CHAPTER 2

LITERATURE REVIEW

The development of new technologies and communication networks stimulates the need for introducing new functionalities in multimedia applications. The current standards in the fields of still image coding are inadequate for producing the best quality of performance in real time image/video applications. To address this issue the Joint Photographic Experts Group came up with a new coding system JPEG 2000, based on the multi-resolution analysis of the DWT as proposed by Mallat (1989). The JPEG 2000 provides superior image quality performance compared to other existing image file formats, viz., JPEG, BMP, GIF etc, as discussed in Fred Harshall (2001).

2.1 OVERVIEW OF THE EXISTING ARCHITECTURES

Various 1-D and 2-D DWT VLSI architectures have been proposed to meet the requirements of real-time processing. The real time implementation of DWT has to process massive amounts of data at high speeds. The Convolutional DWT has traditionally been implemented by convolution or the Finite Impulse Response (FIR) filter bank structures, as discussed in Parhi & Nishitani (1993); Wu&Chen (2001). Filter bank implementations of the DWT require both large number of arithmetic computations and storage, which are not desirable for either high speed or low power image/video processing applications.
The lifting scheme is a new method to construct a wavelet basis, which was first introduced by Daubechies & Swelden (1998). The lifting scheme relies entirely on the spatial domain, and has many advantages compared to the convolutional method, such as lower area, greater power consumption and computational complexity. The objective of this research work is to propose VLSI architecture for high speed and low power implementation of lifting based DWT, with minimum hardware cost and computational complexity.

2.1.1 Existing Architectures for Convolutional & Lifting based DWT

Several VLSI architectures have been proposed for lifting based DWT implementation in FPGA. Design acceleration has been achieved through parallel processing of independent sub-modules, and reusability of image pixel data. Registers were used for temporary storage, and reusability of temporary data. The various architectures differ in terms of the required numbers of multipliers, adders and registers, as well as the amount of accessing external memory, and lead to decrease efficiently the hardware cost and power consumption of design. In spite of improving the efficiency of the existing architectures as discussed in Acharya & Chakrabarti (2006) and Salehi & Sadri (2009), the present requirement is to improve the hardware utilization, and capacity of handling multiple data streams for the calculation of 2-D DWT, suited for high speed real time multimedia applications.

Parhi & Nishitani (1993) presented two classes of novel VLSI architectures, referred to as the folded architecture and the digit-serial architecture for the implementation of 1-D and 2-DConvolutional DWT. In the 1-D folded architecture, the computations of all wavelet levels are folded to the same low-pass and high-pass filters. The number of registers in the folded architecture is minimized by using the generalized life time analysis. The folded architecture is presented in the context of the word- parallel
implementation method. The folded architectures are used in latency critical applications, and digit-serial architectures are used for low complexity and power applications. The combination of the folded and digit-serial architectures is used for the implementation of the 2-D DWT.

Chrysafis & Ortega (2000) presented a line-based, Reduced Memory, and Wavelet Image Compression suited for low memory systems. The context based encoder and decoder utilizes line based scanning of images. The images are read line by line, and only the required minimum numbers of lines are kept in memory. The low memory coder achieves maximum performance with minimum memory requirements. A novel adaptive context based encoder is proposed which requires no global information and stores only a local set of wavelet coefficients. The advantage of this architecture is its low memory requirement with excellent compression performance for the given image coefficients.

Benini et al (2000) proposed dynamic power management methodology for dynamically reconfiguring systems with minimum number of active components. Dynamic Power Management (DPM) includes selectively turning off the system components when they are idle. A survey of several approaches to system level power management is presented for standardizing the hardware/software interface to enable software controlled power management of hardware components.

Wu & Chen (2001) proposed an efficient architecture for the 2-D DWT. It includes a transform module, a RAM module and a multiplexer. In the transform module, the polyphase decomposition technique and the coefficient folding technique to the decimation filter of stages 1 and 2 respectively. The level of decomposition of the wavelet coefficients is level by level approach. The transform module includes two stages: stage 1 performs horizontal filtering and stage 2 performs vertical filtering. The
Coefficient folding technique is used in vertical filtering for reducing the hardware requirements. In comparison with other 2-D DWT architectures, the advantages of the proposed architecture are 100% hardware utilization, fast computing time, regular data flow and low control complexity, making this architecture suitable for next generation JPEG 2000 image compression systems.

Andra et al (2002) proposed generalized VLSI architecture for Lifting-Based Forward and Inverse Wavelet Transform. The highly programmable architecture that supports large set filters proposed in JPEG2000. These filters include (5,3), (9,7), C(13,7), S(13,7), (2,6), (2,10) and (6,10). The data dependencies in the filter computations are represented at most by four stages the architecture consists of two row processors, two column processors and two memory modules. Each processor contains two adders, one multiplier, and one shifter. The multipliers and shifters are interconnected to support the computational structure of specific filter. Each memory module consists of four banks in order to support the high computational bandwidth. The proposed architecture calculates the forward and inverse DWT in a row-column fashion on a block of data of size N×N. At each level of decomposition the outputs are LL, LH, HL and HH coefficients. The LL data is used for the next level of decomposition.

Grangetto et al (2002); Martina et al (2003) proposed optimization and implementation of an image transform coding algorithm based on the integer wavelet transform (IWT). It proposed for the selection of optimal factorizations of the wavelet filter polyphase matrix to be employed within the lifting scheme. The effect of finite word length effect on the wavelet coefficients are discussed so that a small number of bits are kept for mantissa part of the given image coefficients to reduce performance degradation. The
VLSI architecture based on IWT is capable of achieving higher frame rates with moderate gate complexity.

Dillen et al (2003) proposed combined Line-Based Architecture for the 5/3 and 9/7 Wavelet Transform of JPEG2000. The lifting scheme is used to realize a fast wavelet transform. Two lines are processed at a time, and allow minimum memory requirement and fast calculation. The proposed architecture allows continuous processing of images the computation of several levels of decomposition through cascading in time/ space can be done. The pipeline and the optimization of the operations provide speed, while the combination of the two transforms in one structure contributes to saving the area. Boundary treatment of horizontal and vertical filtering is discussed for the signals.

Liao et al (2002); Liao et al (2004) proposed an efficient recursive architectures for 1-D and 2-D Lifting-Based Wavelet Transforms. The traditional DWT architectures compute the current level, j\textsuperscript{th} of decomposition upon completion of previous, (j-1)\textsuperscript{th} level. However in multi-resolution DWT, the number of samples to be processed in each level is always half of the size in previous level. Thus it is possible to process multiple levels of decomposition simultaneously. This is the basic principle of recursive structures proposed by Vishwanath (1994) for convolutional DWT and is applied for lifting based DWT by Liao et al (2002); Liao et al (2004). The recursive and dual scan architectures for 1-D and 2-D DWT based on lifting scheme is proposed. The 1-D recursive architectures exploit inter dependencies among the wavelet coefficients by interleaving on alternate clock cycles using the same data path. It processes two independent data streams using shared functional blocks. The 2-D dual scan architectures process the column and row transform simultaneously and the required memory buffer size is reduced. The computation in higher levels of
decomposition is initiated as soon as the enough intermediate data in low frequency sub band is available for computation.

Huang et al (2004) proposed flipping structure to reduce the critical path and to minimize the memory requirements for the implementation of the lifting based DWT. Latency in lifting based architectures are due to the timing delays from the input node to the computation node in each computing unit, this can be done by eliminating the multipliers on the path from the input node to the computation node. This can be achieved by flipping each computing node with the inverse of the multiplier coefficients. The critical path of the lifting based DWT is reduced and the reduction rate will increase as the number of serially connected computing units becomes larger.

Chen (2004) proposed VLSI Implementation for 1-Dimensional Multilevel Lifting-Based Wavelet Transform. The lifting scheme has been developed as a flexible tool suitable for constructing bi-orthogonal wavelets. The architecture folds the computations of all resolution levels into the same low-pass and high-pass units to achieve higher hardware utilization, because of its modular, regular and flexible structure, the design is scalable for different resolution levels. The architecture has a similar topology to a scan chain; it can be modified as testable scan based chain by adding additional hardware resources.

Jung et al (2004) proposed an efficient line based VLSI Architecture for 2-D Lifting DWT. The architecture computes lifting operation based on state space representation and uses RPA (Recursive Pyramid Algorithm) scheme. To improve hardware utilization, the filter that is responsible for column operations of the first level performs both the row and column operations of the second and following levels. This improves the
hardware utilization of the architecture. For boundary processing of image coefficients periodic symmetric extension is used.

Barua et al (2004) proposed an architecture that is easily scaled to accommodate different numbers of lifting steps. The architecture has regular data flow and low control complexity, and achieves 100% hardware utilization. Symmetric extension of the image to be transformed is with minimum number of clock cycles. The architecture achieves higher throughput and uses less embedded memory than convolutional filter banks.

Xiong et al (2007) uses embedded decimation technique is exploited to optimize the architecture for 1-D Lifting DWT, which is designed to receive an input and generate an output with the low- and high-frequency components of original data being available alternately. Based on 1-D DWT architecture, an efficient line-based architecture for 2-D DWT for high speed is proposed by employing parallel and pipeline techniques, which is mainly composed of two horizontal filter modules and one vertical filter module, working in parallel and pipeline fashion with 100% hardware utilization. This is achieved by exploiting the parallelism among the four sub band coefficients in lifting based 2-D DWT. The multipliers are replaced by shift-add operations for 5/3 lossless transform because the lifting coefficients are integer powers of two.

Angelopoulou et al (2007) proposed implementation and Comparison of the 5/3 lifting 2-D Discrete Wavelet Transform Computation Schedules on FPGAs. For the execution of the multilevel 2-D DWT, several computation schedules based on different input traversal patterns have been proposed. The row–column, the line-based and the block-based architectures can be implemented for real time applications.
Chang et al (2007) proposed adaptive clock gating technique (ACG) which can be used as low power IP core design suitable for System on Chip (SoC) design. Adaptive clock gating technique can automatically enable or disable the Intellectual Property core clock to reduce dynamic and leakage power with power gating technique. Programmable clock gating and adaptive clock gating technique is used to reduce power consumption in IP cores in SoC design.

Cheng & Parhi (2008) proposed a systematic high-speed VLSI implementation of the Separable-DWT based on hardware-efficient parallel FIR filter structures. High-speed 2-D DWT with computation time as low as $N^2/12$ can be easily achieved for an image with controlled increase of hardware cost. This design can save a large amount of multipliers or storage elements and also, be used to implement those 2-D DWT traditionally suitable for lifting or flipping-based designs. The parallel FIR structures are implemented by using cyclic convolution property. The Non separable 2-D DWT architectures are proposed by first transforming the column dimension filtering and downsampling into two FIR filters each with a filtering length of half of the original FIR filters. The 2-D DWT architectures are illustrated for 8x8 image coefficients and for the filter length of four.

Meher et al (2008a) proposed hardware-efficient Systolic-Like Modular Design for 2-DDWT. The overall computation is decomposed into two distinct stages, where column processing is performed in stage-1, while row processing is performed in stage-2. Using a new data-access scheme and a novel folding technique, the computation of both the stages are performed concurrently for transposition-free implementation of convolutional 2-D DWT. The proposed design can offer nearly the same throughput rate and requires the same or less the number of adders and multipliers as the best of the existing structures.
Lamoureux & Luk (2008) proposed an overview of low power techniques for implementation on Field Programmable Gate Arrays (FPGA). It also focuses on ongoing current research on commercial devices on circuit and architectural level techniques for low power implementation. A complete study on system, device, circuit and architectural level of designing of low power applications is discussed. Details regarding the power modeling and computer aided design of low power circuits are given.

Saeed Koko & Agustiawan (2009) presented parallel pipelined VLSI Architectures for Lifting-based 2-D DWT. For real-time applications of 2-D DWT with demanding requirements in terms of speed and throughput, 2-parallel and 4-parallel pipelined lifting based VLSI architectures for lossless 5/3 and lossy 9/7 algorithms are proposed. The overlapped scan method working in parallel for scanning the image coefficients is adopted to derive pipelined architectures. This is done to reduce the internal memory between the row and column processor. The column processor works in parallel with the row processor during the computation of lifting coefficients.

Lai et al (2009) presented a high performance and memory efficient pipelined architecture with parallel scanning method for 2-D lifting based DWT suitable for JPEG 2000 applications. It consists of two 1-D DWT cores and a 2×2 transposition register array. The 1-D DWT core consumes two input data and produces two output coefficients per cycle. The critical path is one multiplier delay only. The coefficients at the same column are scanned along the row direction and fed into the column processor by using parallel scanning method. The internal buffer size and on-chip memory requirement is reduced. The architecture can be used to process 30 frames with HDTV1080p (full HD) pictures per second at 100 MHz.
Shi et al (2009) proposed an efficient folded architecture by converting the serial operations of the data flow of the lifting schemes with the parallel operations by employing parallel and pipeline techniques. The optimized architecture has short critical path latency and efficient folding technique is used to derive an efficient architecture. The hardware utilization is improved by using folding technique. Multipliers are replaced by shift-add operations.

Anirban Das et al (2010) proposed an efficient architecture for 3-D DWT for image and video compression algorithm. This architecture forms the first lifting based 3-DDWT architectures without group of pictures restriction. The new computing technique based on analysis of lifting signal flow graph minimizes the storage requirement. The architecture has reduced memory referencing, low power consumption, low latency and high throughput. The proposed architecture offers a speed of 321MHz suitable for real time image compression with larger frame dimensions.

Mohanty (2010) proposed a high throughput-scalable parallel and pipeline architecture for multilevel 3-DDWT. The computation of 3-D DWT for each level of decomposition is split into three distinct stages, and all the three stages are implemented in parallel by a processing unit consisting of an array of processing modules.

Salehi & Amirfattahi (2011) proposed a block-based 2-D DWT structure with new scan methods for overlapped sections. The main challenge in 2-D DWT structure is the amount of internal memory required to produce wavelet coefficients. It needs fixed memory size depend on input frame size, the memory size is parametric and in trade off with the number of output memory access.
Tian et al (2011) presented an efficient Multi-Input/Multi-Output VLSI architecture (MIMOA) for 2-D DWT. This is a high-speed architecture with computing time as low as $N^2/M$ for an $N \times N$ image with controlled increase of hardware cost, where $'M'$ is the throughput rate. High processing speed can be achieved when multiple row data samples are processed simultaneously. Time multiplexing technique is adopted to control the increase the hardware cost of multiple input, multiple output array architecture.

Ghantous & Bayoumi (2011), proposed a parallel and pipelined VLSI Architecture for 2-D DWT. An area-efficient, parallel and pipelined architecture is proposed with a modified image scan coupled with multiplier-free multiplications. The parallelism is achieved using block based processing. The multipliers are eliminated and replaced by add/shift operations requiring minimal hardware.

Mohanty & Meher (2011) proposed memory efficient modular VLSI architecture for high-throughput and low-latency implementation of multilevel lifting 2-D DWT. A modular and pipeline architecture for lifting-based multilevel 2-D DWT is discussed without using line-buffer and frame-buffer. Overall area-delay product is reduced by appropriate partitioning and scheduling of the computation of individual decomposition-levels. The processing for different levels is performed by a cascaded pipeline structure to maximize the hardware utilization efficiency (HUE).

Kathuria et al (2011) presented a review of clock gating technique to reduce dynamic power by reducing unnecessary clock activities inside the gated combinational circuits. Clock gating can save power by reducing unnecessary clock activities inside the gated module. Latch based clock gating scheme is proposed to overcome the hazards in digital circuits. Glitches and
hazards problem is resolved due to the clock gating technique applied at the controller and target side to save power than other existing methods.

Zhang et al (2012a) proposed a high speed and reduced area 2-D DWT architecture. To reduce the latency in the existing method the intermediate results are recombined and stored to reduce the number of pipeline stages. Two input/ two output parallel scanning architecture uses three registers as transposition buffer between the row-column filters yielding higher efficiency.

Wimer & Koren (2012) proposed optimal fan out of clock network for power minimization by adaptive clock gating techniques. A probabilistic model of clock gating network based on flip flops toggling probabilities and process technology parameters is discussed. The grouping of flip flops for a joint clock gating to reduce switching power consumption is dealt in this architecture. The resulting clock gating methodology achieves 10% power saving of the total clock tree switching power.

Zhang et al (2012) proposed a pipeline architecture for multilevel 2-D DWT using Non-Separable approaches to reduce transposition and frame memory. The Non Separable devices have small latency but the computational time increases with the required number of levels of decomposition.

Mohanty et al (2012) proposed a linear systolic array from the suitably segmented dependence graph for parallel and pipeline implementation of lifting based DWT. The latency required is less and does not require complex control signals suitable for different block sizes.

Mohanty & Meher (2013) proposed memory efficient architecture for convolution based generic structure for the computation of three level 2-D Daubechies and bi-orthogonal DWT. The architecture does not require frame
buffer. The limitations of this architecture is that it requires more arithmetic components than the lifting based structures offering significant saving of area and power with smaller memory size and clock period. In this architecture the DWT level are computed concurrently to avoid frame buffer. Parallel data access is applied in each DWT level to reduce memory complexity.

Masahiro & Hitoshi (2013) proposed a Non Separable 2-D DWT with the reduced number of lifting steps and is compatible with the Separable 2-D DWT of JPEG 2000 lossless and lossy standard. The Non Separable has reduced lifting which has reduced latency suitable for high speed image processing applications.

2.1.2 Limitations of the Existing Architectures

The performance and implementation of various Lifting based DWT architectures are influenced by several design factors, like latency, hardware complexity, memory/storage, and speed for multimedia applications. Their impact varies based on the application, the wavelet coefficients, and type of architecture. The traditional implementation of lifting based DWT needs extra memory for storing the intermediate computational results. Moreover, for real time image compression, the DWT has to process massive amounts of data at high speeds. All these issues lead to difficulties in implementing DWT. The finite length of signal processed by using the wavelet filter leads to the edge effect. The JPEG2000 standard employs the symmetric extension at the boundaries to eliminate it. The traditional extension arithmetic needs additional memory units, and the operations consume much power and area.
The various architectures for lifting based DWT differ in terms of the required numbers of the multipliers, adders and registers, as well as the amount of accessing external memory, leading to an increase in the hardware cost and power consumption of design. In addition to improving the efficiency of the existing architectures, the objective of this research work is to improve the hardware utilization, and capacity to handle multiple data streams for the calculation of the Lifting 2-D DWT, for low power and high speed image processing applications.