# Multiple-lifting Scheme: Memory-efficient VLSI Implementation for Line-based 2-D DWT

Chih-Chi Cheng*, Chao-Tsung Huang*, Po-Chih Tseng*, Chia-Ho Pan†, and Liang-Gee Chen*
*DSP/IC Design Lab, Graduate Institute of Electronics Engineering and
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
Email: {ccc,cthuang,pctseng,lgchen}@video.ee.ntu.edu.tw
†Computer&Communications Research Laboratories,
Industrial Technology Research Institute, Hsinchu, Taiwan
Email: chpan@itri.org.tw

***Abstract*— In this paper, a memory-efficient VLSI implementation for line-based 2-D DWT, named multiple-lifting scheme, is proposed. Memory bandwidth and memory size dominate the cost of 2-D DWT and are highly related to the total power and area of 2-D DWT VLSI implementation, respectively. The proposed multiple-lifting scheme can reduce not only the average memory bandwidth but about 50% area of line buffer in 2-D DWT module. The corresponding data scan, M-scan, is proposed to achieve the multiple-lifting scheme and eliminate the data buffer as well.**

## I. Introduction

Due to many good inherent properties, 2-D DWT has been regarded as an efficient tool for image and video processing. From the VLSI implementation perspective, the dominant factor of hardware cost in 1-D DWT module is the number of multipliers and adders. However, memory issues usually dominate the cost of RAM-based 2-D DWT. Memory access is one of the most important sources of total power, and embedded memory also occupies much area in 2-D DWT [1]. Memory-efficient hardware architecture is therefore an important issue in 2-D DWT VLSI design. In recent years, many 2-D DWT architectures for VLSI implementation are proposed [1]–[5]. According to the analysis of [1], external memory access is the most power-consuming component in 2-D DWT. The line-based implementation [6] thus may be preferred if power consumption is a critical issue in 2-D DWT implementation due to its smaller external memory access.

In this paper, the multiple-lifting scheme and the corresponding M-scan are proposed as a memory-efficient VLSI implementation for line-based DWT. In conventional 2-D DWT implementation, there are one read and one write of line buffer every cycle, and the line buffer has to be a two-port RAM. Under the constraint of equal throughput per second, the average memory bandwidth can be reduced to 50% or lower by use of the proposed multiple-lifting scheme. The maximum number of memory access within a clock cycle is also reduced to one access such that the line buffer can be a single-port RAM. The reduction of memory bandwidth in temporal buffer results in the reduction of total power, and the change from two-port RAM to single-port RAM reduces about 50% area of temporal buffer. By use of the proposed M-scan, the data buffer can be eliminated, which avoids the overhead of multiple-lifting scheme.

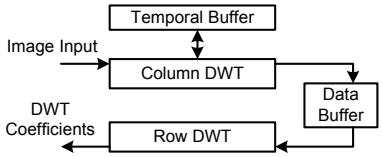


Fig. 1. Generic line-based scheme for column-row 2-D DWT compatible with JPEG2000

This paper is structured as follows. Section II gives an overview of memory structure for line-based DWT. Section III and section IV will present the proposed multiple-lifting scheme and the corresponding M-scan, respectively. Comparisons of implementation results between multiple-lifting scheme and the conventional line-based scheme with eliminated data buffer will be presented in section V. Finally, section VI summarizes this paper.

## II. Overview of line-based DWT

Fig. 1 shows a generic scheme for line-based 2-D DWT, and the 2-D DWT is performed in column-row order to be compatible with JPEG2000 standard. The line buffer needed in line-based DWT can be decomposed into data buffer and temporal buffer as defined in [2]. As shown in Fig. 1, data buffer is used to buffer the intermediate coefficients after column DWT and temporal buffer is used to buffer the register values inside column DWT core. For example, if the lifting-based (9,7) filter is adopted, the number of registers is four [7]. The temporal buffer may have to buffer the register values in 1-D DWT core at every column. The temporal buffer size is thus the product of image width and the number of registers within the 1-D DWT core.

One memory-efficient VLSI implementation is using a proper Z-scan to eliminate the data buffer [8]. However, because the temporal buffer has to represent registers inside 1-D DWT module, its maximum number of memory access within one cycle is one read and one write so that the temporal buffer has to be a two-port RAM, as the lifting-based architecture in column DWT shown in Fig. 2. The computation nodes in Fig. 2 are combinational components such as adders, multipliers and nets. The (9,7) filter is used for illustration throughout this paper.

0-7803-8834-8/05/$20.00 ©2005 IEEE.

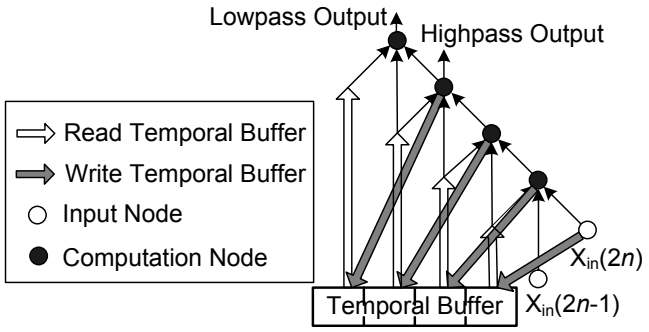


Fig. 2. Lifting-based (9,7) filter for line-based column DWT

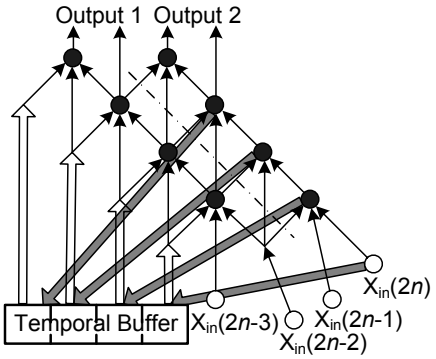


Fig. 3. Two-parallel processing implementation

## III. Proposed Multiple-lifting Scheme

In this section, the proposed multiple-lifting scheme is presented. There are two important goals of multiple-lifting scheme. The first one is to reduce the average memory bandwidth of temporary buffer and thus reduce the power consumption. The other one is to reduce the maximum number of memory access within one cycle in temporal buffer to one read/write such that the temporal buffer can be a single-port RAM and the memory area can thus be reduced.

The discussion throughout this paper will focus on 2-D DWT in column-row order. The comparisons are all under the assumption that the throughput per second is equal in each scheme. All proposed schemes and data scans can be easily modified to be suitable to 2-D DWT in row-column order.

### *A. Reducing average memory bandwidth using parallel processing*

One way to reduce the memory bandwidth is to use parallel processing to raise the throughput per clock cycle such that the total number of memory access for the same number of input pixels will be reduced. Under the constraint of the same throughput per second, the clock rate and memory bandwidth of parallel processing thus can be reduced.

Fig. 3 shows such a two-parallel processing implementation in which two sets of processing element (PE) are adopted. A set of PE is defined as the combinational circuits including adders and multipliers in 1-D DWT of Fig. 2. One action of reading data from temporal buffer can produce two lowpass coefficients and two highpass coefficients. The clock rate and average memory bandwidth of temporal buffer thus can be halved compared with the architecture in Fig. 2. Continuing adding PEs in this way will result in similar multiple-parallel processing implementation, the average memory bandwidth of temporal buffer can be arbitrarily decreased.

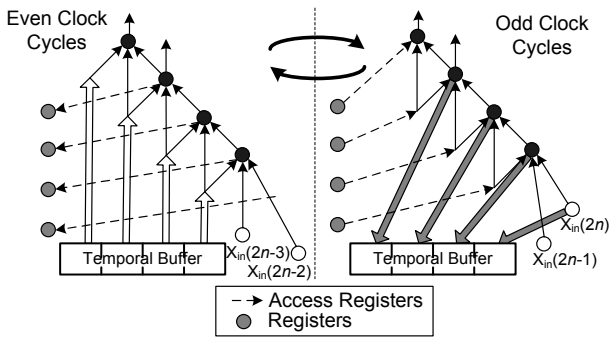


Fig. 4. The proposed two-lifting scheme

The multiple-parallel processing implementation, however, results in increasing PE number that raises the cost of DWT core. Moreover, because one-read and one-write are still required within one cycle in multiple-parallel processing implementation, the temporal buffer still has to be a two-port RAM. The multiple-parallel processing thus trades the increase of area for the decrease of RAM access frequency.

### *B. Proposed two-lifting scheme*

To overcome the disadvantages of parallel processing implementation, the multiple-lifting scheme is proposed. The concept is to maintain the clock rate as conventional line-based implementation discussed in section II and average out the concentrated processing of multiple-parallel processing implementation into every cycle, thus only one set of PE is required and the maximum memory access within one cycle can be reduced.

Take Fig. 3 as an example, if the signals "Output 1" and "Output 2" show up in different cycles, only one set of PE is required. To maintain the low average memory bandwidth, registers are used to buffer the calculated data along the dashed line in Fig. 3. Moreover, the reading and writing of memory can be arranged in different clock cycles such that the maximum memory access within one clock cycle is one read/write, and the temporal buffer can be a single-port RAM.

The proposed two-lifting scheme is shown in Fig. 4. In even cycles, data are read from temporal buffer, and the calculated data are buffered in registers. In odd cycles, data are read from registers, and the calculated data are write into temporal buffer. The average memory bandwidth of temporal buffer is halved, and only one read or one write of temporal buffer is required per clock cycle. The proposed two-lifting scheme thus combines three advantages: halved average memory bandwidth, only one set of PE, and single-port temporal buffer.

### *C. Proposed N-lifting scheme*

The concept of the proposed N-lifting scheme is quite similar to that of the two-lifting scheme. The N sets of PE in N-parallel processing implementation are folded into one PE, and this results in the proposed N-lifting scheme. The scheduling of temporal buffer in four-lifting scheme is shown in Fig. 5 as an example of the proposed multiple-lifting scheme. The average memory bandwidth of temporal buffer

TABLE I

COMPARISONS OF TEMPORAL BUFFER IN DIFFERENT SCHEMES WITH THE SAME OUTPUT THROUGHPUT PER SECOND. B IS THE AVERAGE MEMORY BANDWIDTH (NUMBER OF MEMORY ACCESS PER SECOND) OF TEMPORAL BUFFER IN CONVENTIONAL LINE-BASED IMPLEMENTATION AS FIG. 2

| | Average Memory Bandwidth | Required Number of Processing Element in Column DWT | Maximal Memory Access within One Cycle | Required Type of Temporal Buffer |
|---|---|---|---|---|
| Conventional Line-based | B | 1 | One Read and One Write | Two-port |
| Two-parallel Processing | B/2 | 2 | One Read and One Write | Two-port |
| N-parallel Processing | B/N | N | One Read and One Write | Two-port |
| Proposed Two-lifting Scheme | B/2 | 1 | One Read/Write | Single-port |
| Proposed Four-lifting Scheme | B/4 | 1 | One Read/Write | Single-port |
| Proposed N-lifting Scheme | B/N | 1 | One Read/Write | Single-port |

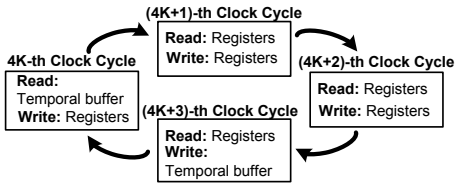


Fig. 5. The scheduling of memory access in proposed four-lifting scheme

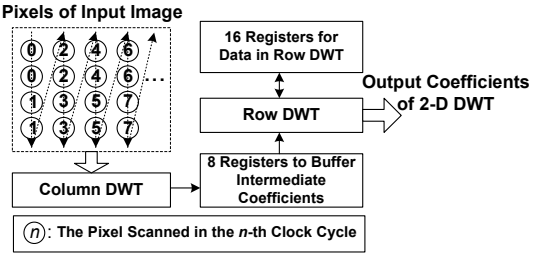


Fig. 6. The proposed M-scan and 2-D implementation for two-lifting scheme with (9,7) filter

in four-lifting scheme thus can be further reduced to half of that in two-lifting scheme while still one full-utilized PE is needed.

In Table I, the proposed multiple-lifting schemes are compared with the conventional line-based scheme in Fig. 2 and multiple-parallel processing implementation in Fig. 3. The proposed multiple-lifting schemes reduce the average memory bandwidth and change the temporal buffer from two-port RAM to single-port RAM while only one set of PE is needed.

## IV. PROPOSED M-SCAN FOR MULTIPLE-LIFTING SCHEME

In section III, the proposed multiple-lifting schemes reduce the power and area of temporal buffer. To eliminate the data buffer, the M-scan suited for multiple-lifting scheme is also proposed in this section.

### A. Proposed M-Scan for two-lifting scheme

Fig. 6 shows the proposed M-scan for two-lifting scheme. The main idea of this M-shape data scan is to use out the intermediate coefficients after column DWT as soon as possible and hence the storage requirement of intermediate coefficients after column DWT can be minimized. Because the scheduling of two-lifting scheme is periodic of two clock cycles as suggested in Fig. 4, the length in column direction of the M-scan has to be four pixels. The row DWT has to be performed right after the intermediate coefficients after column

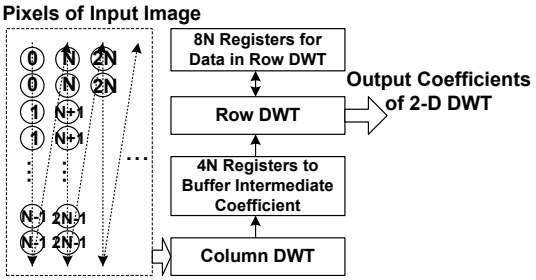


Fig. 7. The proposed M-scan and 2-D implementation for N-lifting scheme with (9,7) filter

DWT are stored in registers for using out coefficients as soon as possible. Because two pixels in row direction have to be fed in row DWT each cycle, eight ($2 \times 4 = 8$) intermediate coefficients after column DWT have to be buffered.

The register values of the row DWT core in each of the four rows in M-scan also have to be buffered. The functionality of this additional buffer in row direction is similar to that of temporal buffer in column direction. The dimension of this additional buffer is four times of the number of registers in row DWT core. If lifting-based (9,7) filter in Fig. 2 is adopted, the size of this buffer is sixteen words ($4 \times 4 = 16$). Registers are sufficient for this small buffer.

### B. Proposed M-Scan for N-lifting Scheme

Fig. 7 shows the proposed M-data scan for N-lifting scheme to eliminate data buffer. As discussed in section III-C, the scheduling of N-lifting scheme is periodic of N clock cycles. The length in column direction of the proposed M-data scan for N-lifting scheme is thus 2N pixels. Because two pixels in row direction have to be fed in row DWT each cycle, 4N ($2 \times 2N = 4N$) intermediate coefficients after column DWT have to be buffered. Register values of row DWT in each of the 2N rows also have to be buffered. The dimension of this additional register array is the product of 2N and the number of registers in 1-D DWT core. If lifting-based (9,7) filter is adopted, the size of this buffer is 8N ($2N \times 4 = 8N$) words as shown in Fig. 7.

## V. EXPERIMENTAL RESULTS

To show the efficiency of the proposed multiple-lifting scheme, three schemes are implemented using (9,7) filter with the same throughput per second. Lifting-based is implemented

TABLE II

COMPARISONS BETWEEN PROPOSED MULTIPLE-LIFTING SCHEME AND CONVENTIONAL LINE-BASED SCHEME WITH ELIMINATED DATA BUFFER UNDER THE SAME THROUGHPUT PER SECOND. ALL THREE SCHEMES HAVE THE SAME CRITICAL PATH AND OPERATE AT CLOCK FREQUENCY = 77MHZ

| Implementation | Total Area ($um^2$) | | Area of RAM ($um^2$) | | Total Power (mW) | | Power of RAM (mW) | | Reduction of RAM - Total Reduction (Overhead) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total Area | Reduction | Area of RAM | Reduction | Total Power | Reduction | Power of RAM | Reduction | Area ($um^2$) | Power (mW) |
| Conventional Line-based | 587737 | | 288786 | | 129 | | 95.95 | | | |
| Proposed Two-lifting Scheme | 422438 | 165298 (28%) | 131435 | 157351 (54%) | 79.44 | 49.56 (38%) | 37.6 | 58.35 (61%) | -7947 (-1%) | 8.79 (6.8%) |
| Proposed Four-lifting Scheme | 452978 | 134758 (23%) | 131435 | 157351 (54%) | 64.52 | 64.48 (50%) | 21.02 | 74.93 (78%) | 21593 (3.7%) | 10.45 (8.1%) |

by flipping structure [9] in all three schemes. The image width is 128 pixels. The first scheme is conventional line-based lifting architecture in Fig. 2 with the non-overlapped stripe-based scan [5] of two pixels stripe width to eliminate the data buffer. The second and third schemes are the proposed two-lifting scheme and four-lifting scheme with the proposed M-scan, respectively. The comparisons of these three schemes are listed in Table II in which Artisan 0.18 $\mu$m cell library and Artisan 0.18 $\mu$m RAM compiler are used. The area information is reported by Synopsys Design Vision and the power information is reported by Synopsys Prime Power. All three schemes are synthesized in the same critical path of 13ns because all three schemes have the same critical path within 1-D DWT core.

Firstly, the total area is reduced by 28% in the proposed two-lifting scheme, which mainly comes from the reduction in the area of RAM. This is due to the required temporal buffer is changed from two-port RAM to single-port RAM. The area of the proposed four-lifting scheme is slightly larger than the proposed two-lifting because there are more registers needed in four-lifting scheme as discussed in section IV-B.

Secondly, the total power is reduced 38% and 50% with proposed two-lifting scheme and four-lifting scheme, respectively. The power of RAM is reduced by 61% in two-lifting scheme because the average memory bandwidth is halved and temporal buffer becomes single-port. The power of RAM in four-lifting is reduced by 78% because the average memory bandwidth is further decreased. The reduced power in RAM is always slightly higher than the reduced total power. Because some registers have to take over the task of buffering when temporal buffer is not accessed, some power is consumed.

Finally, the slight overhead is from the registers buffering the intermediate coefficients and data in row DWT. The overhead of area in two-lifting scheme in Table II is negative because of the variance of circuit synthesis. The power in these registers is reduced greatly by using the clock gating technique. As can be seen in the discussion above, the proposed multiple-lifting scheme can reduce both total power and total area significantly. The temporal buffer size is proportional to image width, but the overhead and cores of 1-D DWT are independent of image width. The reduction ratio will be further increased as longer image width is required.

## VI. CONCLUSION

In this paper, multiple-lifting scheme is proposed to provide a memory-efficient scheme for line-based VLSI implementation. Under the same throughput per second, the proposed two-lifting scheme and four-lifting scheme can halve and quarter the average memory bandwidth of temporal buffer, respectively. The temporal buffer can also be a single-port RAM instead of a two-port RAM. Moreover, the data buffer can still be eliminated with the proposed M-scan. From experiment results, the proposed multiple-lifting can reduce both the total area and total power significantly while maintaining the same throughput per second as anticipated.

## ACKNOWLEDGMENT

This work was supported in part by National Science Council, Republic of China, under the grant number 91-2215-E-002-035 and in part by the MediaTek Fellowship.

## REFERENCES

[1] N.D. Zervas, G.P. Anagnostopoulos, V. Spiliotopoulos, Y. Andreopoulos, and C.E. Goutis, "Evaluation of design alternatives for the 2-D discrete wavelet transform," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, no. 12, pp. 1246 – 1262, Dec. 2001.

[2] P.-C. Tseng, C.-T. Huang, and L.-G. Chen, "Generic RAM-based architecture for two-dimensional discrete wavelet transform with line-based method," in *Asia-Pacific Conference on Circuits and Systems*, 2002, pp. 363–366.

[3] C. Chakrabarti, M. Vishwanath, and R. M. Owens, "Architectures for wavelet transforms: A survey," *Journal of VLSI Signal Processing*, vol. 14, pp. 171–192, 1996.

[4] K. K. Parhi and T. Nishitani, "VLSI architectures for discrete wavelet transforms," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 1, no. 2, pp. 191–202, June 1993.

[5] Chao-Tsung Huang, Po-Chih Tseng, and Liang-Gee Chen, "Memory analysis and architecture for two-dimentional discrete wavelet transform," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 5, pp. 13–16.

[6] C. Chrysafis and A. Ortega, "Line-based, reduced memory, wavelet image compression," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 378–389, Mar. 2000.

[7] W. Sweldens, "The lifting scheme: a custom-design construction of biorthogonal wavelets," *Applied and Computaional Harmonic Analysis*, vol. 3, no. 15, pp. 186–200, 1996.

[8] Mu-Yu Chiu, Kun-Bin Lee, and Chein-Wei Jen, "Optimal data transfer and buffering schemes for JPEG2000 encoder," in *IEEE Workshop on Signal Processing Systems*, 2003, pp. 177–182.

[9] C.-T. Huang, P.-C. Tseng, and L.-G. Chen, "Flipping structure: An efficient VLSI architecture for lifting-based discrete wavelet transform," *IEEE Transactions on Signal Processing*, vol. 52, no. 4, pp. 1080–1089, Apr. 2004.