HIGH SPEED VLSI ARCHITECTURE FOR NON SEPARABLE BLOCK BASED LIFTING WAVELET TRANSFORM

USHA BHANU.1, Dr.A.CHILAMBUCHELVAN2
1Research Scholar, Department of ECE, Saveetha School of Engineering, Chennai, India.
2Professor & Head, Department of EIE, R.M.D Engineering College, Chennai, India.
E-mail: 1bhanu.usha@rediffmail.com, 2chill97@gmail.com

ABSTRACT

The inherent advantage of the in-place computation of the lifting based DWT over the conventional convolution method makes it suitable for efficient hardware implementation with lower computational complexity. The proposed paper uses Non Separable block based method to derive the VLSI architecture for the two dimensional DWT. In separable lifting DWT, the desired filter coefficients at each level are dependent on previous output level and this introduces delay or latency in DWT decomposition. The proposed Non–separable architecture contains floating point multipliers and adders based on IEEE 754 single precision format. It requires six lifting steps for the implementation of 9/7 lossy lifting DWT and the latency is reduced when compared to separable DWT. It is suitable for handling a parallel block of 4×4 input image coefficients at each level, for achieving higher speed and dynamic accuracy. The architecture is coded in Hardware Description Language and implemented in Altera Cyclone II FPGA. The critical path delay for the proposed Non Separable architecture is 3.189 ns, with a speed of 313.58 MHz. The total thermal power dissipation of the architecture is 128.59 mW, suitable for low power embedded multimedia applications.

Keywords: Block Based Scanning, Floating Point Arithmetic, FPGA Implementation, Lifting DWT, Non Separable And Separable DWT.

1. INTRODUCTION


Various efficient VLSI architectures were proposed to meet the real time implementation of the DWT. Wu et al [12] proposed a poly-phase decomposition technique and the coefficient folding technique of decimation filters for the implementation of the convolutional 2D DWT. Gregory Dillen et al [4] proposed a combined Line-Based architecture for the 5/3 and 9/7 Wavelet transform of JPEG2000, using lifting schemes by incorporating pipelining for optimization. In line based architectures the requirement of temporal memory adds constraint in implementing two dimensional DWT efficiently in hardware. Shi et al [14] presented an Efficient Folded Architecture (EFA) with lower computational complexity. The critical path delay affects the performance of this architecture. Huang et al [5] proposed a flipping structure for lifting DWT using five stage pipelining, where the critical path is reduced to one multiplier delay, but it requires a large temporal buffer. Chao Cheng et al [2] proposed a systematic high-speed VLSI implementation of the DWT, based on hardware-efficient parallel FIR filter structures. Xin Tian, et al [11] presented an efficient multi-input/multi-output VLSI architecture (MIMOA) with computing time as low as \( \frac{N^2}{M} \) for an \( N \times N \) image, with controlled increase of hardware cost, where ‘M’ is the throughput rate. Pramod Kumar et al [8] proposed hardware-efficient systolic-like modular design using a new data-access scheme and a novel folding technique for the implementation of the DWT. Mohanty et al [13] proposed block processing and systolic array for lifting based 2D DWT. A high performance,
memory efficient VLSI architecture for Lifting DWT based on parallel scanning method for reducing temporal memory requirement was proposed by Yeong et al [15].

All the above architectures are based on separable 2-D DWT implementation. The 1D DWT is obtained by first applying the lifting steps for the given image coefficients in row-wise to produce L and H sub bands and then column-wise to give four bands LL, LH, HL, HH. It requires 8 lifting steps for 9/7 lossy compression standard of JPEG 2000. In this paper a block based [9] approach for the given floating point image coefficients is done and the order of scanning of the image coefficients is done first column wise and then row-wise to reduce memory requirements. A non-separable VLSI architecture by combining row and column processor as proposed by Mashiro et al [7] can be used for high speed VLSI implementation of lifting 2-D DWT with reduced number of lifting steps for low power embedded multimedia applications. Zhang et al [17] proposed pipelined architecture for multilevel 2-D DWT using Non separable approach by distributing the task of the computations of multiple decomposition levels among the stages of pipeline. To represent the given real image coefficients in binary format for multipliers and adders in the discussed architecture IEEE 754 [16, 18, 19, 20] single precision floating point format is used to improve the dynamic accuracy of the input image signals.

The motivation of this work is to propose a high speed Non-separable VLSI architecture for the given image coefficients for the efficient implementation of block based lifting DWT with reduced number of lifting steps when compared to the existing separable implementation. The lifting algorithm for two dimensional non separable lifting DWT is coded in HDL, and the implementation results in Altera cyclone FPGA show that the hardware implementation of non-separable lifting DWT is suitable for resource constrained high speed low power embedded multimedia applications.

The rest of the paper is structured as follows: Section 2 reviews the basics and mathematics of the lifting algorithm involved in deriving the non-separable lifting DWT for the implementation of the 9/7 lossy wavelet filters and comparison of latency in the existing separable DWT is presented. In section 3 the proposed two dimensional Non separable architecture for 9/7 lossy standard of JPEG 2000 with floating point implementation of multiplier and adders and the corresponding implementation steps for effective hardware utilization is given. Section 4 provides the implementation results of the proposed non-separable architecture and its comparison with the previous existing separable architectures is given. Finally in section 5 the concluding remarks and the possibility of extending this work in future are made.

2. MATHEMATICS INVOLVED IN THE DERIVATION OF NON SEPARABLE VLSI ARCHITECTURE FOR 9/7 LOSSY FILTERS

The conventional in place computation of Lifting DWT depends on the results of previous output at each level. This introduces latency of eight in separable 2-D DWT. This section introduces mathematical background involved in the derivation of separable and Non separable Lifting steps for efficient hardware implementation. The Non separable architecture is obtained by combination of row/column filters. The lossy standard 9/7 filters require six lifting steps and latency is reduced.

2.1 Separable Two Dimensional DWT And Its Latency In Lifting Steps For 9/7 Lossy Standard

The lifting scheme has been developed as a flexible tool suitable for constructing the second generation wavelets [3]. It is composed of three basic operation stages: split, predict and update. The three basic steps in Lifting based DWT are:

- **Split step**: where the signal is split into even and odd points, because the maximum correlation between adjacent pixels can be utilized for the next predict step. For each pair of given input samples x(n) split into even x(2n) and odd coefficients x(2n+1).

- **Predict step**: The even samples are multiplied by the predict factor and then the results are added to the odd samples to generate the detailed coefficients (dj). Detailed coefficients results in high pass filtering.

- **Update step**: The detailed coefficients computed by the predict step are multiplied by the update factors and then the results are added to the even samples to get the coarse coefficients (sj). The coarser coefficients gives low pass filtered output. The JPEG2000 defines 5/3 lossless standard with two lifting steps and 9/7 lossy standard with four
lifting steps for the implementation of 1-D DWT. The 2-D separable DWT is obtained by applying the 1-D image coefficients again to a set of horizontal (row) and vertical (column) filters as shown in Figure 1.

![Figure 1: Representation of Separable 2-D DWT](image)

The aim of this work is to compare the latency involved in the execution of separable and Non-separable lossy 9/7 standard for high speed applications of lifting DWT. The separable method of implementation of 2-D DWT requires temporal memory for storing the intermediate levels of coefficients.

i. 9/7 lossy standard

The lossy 9/7 standard requires four lifting steps in row processing and four lifting steps in column processing for the implementation of separable 2-D DWT. The low pass and high pass filter coefficients for 2D 9/7 row processing is given by,

\[
\begin{bmatrix}
H_1 & H_3 \\
H_2 & H_4
\end{bmatrix}
\begin{bmatrix}
H_1(Z_1) & H_3(Z_1) \\
H_2(Z_1) & H_4(Z_1)
\end{bmatrix}
\]

The column filters are given by,

\[
\begin{bmatrix}
H_3 & H_1 \\
H_4 & H_2
\end{bmatrix}
\begin{bmatrix}
H_3(Z_2) & H_1(Z_2) \\
H_4(Z_2) & H_2(Z_2)
\end{bmatrix}
\]

Where

\[
\begin{bmatrix}
H_1(Z) & 0 \\
0 & H_2(Z)
\end{bmatrix}
\begin{bmatrix}
\alpha & \beta \\
0 & \beta
\end{bmatrix}
\begin{bmatrix}
(1+Z) \\
(1+Z^{-1})
\end{bmatrix}
\]

The computations of filter coefficients in each level are dependent on the results of the previous stage. The latency exists in separable decomposition and is equal to eight for row-column decomposition of separable lifting DWT. Each of the filters \(H_1(Z), H_2(Z), H_3(Z), H_4(Z)\) contains two adders, one multiplier and one delay register with a critical path delay of \(T_m + 2T_a\), where \(T_m\) and \(T_a\) are the delay in the execution time of multipliers and adders respectively.

The filter \(H_1(Z)\) computes predict stage \(P_1\) given by,

\[
D(n) = X_k(n) + \alpha [X_k(n) + X_k(n + 1)]
\]

(5)

\(H_2(Z)\) computes update stage \(U_1\) given by,

\[
S(n) = X_k(n) + \beta [D(n-1) + D(n)]
\]

(6)

\(H_3(Z)\) computes predict stage \(P_2\) resulting in high pass filtering and is given by,

\[
HP(n) = D(n) + \gamma [S(n) + S(n + 1)]
\]

(7)

\(H_4(Z)\) computes update stage \(U_2\) resulting in desired low pass filtering,

\[
LP(n) = S(n) + \delta [HP(n-1) + HP(n)]
\]

(8)

After \(N\) lifting steps, scaling coefficients of \(K\) and \(1/K\) are applied to horizontal and vertical coefficients to obtain the low pass and high pass coefficients. This scaling is done for normalizing the magnitude. The values of constants are \(\alpha = 1.586134, \beta = 0.0529801185, \gamma = 0.882911076, \delta = -0.443506852\) and \(K = 1.149604398\).

2.2 Mathematics in the proposed Non Separable 2-D Lifting DWT and its latency

The Non-separable 2-D DWT can be obtained, by combining the row/column decomposition filters, as discussed by Masahiro [7], so that the latency of
eight lifting steps is reduced to six in the proposed architecture. The low pass and high pass filter coefficients of 2-D 9/7 DWT for row/column processing are given by,

\[
\begin{bmatrix}
H_1H_3 & H_3H_3^* \\
H_2H_2^* & H_4H_4^*
\end{bmatrix} =
\begin{bmatrix}
H_1(1)H_1(1) & H_3(1)^2H_3(1)^2 \\
H_2(1)^2H_2(1) & H_4(1)^2H_4(1)^2
\end{bmatrix}
\]  

(9)

The filter coefficients \(H_1, H_2, H_3, H_4\) are the row processing filters and \(H_{1,2,3,4}\) are the column filters. The in-place matrix representation of the Non separable Lifting is obtained, by multiplying the down sampled version of the input image coefficients with the corresponding filter coefficients, to obtain the sub band decomposed DWT values. The output decomposed values LL, LH, HL and HH are related to the down sampled odd and even coefficients values \(X_{11}, X_{12}, X_{21}, X_{22}\) by:

\[
\begin{bmatrix}
LL \\
LH \\
HL \\
HH
\end{bmatrix} = (K)(N_{H_1,H_2,H_3,H_4})(N_{H_{1,2,3,4}})(X_{11}, X_{12}, X_{21}, X_{22})
\]

(10)

where \(\alpha, \beta, \gamma\) and \(\delta\) are the multiplier coefficients, and \(K\) is the constant for normalizing the magnitude, for floating point representation. The inputs \(X_{odd}\) and \(X_{even}\) are the odd and even pixels of the down sampled input image coefficients. The architecture uses four blocks of a 4×4 input image in parallel, at the given clock period. The output sub bands LL, LH, HL, HH are obtained, based on the input-output relationship of Equation [10], and the general representation of the Non Separable method is shown in Figure 2.

\[
N = \begin{bmatrix}
1 & \beta & \beta \\
\gamma & 1 & 0 \\
\alpha & 0 & 1 \\
\alpha \gamma & \alpha & \gamma & 1
\end{bmatrix}
\]

(11)

\[
K = \text{diag}[K^2, 1, 1, K^2]
\]

(12)

Figure 2: Representation of the Non Separable Lifting DWT
The total number of lifting steps is reduced from eight (separable) to six in the Non Separable implementation of the 2-D lifting DWT. The latency for the execution steps is reduced, and their percentage of reduction when compared to that of the separable method is shown in Table: 1.

<table>
<thead>
<tr>
<th>DWT METHOD</th>
<th>LIFTING STEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5/3 Lossless</td>
</tr>
<tr>
<td></td>
<td>9/7 Lossy</td>
</tr>
<tr>
<td>Separable 2-D</td>
<td>4(100%)</td>
</tr>
<tr>
<td>DWT</td>
<td>8(100%)</td>
</tr>
<tr>
<td>Non Separable 2-</td>
<td></td>
</tr>
<tr>
<td>D DWT</td>
<td>3(75%)</td>
</tr>
<tr>
<td></td>
<td>6(75%)</td>
</tr>
</tbody>
</table>

3 HARDWARE IMPLEMENTATION OF THE NON SEPARABLE LIFTING 2-D DWT

The Non separable Lifting DWT architecture proposed uses the combination of row/column filters. The floating point arithmetic implementation of the functional units in the architecture is done, for handling the real time implementation of image/video signals with improved accuracy. The latency is reduced to six for 2-D lifting DWT.

3.1 Proposed Architecture for the Non Separable Lifting DWT

The row/column filter in the proposed architecture in Figure 3, is implemented, by using 32 bit floating point multipliers, adders and registers. The down sampler divides the given array of input image coefficients in to odd and even coefficients. The inputs $X_{11}$, $X_{12}$, $X_{21}$ and $X_{22}$ of the proposed architecture are the corresponding outputs of the down sampler. The registers used in the architecture are 32 bit registers to hold the result of the execution at each level. These registers are pipelined to achieve parallel processing of the outputs at each level.

Figure 3: Proposed Non Separable 2-D Lifting DWT Architecture
3.2 Method of Scanning of Image Coefficients

The input image coefficients are scanned along the first column of the block of coefficients, and the filtering is done by combining the row–column filters to generate a detailed HH sub band output. The image coefficients are scanned along the second column, to generate the LH and HL sub band coefficient output. The image coefficients are scanned along the third column, to generate the required approximated low pass (LL) output. The same processing steps are repeated along the 4th, 5th and 6th column and multiplied with the scaling factors to generate the final sub band coefficient output. The general order parallel scanning of the image coefficient is shown in Figure. 4 below. In this each circle denotes a parallel block of 4×4 image coefficients at each level.

![Scanning Order Of The Input Image Coefficients In The Proposed Method](image)

The floating point multiplier and adder in the architectural design, as shown in Figure.3 are being efficiently implemented in the hardware, to make Non separable architecture suitable for high speed and low power image processing applications, with improved dynamic accuracy. In the multimedia System on Chips, the design of specialized floating point arithmetic units (FPUs) is critical, with respect to the operating speed and area. An efficient implementation of the IEEE 754 single precision floating point multipliers for FPGA devices is discussed by Ashrafy et al [19]. The pipelining of multipliers for a high speed MAC unit is also discussed. The Multiplication of the floating point numbers $X_1$ with sign $S_1$, exponent $E_1$ and significand $P_1$ and the second number $X_2$ with sign $S_2$, exponent $E_2$ and significand $P_2$ consists of preprocessing of the given floating point number is used to detect whether the given number is not a number (NaN), Infinity (INF), and zero (IsZ) to accommodate the given image floating point value in the given word length, as discussed by Kahan [18]. Leading one anticipation algorithm is a well-known method, for implementing high speed floating point adders, and the architecture for the same was given by Ashwini [20].

4 IMPLEMENTATION RESULTS OF THE PROPOSED NON SEPARABLE ARCHITECTURE

The developed 2-D DWT non separable lifting architecture comprising of the down sampler, 32 bit floating point registers, adders and multipliers, is coded in Verilog HDL and is implemented in EP2C35F672C6 cyclone II Altera FPGA. The computational complexity of the architecture is increased due to the floating point arithmetic, but this disadvantage can be overcome by achieving
reduced critical path latency and low power, suitable for multimedia applications.

4.1 Simulation Results of the Non Separable Architecture

The simulation of the proposed architecture is done in Modelsim Altera. The Verilog code for the architecture is coded for the functional circuits, and test bench is created for the given array of image coefficients. The test bench is created for 256×256 image coefficients, which are loaded into the architecture; where at each level of the inputs a block of 4×4 image coefficients is given. The image is read from the file in the read mode, and the output coefficients are initialized in the write mode in synchronization with the clock period. From the simulation analysis it is clear, that as soon as the image coefficient file is loaded into the architecture the sub band output, a low pass coefficient is obtained at a latency of six clock cycles.

4.2 Implementation Results of the Non Separable architecture

The performance analysis of the Non separable Lifting 2-D DWT architecture is coded in HDL, and is generic for any square input block of image coefficients. The architecture is implemented using EP2C35F672C6 cyclone II Altera FPGA. The performance of the proposed architecture is compared, in terms of the number of multipliers, number of adders, storage size, computing time, control complexity and hardware utilization. The computing time has been normalized to the internal clocking rate. Every 4×4 block of coefficients at each level generates the sub band output blocks of LL, LH, HL and HH coefficients, with a latency of six clock cycles. The RTL schematic of the Non separable DWT architecture is shown in Figure. 5; it consists of an input image coefficient X of 32 bit floating point data, along with the down sampler, registers and the required sub band outputs LL, LH, HL and HH coefficients.

The sample of the floating point adders, and multipliers, along with the required down sampled input image coefficients and registers are shown in Figure. 6 below. . The implementation results given in Tables: 2 & 3 show that the hardware implementation of the lifting schemes based on the Non Separable DWT is suitable for real time high speed low power multimedia applications.
From the implementation results the critical path latency of 3.189 ns is achieved with a 313.58 MHz of frequency of operation. The total power dissipation of the proposed architectures is 128.59 mW, suitable for low power applications.

### 4.3 Comparison of architectures

The comparison of the 2-D Non separable DWT architecture is compared with other existing separable architectures. The parameters taken for comparison are hardware complexity (measured by number of multipliers, adders and registers); the DWT scheme used, and operating frequency of the architectures. From the graph shown in Figure.7 it is clear that the proposed design method provides high speed when compared to the other existing architectures.
The block based lifting DWT architecture [9] based on overlapped scanning for reducing the memory requirements for the implementation of lossless (5, 3) DWT achieves a maximum operating frequency of 97MHz. The Multi- input/ Multi-output [11] 2D VLSI architecture for 9/7 lifting DWT achieves an operating frequency of 64.25 MHz for varying block sizes of 2, 4, and 8 respectively. The high speed 2-D convolution based parallel FIR structures [2] for hardware efficient DWT achieves a maximum operating frequency of 58.73 MHz with a computation time of $\frac{N^2}{12}$ where $N$ is the dimension of the given image. The hardware efficient systolic like modular [8] design for the implementation of 2-D row/column decomposition of separable convolution DWT achieves an operating frequency of 230.3MHz. The architecture uses folding technique and new data access scheme for the implementation. From the implementation results shown it is clear that the proposed Non- Separable VLSI architecture based on combined row/column filtering and with reduced lifting steps of six when compared to other existing separable lifting DWT achieves a high speed of 313.58 MHz suited for real time image/video applications.

5. CONCLUSION

The high speed Non Separable VLSI architecture for efficient implementation of lifting based DWT with reduced number of lifting steps when compared to the existing separable architectures is proposed in this paper. The architecture is capable of handling a parallel blocks of input image coefficients at each level to generate the sub band filtered outputs. The floating point implementation of multipliers and adders are used in the architecture for handling real time image signals with the controlled increase of hardware complexity. The parallel processing of input image coefficients is done to achieve a high speed of 313.58 MHz with a reduced critical path latency of 3.189 ns from input to output suitable for handling high speed, low power multimedia applications. The advantage of the discussed architecture is that it does not require additional memory for storing the intermediate coefficients. The Non separable architecture can be extended for varying block sizes with overlapped scanning of image coefficients thus reducing accessing time and memory requirements for storing the intermediate coefficients.

REFERENCES


