Energy- and Area-Efficient Parameterized Lifting-Based 2-D DWT Architecture on FPGA

Yusong Hu[†‡] and Viktor K. Prasanna[‡]

‡ Ming Hsieh Department of Electrical Engineering[†]
University of Southern California
Los Angeles, USA, 90089

†School of Electrical and Electronic Engineering[*]
Nanyang Technological University
Singapore, 639798

Email: {ysh_055, prasanna}@usc.edu

Abstract—**State-of-the-art DWT designs focus on improving hardware utilization and memory efficiency of DWT. In this paper, we consider energy efficiency as the key performance metric. Memory (external memory and on-chip memory) energy dominates the total energy consumption. We propose a DWT architecture with an overlapped block-based image scanning method that optimizes the number of external memory accesses and the on-chip memory size. Using the overlapped block-based scanning method, the required number of external memory accesses of the proposed architecture is reduced by up to 50% when compared with state-of-the-art designs. Besides, the on-chip memory size is also reduced. We implement the proposed architecture on a state-of-the-art FPGA for various image sizes. Our design sustains up to 80.2% of the peak energy efficiency of the device. Compared with the state-of-the-art design, the proposed architecture achieves up to 58.1% energy efficiency improvement.**

Keywords—Discrete wavelet transform (DWT), FPGA architecture, energy efficiency.

I. Introduction

The discrete wavelet transform (DWT) is widely used in image analysis, signal processing and computer graphics since its introduction in [1]. The JPEG 2000 standard [2] has chosen DWT instead of the block discrete cosine transform due to its good energy compaction and inherent multi-resolution image presentation ability. The state-of-the-art FPGAs have become an attractive choice for implementing signal processing applications such as DWT, as they offer unprecedented logic density and high customizability.

Recently, energy efficiency has emerged as one of the most important performance metrics in computing [3]. Although the DWT architecture has been extensively studied, most existing works focus on improving the hardware utilization and reducing the on-chip memory size and critical path length. To the best of our knowledge, energy efficiency has not been considered as a key performance metric. In this bwork, we observe that the external memory and on-chip memory dominates the total energy consumption. By employing a block-based image scanning method, we significantly reduce the required number of external memory accesses without increasing the on-chip memory size. As a result, the energy consumption of the proposed architecture is significantly reduced. The proposed architecture is also parameterized and achieves high memory efficiency. While the on-chip memory size of the existing designs depends on the input image width or height, the proposed architecture consumes only a constant number of on-chip memory.

In this paper, we make the following contributions:

1) An overlapped block-based image scanning method to improve the energy efficiency by optimizing the number of external memory accesses and on-chip memory size (Section III.B).
2) An energy-efficient parameterized DWT architecture for the lifting 5/3 and 9/7 filter (Section III and IV).
3) A DRAM activation schedule to reduce the energy consumption of external memory (Section III.C).
4) Optimized design achieving up to 58.1% improvement in energy efficiency compared with the state-of-the-art design (Section IV.C).

The rest of the paper is organized as follows. Section II covers the background and related work. Section III introduces the block-based image scanning method and DWT architecture on FPGA. Section IV presents the experimental results and analyzes the energy efficiency of the proposed DWT architecture. Section V concludes the paper.

II. Background and Related Work

A. Lifting scheme

The existing DWT architectures can be classified into two categories, namely convolution-based and lifting-based [5]. Compared with the convolution-based architecture, the

[†]This work has been funded by DARPA under grant number HR0011-12-2-0023 and the U.S. National Science Foundation under grant CCF-1320211. Equipment grant from Xilinx, Inc. is gratefully acknowledged.
*This work is supported by Nanyang Technological University under a Research Scholarship.

978-1-4799-6233-4/14/$31.00 ©2014 IEEE

lifting-based architecture has several advantages with respect to energy efficiency, such as lower computation complexity and memory-efficient in-place computation [4]. Lifting-based 2-D DWT can be decomposed into two steps, namely row-wise DWT (rDWT) and column-wise DWT (cDWT). The rDWT for the 9/7 filter can be represented by four lifting steps as below.

$$d_h(m,n)=x(m,2n+1)+\alpha(x(m,2n)+x(m,2n+2)) \quad (1)$$

$$d_l(m,n)=x(m,2n)+\beta(d_h(m,n-1)+d_h(m,n)) \quad (2)$$

$$H(m,n)=d_h(m,n)+\gamma(d_l(m,n-1)+d_l(m,n)) \quad (3)$$

$$L(m,n)=d_l(m,n)+\delta(H(m,n-1)+H(m,n)) \quad (4)$$

where x, H and L represent the input image, the high-pass intermediate results and the low-pass intermediate results, respectively. d_h and d_l are the partial results (H is also regarded as partial results). α, β, γ and δ are the lifting coefficients. We assume M and N are the width and height of the input image, respectively ($0 \le n \le N-1$ and $0 \le m \le M-1$). The cDWT has the similar formulation as rDWT. The H and L are used as input to the cDWT to generate four subbands, namely HH, HL, LH and LL. The LL and HH subbands are scaled by the scaling factors K^2, $1/K^2$, respectively. The 5/3 DWT filter has a similar formation to the 9/7 DWT filter except that it requires only two lifting steps and no scaling step.

B. Related work

To the best of our knowledge, none of the existing DWT architectures use energy efficiency as their main performance metric. Most of the existing works focus on reducing the on-chip memory size. The existing DWT architectures can be classified according to the image scanning methods employed.

The line-based architectures [5]-[8] read the image in line-by-line order. A high-throughput line-based architecture for multi-level DWT is proposed in [5], with a transposition memory of length 2.5M and a temporal memory of length 3M for an image of size MN. The designs proposed in [6], [7] and [8] scan two row concurrently so that to facilitate the reuse of intermediate results. Consequently, the transposition memory is eliminated whereas the temporal memory size is 4M. In [9], an architecture based on the block-based scanning method is proposed. The input image is partitioned into blocks and scanned in a line-by-line order within each block. Although the throughput of the architecture is high, it is a convolution-based architecture and requires a large number of arithmetic resources.

In order to reduce the on-chip memory size, an overlapped scanning method is introduced, which reads some pixels multiple times. An overlapped stripe-based scanning method is proposed in [10] to explore the tradeoff between the number of external memory reads and on-chip memory size. However, within each stripe, the image is scanned in a line-by-line order and therefore a large transposition memory is required. Multi-level DWT architectures based on the overlapped stripe-based scanning method are proposed in [12] and [13], where the temporal memory of the first level DWT is eliminated to achieve a high memory efficiency. On the other hand, the external memory bandwidth is significantly increased. The state-of-the-art architecture for single-level DWT is presented in [11], where the temporal memory size is reduced to 3N at the expense of one pixel overlap between two adjacent stripes.

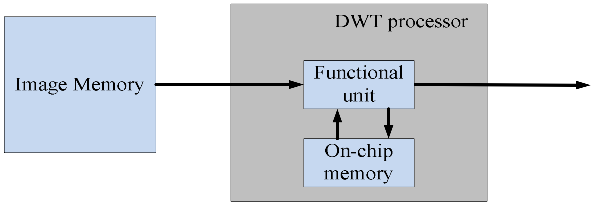

Fig. 1 High level abstraction of DWT architecture

Although the on-chip memory size can be reduced by overlapped scanning methods, it incurs significant external memory energy overhead. Compared with the existing designs, we focus on improving the energy efficiency of DWT architecture. With the overlapped block-based scanning method, we improve the energy efficiency by reducing the number of external memory accesses and the on-chip memory size.

III. Energy-efficient Parameterized 2-D DWT Architecture

A. Design assumptions

DWT architecture: As shown in Fig. 1, the DWT architecture is composed of an image memory (external memory) and a DWT processor. The image memory outputs the input image to the DWT processor on FPGA, which performs DWT to the input image. On-chip memory refers to the employed memory resources on FPGA.

Operations: There are two types of operations, namely computation operations and memory-access operations. A computation operation for DWT architecture is defined as one multiplication. A memory operation is defined as a read/write operation to the external memory or the on-chip memory. For image of size MN, a minimum of 2MN and 4.5MN computation operations are required for 5/3 and 9/7 filters, respectively. At least MN image memory read operations are required for both the 5/3 and 9/7 filter.

Algorithm-mapping parameters: Two algorithm-mapping parameters are introduced to characterize the DWT architecture, namely parallelism (L) and height of block (R). The parallelism is defined as the number of computation operations performed every clock cycle, while the height of block is defined as the height of each image block.

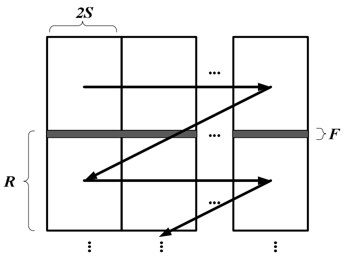

Fig. 2 Overlapped block-based image scanning method

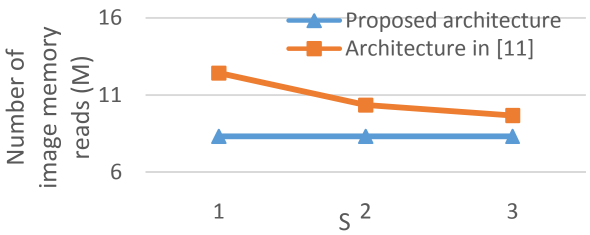

Fig. 3 Number of image memory accesses (5/3 filter, input size of 2160×3840, R=543)

Energy efficiency: Energy efficiency is defined as the number of operations that can be performed by a unit of energy (GOPS/W). It is a widely accepted performance metric for computing [3]. In this paper, only the dynamic power is considered. Let the minimum required computation be C, the energy consumed by processing one frame be E, the average power of the convolution architecture be P and the processing time per frame be t. Therefore, the energy efficiency η is obtained by:

$$\eta = C / E = C / Pt \tag{5}$$

Energy×Area×Time (EAT): EAT is defined as the product of the energy consumption per frame, the area usage of the architecture and the computation time per frame [14]. We use EAT to evaluate the impact of energy efficiency on area and throughput. We estimate the area of the design based on the resource utilization, i.e. the number of LUTs or flip-flops (select the larger one) used by the design. The BRAM blocks are transferred to certain amount of LUT slices according to the memory size [14].

Peak energy efficiency: We introduce peak energy efficiency to evaluate if an architecture is well optimized with respect to energy efficiency. The energy efficiency of a specific target device is upper bounded by the peak energy efficiency. To compute the peak energy efficiency of an algorithm on a given target device, we consider the energy consumed to perform the minimum number of required operations (e.g. computation and memory accesses) while ignoring all the overheads such as I/O, control logic, routing and on-chip buffers that may be employed by an implementation. For 2-D DWT, assume the average power dissipation of the required computation and external memory on a given target device are P_c and P_i, respectively. The peak energy efficiency η_{peak} is obtained by:

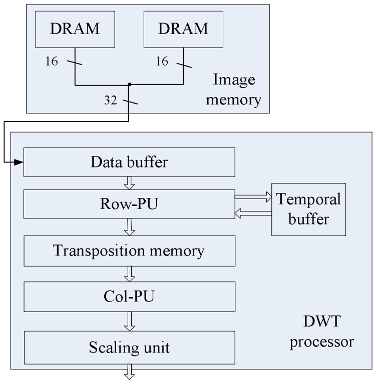

Fig. 4 Proposed DWT processor

$$\eta_{peak} = C / (P_c + P_i)\ t \tag{6}$$

The peak energy efficiency of the target device considered in this paragraph is measured in Section IV.B.

B. Overlapped block-based scanning method

The overlapped block-based scanning method is shown in Fig. 2. The input image is partitioned into image blocks of width $2S$ and height R as presented by the rectangular with thick border. There are F rows of overlap (gray stripe in Fig. 2) between vertically adjacent blocks. F = 3 and 7 for the 5/3 and the 9/7 DWT filter, respectively. The blocks are processed one-by-one in a line-based order from left to right. After a line of the blocks is processed, the process continues from the next line of blocks. Within each block, the pixels are scanned in a top-down order. Every clock cycle, $2S$ pixels are concurrently output to the DWT processor. In order to process an image of size MN, $MN + MF\frac{N-R}{R-F}$ image memory accesses are required. Consequently, for each input frame, $2(MN + MF\frac{N-R}{R-F})$ and $4.5(MN + MF\frac{N-R}{R-F})$ computations have to be performed for 5/3 and 9/7 filters, respectively. For the best existing architecture in [11], $MN + N\frac{M}{2S}$ image memory accesses are required per frame. As shown in the Fig. 3, the number of image memory accesses is significantly reduced compared with the state-of-the-art architecture [11], resulting in higher energy efficiency and lower EAT as demonstrated in Section IV.C. Note that in practical applications, a design with small S can offer enough throughput ($S \le 3$).

C. Image memory and DRAM power schedule

Based on the overlapped block-based scanning method, the proposed parameterized DWT architecture for the 9/7 DWT filter is shown in Fig. 4. The DWT architecture is composed of an image memory and a DWT processor. The image memory is a DRAM bank composed of two DRAM

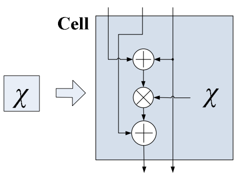

Fig. 5 Detailed structure of Cell

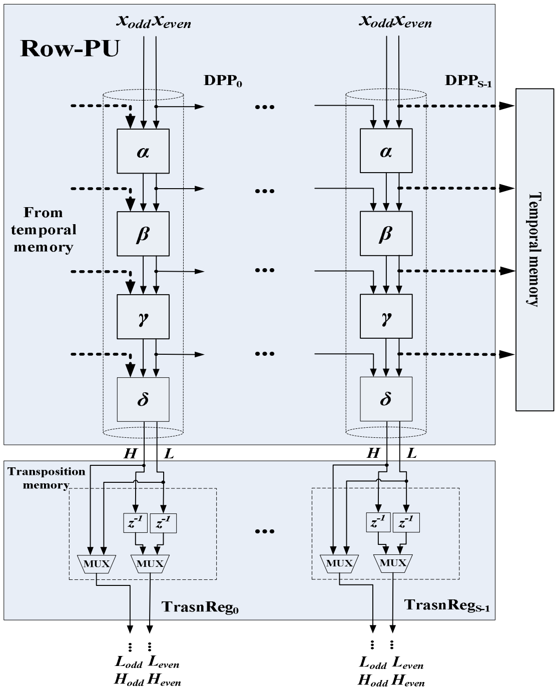

Fig. 6 Row-PU and transposition memory for 9/7 filter

chips to offer enough external bandwidth for high throughput. Each DRAM has 16-bit output.

The DRAM power is composed of active power, read/write term power and background power [19]. While the active and read/write term power is dependent on the number of DRAM reads/writes, the background power is decided by the power mode in which the DRAM operates. DRAM can be in two states, namely pre-charge and active. For each of the state, the DRAM can be operated in either power-down mode or standby mode. In order to perform read/write operations, the DRAM has to be in the active_standby mode, which has the highest power dissipation. In order to reduce the DRAM power dissipation, we schedule the DRAM into pre-charge_power-down mode when the DRAM is not accessed.

D. Parameterized DWT processor

As shown in Fig. 4, the DWT processor is composed of a data buffer, a row processing unit (Row-PU), a column processing unit (Col-PU), a transposition memory, a temporal memory and a scaling unit. The data buffer is used to balance the throughput of the image memory and the DWT processor. The Row-PU and Col-PU are in charge of performing rDWT and cDWT, respectively. The temporal memory stores the partial results generated by Row-PU. The transposition memory transposes the intermediate results H and L generated by the Row-PU and outputs these results to the Col-PU. The H and L are processed by the Col-PU in an interleaved order. The four subbands HH, HL, LH, LL generated by the Col-PU are then scaled by the scaling unit.

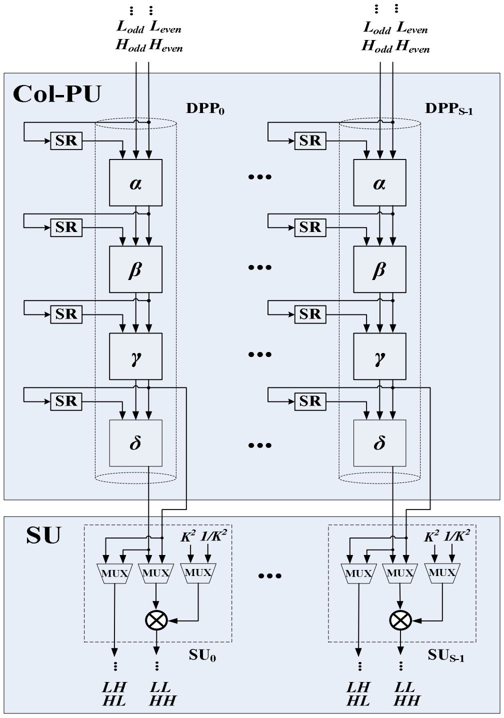

Fig. 7 Col-PU and scaling unit for 9/7 filter

The detailed structure of the proposed design is shown in Fig. 5, 6 and 7. The lifting step represented by equations (1)-(4) is performed by the Cell (Fig. 5), which is composed of one multiplier and two adders. Four Cells with different lifting coefficients are connected to constitute the data processing pipe (DPP) to perform 1-D DWT as shown in Fig. 6 and 7. There are S DPPs in the Row-PU and the Col-PU. The Row-PU consumes 2S pixels from the image memory and 4 partial results from the temporal memory every clock cycle. Meanwhile, 4 partial results are generated by the S-1th DPP of the Row-PU and shifted into the temporal memory. The temporal buffer is composed of four shift registers of length R. While the DPPs in the Row-PU receive/send the partial results from/to their adjacent DPPs, the DPPs in the Col-PU output the four partial results to four shift registers and consume them 2 clock cycles later. The shift registers in the Col-PU have a length of 2. Compared with the existing designs, the proposed architecture achieves a constant on-chip memory size, resulting in small EAT as demonstrated in Section IV.C.

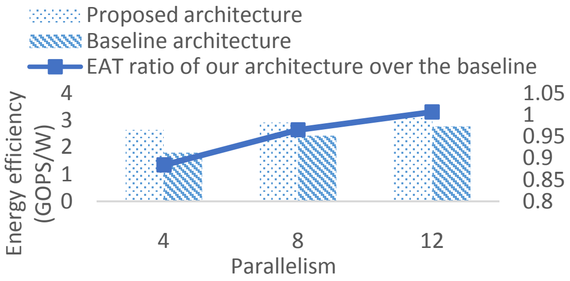

Fig. 8 Energy efficiency and EAT comparison (5/3 filter, small size)

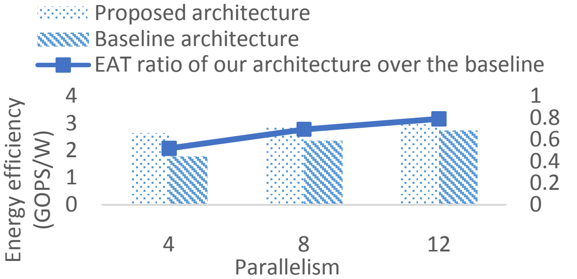

Fig. 9 Energy efficiency and EAT comparison (5/3 filter, medium size)

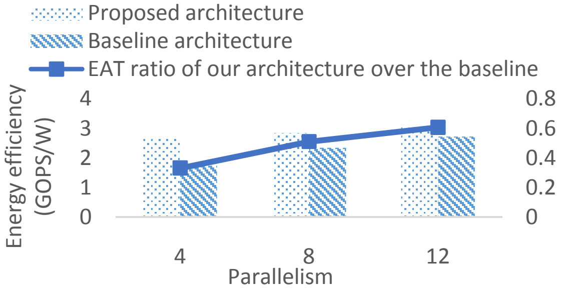

Fig. 10 Energy efficiency and EAT comparison (5/3 filter, large size)

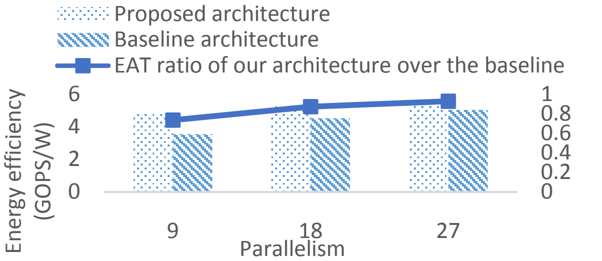

Fig. 11 Energy efficiency and EAT comparison (9/7 filter, small size)

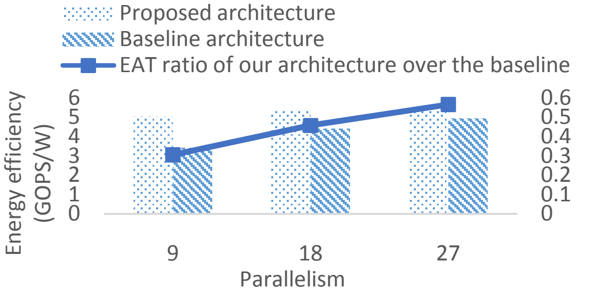

Fig. 12 Energy efficiency and EAT comparison (9/7 filter, medium size)

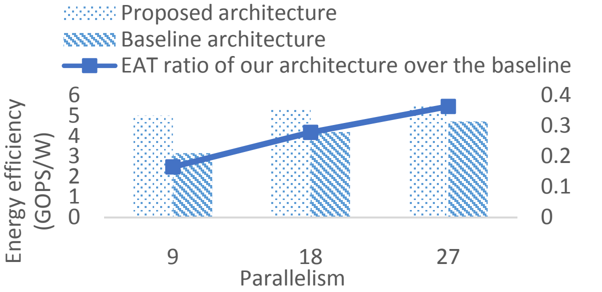

Fig. 13 Energy efficiency and EAT comparison (9/7 filter, large size)

The DWT architecture for 5/3 filter has the similar structure to 9/7 filter, except that there are only two Cells contained in the DPP and there is no scaling unit. Consequently, there are only 2 shift registers of size *R* contained in the temporal memory. For 5/3 and 9/7 DWT filter, the parallelism $L = 4S$ and $9S$, respectively.

IV. Experimental Results and Performance Comparison

A. Experimental setup

In this section, the proposed architecture and baseline architecture (state-of-the-art architecture in [11]) for 5/3, 9/7 DWT filter with various block and input image sizes are implemented. The energy efficiency and EAT are measured and compared. In the following, we choose three image sizes as the specification for the sake of comparison. The small size image (640×480) is used in regular definition video [15] while the medium (1920×1080) and large (3840×2160) size image are employed in recently proposed high efficiency video coding standard (HEVC) [16]. The input image and data path have 32-bit depth.

The designs were implemented in Verilog on Virtex 7 family (XCVX980 with –2L speed grade) using the Xilinx Vivado 2014.1 development tools. We used the Vivado Power Analysis tool to measure the power dissipation based on the post place-and-route simulation. The input matrices were randomly generated with an average toggle rate of 50%. The simulation results were recorded by the switching activity interchange format file (SAIF) and used as input to the Xilinx Power Analysis tool. The Micron DDR3 chip with 16-bit output and 1GB density was used to estimate the power of the image memory. We used the Micron DDR3 power calculator [17] to calculate the power dissipation. The operating frequencies of the image memory and DWT processor were set at 800MHz and 200MHz, respectively.

B. Peak energy efficiency

We compute the peak energy efficiency of the target device (Xilinx Virtex-7 FPGAs). For this, we assume an ideal scenario and consider only the memory access and computation power and ignore all overheads. The power dissipation of the 32-bit multiplier at 200 MHz is 6mW. As the DRAM dissipates a significant amount of power even in power-down mode, we assume the image memory is accessed at its peak bandwidth (1600M 32-bit pixels/sec.) when computing the peak energy efficiency. In order to consume all the input pixels every second, $\frac{1600\times 2}{200}=16$ and

$\frac{1600\times 4.5}{200}=36$ multipliers are required for 5/3 and 9/7 DWT filters, respectively. Consequently, according to equation (6), the peak energy efficiency of 9/7 DWT filter on Virtex-7 FPGAs can be obtained by:

$$\eta_{peak}=\frac{4.5MN}{(36\times 6+2\times 367.9)\frac{4.5MN}{4.5\times 1600}}=7.56 \text{ GOPS/W}$$

With the same method, the peak energy efficiency of 5/3 DWT filter on Virtex-7 FPGAs is obtained as 3.85 GOPS/W.

C. Performance comparison

We compare the energy efficiency and EAT of the proposed architecture with the baseline architecture for the 5/3 and the 9/7 DWT filters. We choose block height *R*=243, 543 and 543 for small, medium and large size input for the 5/3 filter as shown in Fig. 8, 9 and 10. For various parallelism, the proposed architecture for the 5/3 filter achieves 2.63~3.09 GOPS/W and improves the energy efficiency by up to 52.5% compared with the baseline architecture. The energy efficiency for 9/7 filter is shown in Fig. 11, 12 and 13, where the block height R=247, 547 and 547 for small, medium and large input size, respectively. For various parallelism, the proposed architecture for the 9/7 filter achieves 4.88~5.52 GOPS/W energy efficiency and improves the energy efficiency by up to 58.1% compared with the baseline architecture. Compared with the baseline architecture, the proposed architecture has higher energy efficiency due to the lower image memory energy consumption. For both the proposed and the baseline architecture, as parallelism increases, more partial results are used immediately after they are generated, reducing the number of temporal buffer reads/writes and the on-chip memory energy consumption. With increasing parallelism, the DRAM is accessed at higher bandwidth resulting in less energy being consumed in idle state. Therefore, the energy efficiency increases with the increase in parallelism. As the image size increases, the on-chip memory size of the baseline architecture increases with height of the image whereas the on-chip memory size of the proposed design is decided by the height of the block. Consequently, the area of the proposed design is smaller than the baseline architecture and the EAT ratio decreases as the input size increase. Compared with the upper bound on energy efficiency, the proposed architecture achieves up to 80.2% and 72.9% of the peak energy efficiency for 5/3 and 9/7 DWT filters, respectively.

V. Conclusion

In this work, a lifting-based parameterized DWT architecture for the 5/3 and the 9/7 DWT filter is proposed. Compared with prior designs, the energy efficiency is used as main performance metric. The proposed architecture achieves high energy and area efficiency by introducing an overlapped block-based image scanning method which optimizes the number of external memory reads and the on-chip memory size. The experiments show that, by reducing the external memory and on-chip memory energy, our proposed architecture sustains high energy efficiency and low EAT for various input sizes.

VI. References

[1] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 11, pp. 674-693, 1989.

[2] M. Rabbani and R. Joshi, “An overview of the JPEG 2000 still image compression standard,” *Signal Process.: Image Commun.*, vol. 17, no.1, pp. 3–48, Jan. 2002.

[3] “The Green500 List – November 2013,” Available: http://www.green500.org/news/green500-list-november-2013.

[4] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *Journal of Fourier Analysis and Applications,* vol. 4, pp. 247-269, 1998.

[5] B. K. Mohanty and P. K. Meher, "Memory Efficient Modular VLSI Architecture for Highthroughput and Low-Latency Implementation of Multilevel Lifting 2-D DWT," *Signal Processing, IEEE Transactions on,* vol. 59, pp. 2072-2084, 2011.

[6] W. Zhang, Z. Jiang, Z. Gao, and Y. Liu, "An Efficient VLSI Architecture for Lifting-Based Discrete Wavelet Transform," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 59, pp. 158-162, 2012.

[7] Y.-K Lai., L.-F. Lien, and Y.-C. Shih, "A high-performance and memory-efficient VLSI architecture with parallel scanning method for 2-D lifting-based discrete wavelet transform," *Consumer Electronics, IEEE Transactions on,* vol. 55, pp. 400-407, 2009.

[8] B. K. Mohanty, A. Mahajan, and P. K. Meher, "Area- and Power-Efficient Architecture for High-Throughput Implementation of Lifting 2-D DWT," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 59, pp. 434-438, 2012.

[9] C. Cheng and K. K. Parhi, "High-Speed VLSI Implementation of 2-D Discrete Wavelet Transform," *Signal Processing, IEEE Transactions on,* vol. 56, pp. 393-403, 2008.

[10] C.-T. Huang, P.-C. Tseng, and L.-G. Chen, "Analysis and VLSI architecture for 1-D and 2-D discrete wavelet transform," *Signal Processing, IEEE Transactions on,* vol. 53, pp. 1575-1586, 2005.

[11] Y. Hu and C. C. Jong, "A Memory-Efficient Scalable Architecture for Lifting-Based Discrete Wavelet Transform," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol.60, no.8, pp.502-506, Aug. 2013.

[12] B. K. Mohanty, and P. K. Meher, "Memory-Efficient High-Speed Convolution-based Generic Structure for Multilevel 2-D DWT," *Circuits and Systems for Video Technology, IEEE Transactions on,* on-line version, 2012.

[13] Y. Hu and C. C. Jong, "A Memory-Efficient High-Throughput Architecture for Lifting-Based Multi-Level 2-D DWT," *Signal Processing, IEEE Transactions on*, vol.61, no.20, pp.4975,4987, Oct.15, 2013

[14] C. Ren, L. Hoang and V.K Prasanna, "Energy efficient parameterized FFT architecture," *Field Programmable Logic and Applications (FPL), 2013 23rd International Conference on* , vol., no., pp.1,7, 2-4 Sept. 2013

[15] G.A. Davidson, M.A. Isnardi, L.D. Fielder, M.S. Goldman and C.C. Todd, "ATSC Video and Audio Coding," *Proceedings of the IEEE* , vol.94, no.1, pp.60,76, Jan. 2006.

[16] G.J. Sullivan, J. Ohm, W.-J. Han, T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *Circuits and Systems for Video Technology, IEEE Transactions on* , vol.22, no.12, pp.1649,1668, Dec. 2012.

[17] "Micron DDR3 SDRAM System-Power Calculator," Available: http://www.micron.com/products/support/ power-calc.