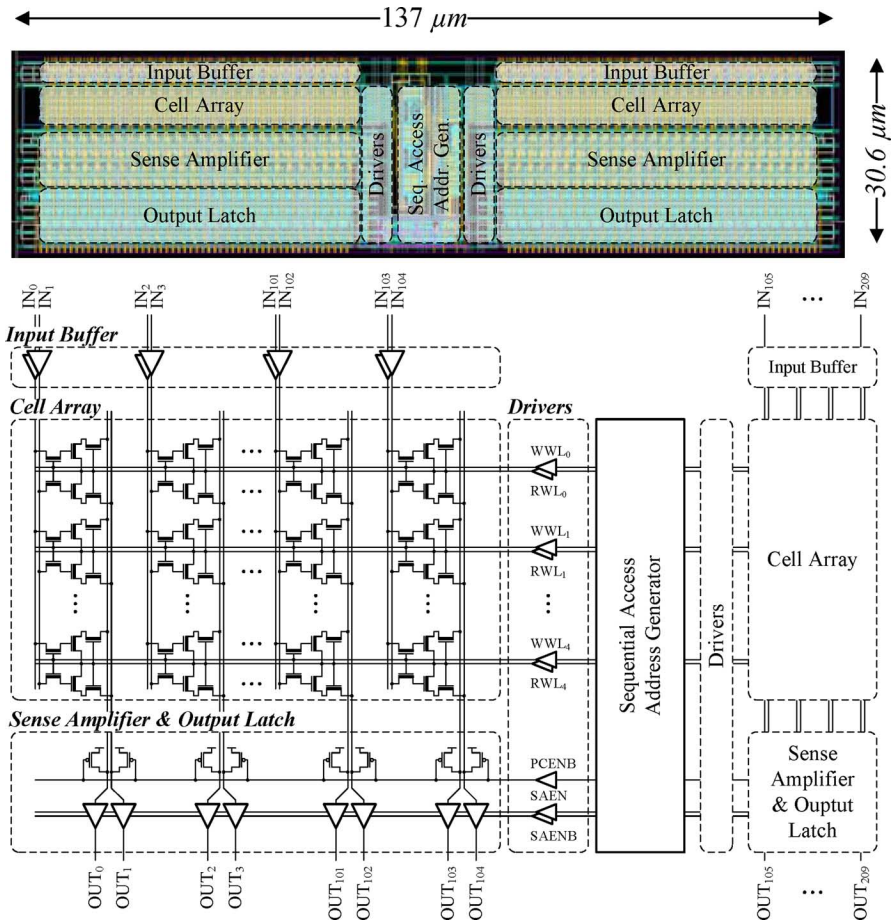


Fig. 10. Layout and schematic illustration of a 5 × 210 eDRAM array including cell array and peripherals.

memory access sequence can be understood using the multi-iteration pipeline chart in Fig. 7. For the V2C memory, in cycle 0 to cycle 3, V2C messages are written to row[0] to row[3]. Starting from cycle 4, there will be one read and one write in every cycle. In cycle 4, one V2C message is written to row[4], and another is read from row[0]. In cycle 5, one V2C message is written to row[0], and another is read from row[1], and so on.

We take advantage of the sequential access to simplify the address generation using a circular 5-stage shift register [41]. The output of each register is attached to one write enable (WE) and one read enable (RE). Only one of the registers is set to 1 in any given cycle and the 1 is propagated around the ring to enable each word serially. The simple sequential address generation saves both power and area.

B. Simulation Results

The complete 5 row × 210 column eDRAM array layout is shown in Fig. 10. The simulation results of the read access time and power consumption of the memory are plotted in Fig. 11. At the nominal supply voltage of 1.0 V and WWL voltage of −300 mV, the read access time is 0.68 ns at 25 °C. A higher temperature of 125 °C decreases the read access time to 0.57 ns, due to the increasing leakage of the sense amplifier that accelerates the charging of the bitline. This effect on read access time becomes more significant when the supply voltage is lowered.

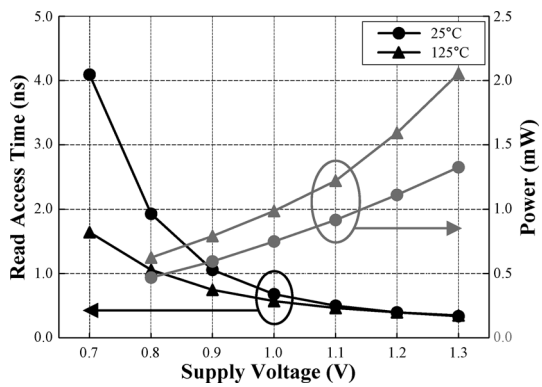


Fig. 11. Simulated read access time (in black) and power consumption (in grey) of the eDRAM array at 25 °C and 125 °C. Results are based on post-layout simulation using a −300 mV WWL and power is measured at a 180 MHz clock frequency.

At 0.7 V, the read access time is 4.1 ns at 25 °C and 1.6 ns at 125 °C.

The IEEE 802.11ad LDPC decoder requires 32 5×210 eDRAM arrays, two for each of the 16 VNGs as V2C memory and C2V memory. To achieve the highest required throughput of 6 Gb/s, the clock period is set to 2.8 ns, and the memory supply voltage has to be set to about 0.9 V.

TABLE I
MEASUREMENT SUMMARY OF THE LDPC DECODER AT 5.0 dB SNR AND 10 DECODING ITERATIONS

Frequency (MHz)		30	60	90	180	270	360	450	540
Core	Supply (V)	0.41	0.45	0.51	0.64	0.76	0.94	1.06	1.15
	Power (mW)	5.6	11.0	21.0	68.2	142.8	285.8	480.1	620.1
eDRAM	Supply (V)	0.69	0.73	0.80	0.92	1.03	1.11	1.22	1.30
	Power (mW)	6.2	10.2	16.7	37.6	64.8	87.8	130.8	162.8
Total Power (mW)		11.8	21.2	37.7	105.8	207.6	373.6	610.9	782.9
eDRAM Fraction (%)		52	48	44	36	31	23	21	21
Throughput (Gb/s)		0.5	1.0	1.5	3.0	4.5	6.0	7.5	9.0
Energy Efficiency (pJ/bit)		21.0	21.9	35.6	34.5	44.8	61.7	76.4	89.5
Area Efficiency (Gb/s/mm^2)		0.31	0.63	0.94	1.88	2.81	3.75	4.69	5.63

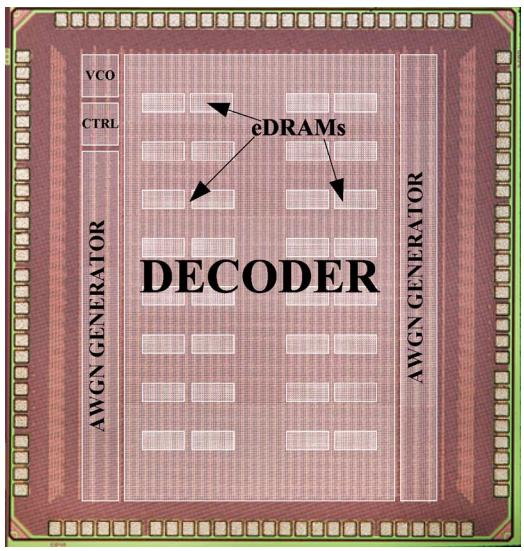


Fig. 12. Chip microphotograph. Locations of the 32 eDRAM arrays inside the LDPC decoder and the testing peripherals are labeled.

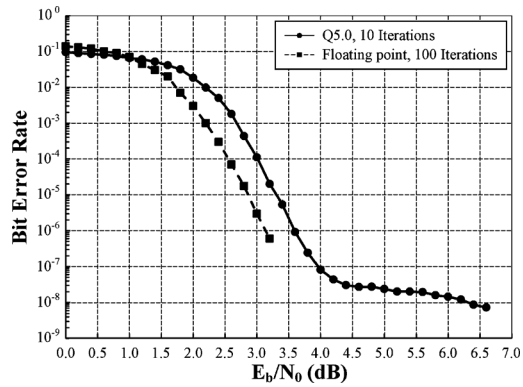


Fig. 13. Bit error rate performance of the rate-1/2 LDPC code of the IEEE 802.11ad standard using a 5-bit quantization with 10 decoding iterations and floating point with 100 iterations.

VI. DECODER CHIP IMPLEMENTATION AND MEASUREMENTS

A decoder test chip was implemented in a TSMC 65 nm 9-metal general-purpose CMOS technology [42]. It was designed as a proof-of-concept to support the rate-1/2 (672, 336) LDPC code for the IEEE 802.11ad standard, but the architecture also accommodates the three higher rate codes. The chip microphotograph is shown in Fig. 12. The test chip measures 1.94 mm$\times$1.84 mm and the core measures 1.6 mm$\times$1.0 mm including 32 5$\times$210 eDRAM arrays.

The decoder test chip uses separate supply voltages for the decoder core logic and eDRAM memory arrays to allow each supply voltage to be independently set to achieve the throughput targets with the lowest power. Clock is generated on-chip, and it can also be provided through an external source. The decoder incorporates AWGN generators to model the communication channel and provide input vectors in real time. Decoding errors are collected on-chip to compute the bit error rate (BER) and frame error rate (FER).

The decoder supports two test modes: a scan mode for debugging and an automated mode for gathering error statistics. In the scan mode, input vectors are fed through scan chains and the decoding decisions are scanned out for inspection. In the automated mode, the decoder takes inputs from the on-chip AWGN generators, and decoding decisions are checked on-chip for errors. The AWGN noise variance and scaling factors are tuned to provide a range of signal-to-noise ratio (SNR). We step through a number of SNR points and collect sufficient error statistics to plot BER against SNR waterfall curves. The waterfall curves are checked against the reference waterfall curve obtained by software simulation.

A. Chip Measurements

The test chip operates over a wide range of clock frequencies from 30 MHz up to 540 MHz, which translate to a throughput from 0.5 Gb/s up to 9 Gb/s using a fixed 10 decoding iterations. Early termination is built-in to increase throughput at high SNR if needed. The decoder BER is shown in Fig. 13. An excellent error-correction performance is achieved down to a BER of 10^{-7}, which is sufficient for the application.

Fig. 14 shows the measured power consumption of the decoder chip, the core and the eDRAM arrays at each clock frequency. The decoder consumes 38 mW, 106 mW, and 374 mW to achieve a throughput of 1.5 Gb/s, 3 Gb/s, and 6 Gb/s, respectively, at the optimal core and memory supply voltages listed in Table I. The power consumption of the non-refresh eDRAM increases almost linearly with frequency compared to the quadratic increase in core logic power, 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 power consumption over the SNR range of interest is shown in Fig. 15.

TABLE II
COMPARISON OF STATE-OF-THE-ART LDPC DECODERS

	This Work			JSSC'12 [11]	JSSC'11 [10]		JSSC'10 [12]		ASSCC'11 [25]	ASSCC'10 [24]	ASSCC'10 [16]	
Technology	65nm			65nm	130nm		65nm		65nm	90nm	90nm	
Block Length	672			672	576-2304		2048		576-2304	648-1944	2048	
Code Rate	1/2			1/2-7/8	1/2-5/6		0.84		1/2-5/6	1/2-5/6	0.84	
Decoding Algorithm	Offset Min-Sum			Layered Normalized Min-Sum	Layered Normalized Min-Sum		Offset Min-Sum		Layered Offset Min-Sum	Layered Offset Min-Sum	Layered Offset Min-Sum	
Core Area (mm^2)	1.60			1.56	3.03		5.35		3.36	1.77	5.35	
Iterations	10			5	10		8		10	10	4	
Input Quantization (bit)	5			6	6		4		6	5	7	
Core Supply (V)	0.41	0.94	1.15	1.0	1.2		0.7	1.2	1.2	1.0	0.8	1.2
Memory Supply (V)	0.69	1.11	1.30									
Clock Frequency (MHz)	30	360	540	197	214		100	700	110	346	84.7	137
Throughput (Gb/s)	0.5	6.0	9.0	5.79	0.874	0.955	6.67[a]	47.7[a]	1.056	0.679	7.23	11.69
Power (mW)	11.8	373.6	782.9	361	342	397	144	2800	115	107.4	386.8	1559
Norm. Throughput (Gb/s)[b]	0.5	6.0	9.0	5.79	1.748	1.91	2.335[c]	16.695[c]	2.112	1.36	5.784	9.352
Norm. Energy Eff. (pJ/bit)[d]	21.0	61.7	89.5	62.4	195.7	207.9	61.7	167.7	54.9	79	66.9	166.7
Norm. Area Eff. (Gb/s/mm^2)[d]	0.31	3.75	5.63	3.70	0.58	0.63	0.44	3.12	0.63	0.77	1.08	1.75

[a] Early termination enabled.
[b] Throughput is normalized to 10 decoding iterations for flooding decoders and 5 decoding iterations for layered decoders.
[c] Early termination requires an average of 2.5 iterations at a 5.5dB SNR. One additional iteration is needed for convergence detection. [12]
[d] Energy and area efficiency are computed based on the normalized throughput.

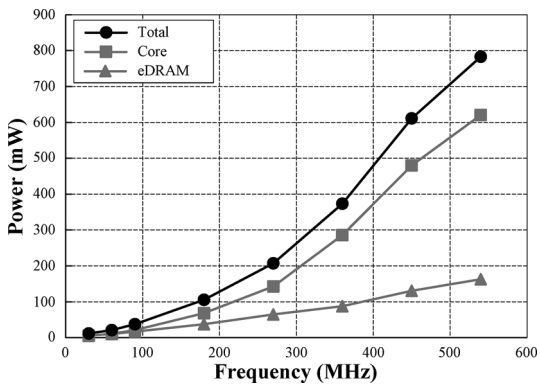


Fig. 14. Measured LDPC decoder power at 5.0 dB SNR and 10 decoding iterations. The total power is divided into core and eDRAM power. Voltage scaling is used for the optimal core and eDRAM power.

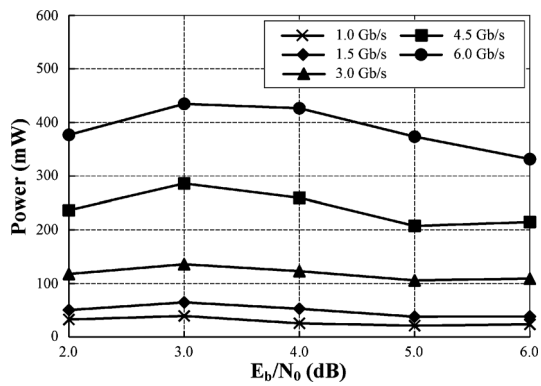


Fig. 15. Measured LDPC decoder power across SNR range of interest at 10 decoding iterations. Voltage scaling is used for optimal core and eDRAM power.

The power is the highest when the decoder is operating near the middle of the waterfall region, a result of high switching activities. The power decreases in the high SNR region due to the improved channel condition that leads to fewer switching activities.

B. Comparison With State-of-the-Art

The three metrics of an LDPC decoder implementation are throughput, power and silicon area. Two efficiency measures can be derived based on the three metrics: power/throughput (in pJ/b) gives energy efficiency, and throughput/area (in b/s/mm^2) gives area efficiency. Table II summarizes the results of the test chip along with other state-of-the-art LDPC decoders published in the last three years. For a fair comparison, we normalize the throughput to 10 iterations for a flooding decoder and 5 iterations for a layered decoder that converges faster.

As Table II shows, our results have advanced the state of the art by improving the best energy efficiency to 21 pJ/b in the low power mode and the best area efficiency to 5.63 Gb/s/mm^2 in the high performance mode. We provide a range of operating points in Table I to show the tradeoff space between energy efficiency and area efficiency.

VII. Conclusion

We present a low-power logic-compatible eDRAM design for a high-throughput LDPC decoder. The eDRAM retains storage for the necessary data access window, eliminating refresh for a significant power reduction. A new 3T LVT NMOS eDRAM cell design trades off the excessive retention time for a fast 0.68 ns read access at 1.0 V. To ensure a reliable storage, the coupling noise is mitigated by balancing the write and read wordline coupling, and the subthreshold leakage is minimized by a negative write wordline.

A row-parallel LDPC decoder is designed using 32 5×210 non-refresh eDRAM arrays for the (672, 336) LDPC code suitable for the IEEE 802.11ad standard. We use row merging and dual-frame processing to increase hardware utilization and remove pipeline stalls, resulting in a significant reduction of the clock frequency from 1.07 GHz to 360 MHz. The 1.6 mm^2 65 nm LDPC decoder test chip achieves a peak throughput of 9 Gb/s at 89.5 pJ/b, of which only 21% is spent on eDRAMs. With voltage and frequency scaling, the energy efficiency is improved to 35.6 pJ/b for a 1.5 Gb/s throughput.

Acknowledgment

The authors would like to thank Y. Lee for advice on designing the eDRAM, and E. Yeo and P. Urard for suggestions on the LDPC decoder design.

References

[1] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA, USA: MIT Press, 1963.

[2] D. J. C. Mackay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 399–431, Mar. 1999.

[3] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[4] *IEEE Draft Standard for Information Technology-Telecommunications and Information Exchange between Systems-Local and Metropolitan Area Networks-Specific Requirements, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment: Enhancements for Higher Throughput*, IEEE Std. 802.11n/D2.00, Feb. 2007.

[5] *IEEE Standard for Local and Metropolitan Area Networks: Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1*, IEEE 802.16e, Feb. 2006.

[6] *ETSI Standard TR 102 376 V1.1.1: Digital Video Broadcasting (DVB) User Guidelines for the Second Generation System for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications (DVB-S2)*, ETSI Std. TR 102 376, Feb. 2005.

[7] *IEEE Standard for Information Technology-Telecommunications and Information Exchange between Systems-Local and Metropolitan Area Networks-Specific Requirements, Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, IEEE Std. 802.3an, Sep. 2006.

[8] A. Kavčić and A. Patapoutian, "The read channel," *Proc. IEEE*, vol. 96, no. 11, pp. 1761–1774, Nov. 2008.

[9] G. Dong, N. Xie, and T. Zhang, "On the use of soft-decision error-correction codes in NAND flash memory," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 58, no. 2, pp. 429–439, Feb. 2011.

[10] B. Xiang, D. Bao, S. Huang, and X. Zeng, "An 847–955 Mb/s 342–397 mW dual-path fully-overlapped QC-LDPC decoder for WiMAX system in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1416–1432, Jun. 2011.

[11] S.-W. Yen, S.-Y. Hung, C.-H. Chen, H.-C. Chang, S.-J. Jou, and C.-Y. Lee, "A 5.79-Gb/s energy-efficient multirate LDPC codec chip for IEEE 802.15.3c applications," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 2246–2256, Sep. 2012.

[12] Z. Zhang, V. Anantharam, M. J. Wainwright, and B. Nikolić, "An efficient 10GBASE-T wthernet LDPC decoder design with low error floors," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 843–855, Apr. 2010.

[13] A. J. Blanksby and C. J. Howland, "A 690-mW 1-Gb/s 1024-b, rate-1/2 low-density parity-check code decoder," *IEEE J. Solid-State Circuits*, vol. 37, no. 3, pp. 404–412, Mar. 2002.

[14] A. Darabiha, A. C. Carusone, and F. R. Kschischang, "Power reduction techniques for LDPC decoders," *IEEE J. Solid-State Circuits*, vol. 43, no. 8, pp. 1835–1845, Aug. 2008.

[15] Z. Zhang, V. Anantharam, M. J. Wainwright, and B. Nikolić, "A 47 Gb/s LDPC decoder with improved low error rate performance," in *IEEE Symp. VLSI Circuits Dig.*, Kyoto, Japan, Jun. 2009, pp. 286–287.

[16] A. Cevrero, Y. Leblebici, P. Ienne, and A. Burg, "A 5.35 mm^2 10GBASE-T Ethernet LDPC decoder chip in 90 nm CMOS," in *Proc. 2010 IEEE Asian Solid-State Circuits Conf.*, Beijing, China, Nov. 2010, pp. 317–320.

[17] M. Korb and T. G. Noll, "Area- and energy-efficient high-throughput LDPC decoders with low block latency," in *Proc. IEEE Eur. Solid-State Circuits Conf., ESSCIRC'11*, Helsinki, Finland, Sep. 2011, pp. 75–78.

[18] P. Urard, L. Paumier, V. Heinrich, N. Raina, and N. Chawla, "A 360 mW 105 Mb/s DVB-S2 compliant codec based on 64800 b LDPC and BCH codes enabling satellite-transmission portable devices," in *2008 IEEE Int. Solid-State Circuits Conf. Dig.*, San Francisco, CA, USA, Feb. 2008, pp. 310–311.

[19] M. M. Mansour and N. R. Shanbhag, "A 640-Mb/s 2048-bit programmable LDPC decoder chip," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 684–698, Mar. 2006.

[20] X.-Y. Shi, C.-Z. Zhan, C.-H. Lin, and A.-Y. Wu, "An 8.29 mm^2 52 mW multi-mode LDPC decoder design for mobile WiMAX system in 0.13 μm CMOS process," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 672–683, Mar. 2008.

[21] C.-H. Liu, S.-W. Yen, C.-L. Chen, H.-C. Chang, C.-Y. Lee, Y.-S. Hsu, and S.-J. Jou, "An LDPC decoder chip based on self-routing network for IEEE 802.16e applications," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 684–694, Mar. 2008.

[22] C.-L. Chen, K.-S. Lin, H.-C. Chang, W.-C. Fang, and C.-Y. Lee, "A 11.5-Gbps LDPC decoder based on CP-PEG code construction," in *Proc. IEEE Eur. Solid-State Circuits Conf., ESSCIRC'09*, Athens, Greece, Sep. 2009, pp. 412–415.

[23] F. Naessens, V. Derudder, H. Cappelle, L. Hollevoet, P. Raghavan, M. Desmet, A. M. AbdelHamid, I. Vos, L. Folens, S. O'Loughlin, S. Singirikonda, S. Dupont, J.-W. Weijers, A. Dejonghe, and L. V. der Perre, "A 10.37 mm^2 675 mW reconfigurable LDPC and turbo encoder and decoder for 802.11n, 802.16e and 3GPP-LTE," in *IEEE Symp. VLSI Circuits Dig.*, Honolulu, HI, USA, Jun. 2010, pp. 213–214.

[24] C. Roth, P. Meinerzhagen, C. Studer, and A. Burg, "A 15.8 pJ/bit/iter quasi-cyclic LDPC decoder for IEEE 802.11n in 90 nm CMOS," in *IEEE Asian Solid-State Circuits Conf.*, Beijing, China, Nov. 2010, pp. 313–316.

[25] X. Peng, Z. Chen, X. Zhao, D. Zhou, and S. Goto, "A 115 mW 1 Gbps QC-LDPC decoder ASIC for WiMAX in 65 nm CMOS," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Jeju, Korea, Nov. 2011, pp. 317–320.

[26] M. Weiner, B. Nikolić, and Z. Zhang, "LDPC decoder architecture for high-data rate personal-area networks," in *Proc. IEEE Int. Symp. Circuits and Systems*, Rio de Janeiro, Brazil, May 2011, pp. 1784–1787.

[27] W. M. Regitz and J. A. Karp, "Three-transistor-cell 1024-bit 500-ns MOS RAM," *IEEE J. Solid-State Circuits*, vol. SC-5, no. 5, pp. 181–186, Oct. 1970.

[28] D. Somasekhar, Y. D. Ye, P. Aseron, S.-L. Lu, M. M. Khellah, G. R. J. Howard, T. Karnik, S. Borkar, V. K. De, and A. Keshavarzi, "2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 174–185, Jan. 2009.

[29] K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, Jun. 2011.

[30] K. C. Chun, P. Jain, T.-H. Kim, and C. H. Kim, "A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches," *IEEE J. Solid-State Circuits*, vol. 47, no. 2, pp. 547–559, Feb. 2012.

[31] *IEEE Standard for Information Technology-Telecommunications and Information Exchange between Systems-Local and Metropolitan Area Networks-Specific Requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band*, IEEE Std. 802.11ad, Dec. 2012.

[32] M. P. C. Fossorier, M. Mihaljević, and H. Imai, "Reduced complexity iterative decoding of low-density parity check codes based on belief propagation," *IEEE Trans. Commun.*, vol. 47, no. 5, pp. 673–680, May 1999.

[33] J. Chen, A. Dholakia, E. Eleftheriou, M. P. C. Fossorier, and X.-Y. Hu, "Reduced-complexity decoding of LDPC codes," *IEEE Trans. Commun.*, vol. 53, no. 8, pp. 1288–1299, Aug. 2005.

[34] C. Roth, A. Cevrero, C. Studer, Y. Leblebici, and A. Burg, "Area, throughput, and energy-efficiency trade-offs in the VLSI implementation of LDPC decoders," in *Proc. IEEE Int. Symp. Circuits and Systems*, Rio de Janeiro, Brazil, May 2011, pp. 1772–1775.

[35] J. Barth, W. R. Reohr, P. Parries, G. Fredeman, J. Golz, S. E. Schuster, R. E. Matick, H. Hunter, C. C. Tanner, J. Harig, H. Kim, B. Khan, J. Griesemer, R. P. Havreluk, K. Yanagisawa, T. Kirihata, and S. S. Iyer, "A 500 MHz random cycle, 1.5 ns latency, SOI embedded DRAM macro featuring a three-transistor micro sense amplifier," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 86–95, Jan. 2008.

[36] P. J. Klim, J. Barth, W. R. Reohr, D. Dick, G. Fredeman, G. Koch, H. M. Le, A. Khargonekar, P. Wilcox, J. Golz, J. B. Kuang, A. Mathews, J. C. Law, T. Luong, H. C. Ngo, R. Freese, H. C. Hunter, E. Nelson, P. Parries, T. Kirihata, and S. S. Iyer, "A 1 MB cache subsystem prototype with 1.8 ns embedded DRAMs in 45 nm SOI CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1216–1226, Apr. 2009.

[37] J. Barth, D. Plass, E. Nelson, C. Hwang, G. Fredeman, M. Sperling, A. Mathews, T. Kirihata, W. R. Reohr, K. Nair, and N. Cao, "A 45 nm SOI embedded DRAM macro for the POWER™ processor 32 MByte on-chip L3 cache," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 64–75, Jan. 2011.

[38] W. K. Luk, J. Cai, R. H. Dennard, M. J. Immediato, and S. V. Kosonocky, "A 3-transistor DRAM cell with gated diode for enhanced speed and retention time," in *IEEE Symp. VLSI Circuits Dig.*, Honolulu, HI, USA, Jun. 2006, pp. 184–185.

[39] P. Meinerzhagen, A. Teman, R. Giterman, A. Burg, and A. Fish, "Exploration of sub-VT and near-VT 2T gain-cell memories for ultra-low power applications under technology scaling," *J. Low Power Electron. Applicat.*, vol. 3, no. 2, pp. 54–72, Apr. 2013.

[40] S. Satpathy, Z. Foo, B. Giridhar, R. Dreslinski, D. Sylvester, T. Mudge, and D. Blaauw, "A 1.07 Tbit/s 128 × 128 swizzle network for SIMD processors," in *IEEE Symp. VLSI Circuits Dig.*, Honolulu, HI, USA, Jun. 2010, pp. 81–82.

[41] S.-M. Yoo, J. M. Han, E. Haq, S. S. Yoon, S.-J. Jeong, B. C. Kim, J.-H. Lee, T.-S. Jang, H.-D. Kim, C. J. Park, D. I. Seo, C. S. Choi, S.-I. Cho, and C. G. Hwang, "A 256M DRAM with simplified register control for low power self refresh and rapid burn-in," in *IEEE Symp. VLSI Circuits Dig.*, Honolulu, HI, USA, Jun. 1994, pp. 85–86.

[42] Y. S. Park, D. Blaauw, D. Sylvester, and Z. Zhang, "A 1.6-mm^2 38-mW 1.5-Gb/s LDPC decoder enabled by refresh-free embedded DRAM," in *IEEE Symp. VLSI Circuits Dig.*, Honolulu, HI, USA, Jun. 2012, pp. 114–115.

Youn Sung Park (S'10) received the B.A.Sc. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2008, and the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Michigan, Ann Arbor, where he is a member of the Michigan Integrated Circuits Laboratory (MICL).

His research interest is in the design of energy-efficient digital signal processors through algorithm, architecture, and circuit co-optimization.

David Blaauw (M'94–SM'07–F'12) received the B.S. degree in physics and computer science from Duke University, Durham, NC, USA, in 1986, and the Ph.D. degree in computer science from the University of Illinois, Urbana, IL, USA, in 1991.

Until August 2001, he worked for Motorola, Inc., Austin, TX, USA, where he was the manager of the High Performance Design Technology group. Since August 2001, he has been on the faculty at the University of Michigan where he is a Professor. He has published over 400 papers and holds 40 patents. His work has focused on VLSI design with particular emphasis on ultra-low-power and high-performance design.

Prof. Blaauw was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronic and Design. He was also the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference and a member of the ISSCC Technical Program Committee.

Dennis Sylvester (S'95–M'00–SM'04–F'11) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, CA, USA, where his dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC Berkeley EECS department.

He is a Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, MI, USA, and Director of the Michigan Integrated Circuits Laboratory (MICL), a group of ten faculty and 70+ graduate students. He has held research staff positions in the Advanced Technology Group of Synopsys in Mountain View, CA, Hewlett-Packard Laboratories in Palo Alto, CA, and a visiting professorship at the National University of Singapore. He has published over 350 articles along with one book and several book chapters. His research interests include the design of millimeter-scale computing systems and energy efficient near-threshold computing. He holds 19 US patents. He also serves as a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas. He co-founded Ambiq Micro, a fabless semiconductor company developing ultra-low power mixed-signal solutions for compact wireless devices.

Dr. Sylvester has received an NSF CAREER Award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and eight best paper awards and nominations. He was a recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of major design automation and circuit design conferences, the executive committee of the ACM/IEEE Design Automation Conference, and the steering committee of the ACM/IEEE International Symposium on Physical Design. He has served as Associate Editor for IEEE TRANSACTIONS ON CAD and IEEE TRANSACTIONS ON VLSI SYSTEMS, and as Guest Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II.

Zhengya Zhang (S'02–M'09) received the B.A.Sc. degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2003, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, CA, USA, in 2005 and 2009, respectively.

Since 2009, he has been on the faculty of the University of Michigan, Ann Arbor, MI, USA, as an Assistant Professor in the Department of Electrical Engineering and Computer Science. His current research interests include low-power and high-performance VLSI circuits and systems for computing, communications and signal processing.

Dr. Zhang was a recipient of the National Science Foundation CAREER Award in 2011, the Intel Early Career Faculty Honor Program Award in 2013, the David J. Sakrison Memorial Prize for outstanding doctoral research in EECS at UC Berkeley, and the Best Student Paper Award at the Symposium on VLSI Circuits. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS.