Regular Article

An Efficient Architecture of Forward Transforms and Quantization for H.264/AVC Codecs

Xuan-Tu Tran, Van-Huan Tran

VLSI Systems Design Group, SIS Laboratory, VNU University of Engineering and Technology, Hanoi, Vietnam

Correspondence: Xuan-Tu Tran, tutx@vnu.edu.vn

Manuscript communication: received 29 May 2011, revised 26 June 2011, accepted 28 June 2011

***Abstract*– Thanks to many novel coding tools, H.264/AVC has become the most efficient video compression standard providing much better performance than previous standards. However, this standard comes with an extraordinary computational complexity and a huge memory access requirement, which make the hardware architecture design much more difficult and costly, especially for realtime applications.**

In the framework of H.264 codec hardware architecture project, this paper presents an efficient architecture of Forward Transform and Quantization (FTQ) for H.264/AVC codecs in mobile applications. To reduce the hardware implementation overhead, the proposed design uses only one unified architecture of 1-D transform engine to perform all required transform processes, including discrete cosine transform and Walsh Hadamard transform. This design also enables to share the common parts among multipliers that have the same multiplicands. The performance of the design is taken into consideration and improved by using a fast architecture of the multiplier in the quantizer, the most critical component in the design. Experimental results show that our architecture can completely finish transform and quantization processes for a 4:2:0 macroblock in 228 clock cycles and the achieved throughput is $445Msamples/s$ at $250MHz$ operating frequency while the area overhead is very small, $147755\mu m^2$ (approximate $15KGates$), with the $130nm$ TSMC CMOS technology.

***Keywords*– ISO/IEC 14496-10 AVC, ITU-T Recommendation H.264, VLSI architecture for H.264 codec, Forward Transforms and Quantization, integer DCT, quantization.**

This work has been partly presented at the 2011 IEEE Symposium on Design and Diagnostics of Electronic Circuits and Systems (IEEE DDECS 2011) at Cottbus, Germany in April 2011.

1 Introduction

THE H.264 Advanced Video Coding (H.264/AVC) [1] is known as the lastest and most efficient video compression standard providing better video quality at a lower bit-rate than previous standards. The standard is recommended by the Joint Video Team (JVT) formed by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). It contains a rich set of video coding tools to support a variety of applications ranging from mobile services and video conferencing, digital broadcast, to IPTV, HDTV and digital storage media. The general architecture composed of functional blocks of H.264/AVC encoder is depicted in figure 1. Compared with previous standards such as MPEG-4 [2], H.263 [3], and MPEG-2 [4], H.264/AVC can achieve 39%, 49%, and 64% of bit-rate reduction, respectively [5]. To achieve high compression ratio, H.264/AVC has adopted several advances in coding technology to remove spatial and temporal redundancies. However, it makes the hardware implementation becomes more complex and costly. Many designers have paid much attention to the complexity reduction of motion estimation and/or entropy coding parts [6, 7]. However, the

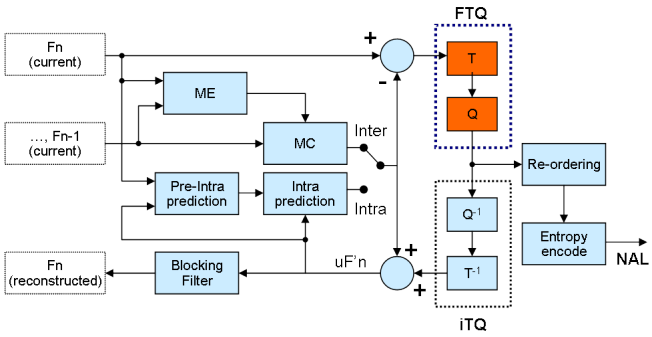

Figure 1. H264/AVC encoder – a functional architecture.

design and implementation of forward transforms and quantization (FTQ) part should also be considered for minimizing the system hardware overhead.

By specifying a set of integer transforms for small block-sizes, which are integer Discrete Cosine Transform (DCT) and Walsh Hadamard transform, it has really reduced the computational complexity as important as blocking artifacts [8]. In addition, thanks to the advanced procedures prensented in [9], these transforms can be easily realized with some required shifting and adding operations. In fact, in H.264/AVC standard, the size of transforms is variable depend-

ing on the profile used in the encoders, where 4×4 block-size and 2×2 block-size transforms are primitive components. Larger transforms, which used in adaptive block-size transforms (ABTs), are more suitable for High-Definition (HD) video. In this paper, only the 4×4 block-size transforms will be discussed but the same principle can be applied for the other sizes of transforms, or even larger transforms can be converted to the 4×4 block-size transforms [10].

Previous works have already been successes in hardware implementation of transforms and quantization. Chih-Peng Fan and Yu-Lin Cheng [11] proposed a design with a high through-put and low latency architecture using Canonical Signed Digit (CSD) multiplier for shared quantization/inverse-quantization. In [12], Yu-Ting Kuo presented an area-efficient architecture using direct 2-D transform method. Whereas, in [13–15] is proposed a multi-transform architecture used for variable adaptive block-size transforms. Generally, these works used two separate 1-D transforms in cascading to carry out a 2-D transform or to implement a direct 2-D transform. Obviously, the advantage of these methods is that they can achieve a high throughput in transforms. However, the disadvantage is that their hardware implementation area could not be significantly reduced, even in [15] the quantization parts are intently integrated into transform steps. In addition, the fact is that the bottleneck of encoders mostly comes from motion estimation and/or entropy coding modules rather than transforms and quantization. Optimizing the design for throughput is therefore less important than other objectives such as overal performance providing real-time processing capacity or hardware implementation area of the whole system.

In this work, we propose a transform architecture using only one unified 1-D transform module in order to trade-off between throughput and area overhead. With some improvements in control part, this architecture is able to perform integer DCT-based transforms as well as Hadamard transforms. In addition, to improve the system performance and get more area-efficiency, we also present a particular architecture of multiplier in the quantizer. In where, a shared module (called pre-multiplier) is intently used for multipliers have the same multiplicands. The FTQ architecture is designed to be used as part of a low power H.264/AVC encoder for mobile applications. The proposed architecture is then implemented using the $130nm$ TSMC CMOS technologies. It can completly finish the transform and quantization processes for a 4:2:0 macroblock in 228 clock cycles and therefore can achieve a throughput of $445Msamples/s$ at $250MHz$ operating frequency while the area overhead is very small, approximate $15KGates$.

The remaining part of this paper is organized as follows: Section 2 briefly recalls some backgrounds of transform and quantization algorithms in H.264/AVC coding, including integer DCT-based transform and Hadamard transform. The proposed architecture for a forward transform and quantization will be presented in Section 3. Experimental results will be presented and discussed in Section 4. Finally, conclusions and perspectives will be given in Section 5.

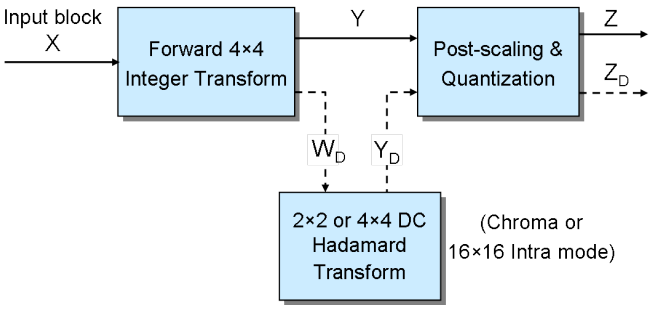

Figure 2. Transform and quantization flow diagram.

2 Forward Transforms and Quantization Algorithms in H.264/AVC

In H.264/AVC video coding standard, the residual frame of the prediction, which is the difference of the original frame and the predicted frame, is partitioned into fixed-size of macroblocks. As usually, a macroblock is composed of 16×16 luminance (Y) samples, 8×8 chroma blue (C_b) samples, and 8×8 chroma red (C_r) samples in the case of 4:2:0 chroma subsampling format. At a smaller level, macroblocks are subdivided into blocks of 4×4 samples for encoding. Each macroblock has its own information on quantization parameters (QPs), coded type (Intra mode or Inter mode) and prediction mode. The transform and quantization flow for those blocks can be illustrated in Figure 2.

According to this flow, the input block X is first transformed using integer DCT-base method. The transformed coefficients are then post-scaled and quantized. In the 16×16 Intra-prediction mode, DC coefficients of all transformed residual blocks are grouped into an array of 4×4 before being sent to Hadamard transform. Details of these processes are described in mathematical models in the following.

2.1 4×4 Forward Transforms

Integer DCT-based Transform: The integer DCT-based transform, which applied to a residual 4×4 blocks (denoted by matrix X), is defined in H.264/AVC as the following:

$$Y = CXC^T = C(CX)^T, \tag{1}$$

where

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}.$$

Hadamard Transform: The Hadamard transform which applied to a 4×4 luminance DC block (denoted by matrix W_D) in 16×16 Intra-prediction mode, is defined as the following:

$$Y_D = HW_DH^T = H(HW_D)^T, \tag{2}$$

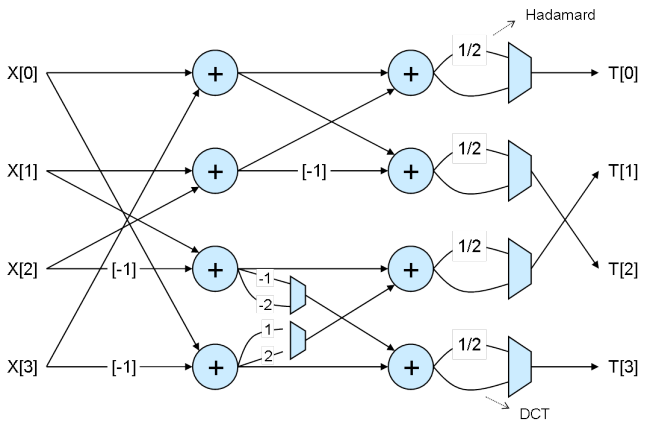

Figure 3. Hybrid 1-D transform architecture for integer DCT-based and Hadamard transforms.

where

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} /2.$$

In general, both integer DCT-based transform and Hadamard transform are normally formed by two duplicated *cores* of 1-D transform, where the *core* is a matrix multiplication either C^T or H^T. The 2-D transforms are carried out by applying the input block to the *core*; the immediate results are re-arranged by transposing operations and then re-applied to the *core*. Obviously, the specification of the matrixes C and H in which only coefficients of ± 1 and/or ± 2 are available. These transformations are multiplication-less and purely require a few of *add* and *logical shift* operations. On the other hand, the dynamic range of data is also estimated to reduce the overhead in computating processes. With 8-bit precision of the pixel data, the dynamic range of outcomes of integer DCT-based transform is 16-bit.

In here, we have already modified the matrix H by scaling of $1/2$ to preserve the arithmetic operations of Hadamard transform in 16-bit precision as of integer DCT-based transform. Then, in the quantization process of the DC blocks, the result will be rescaled of 2. By this way, all 4×4 forward transforms are completely handled in 16-bit precision.

Figure 3 shows a hybrid and fast 1-D transform diagram for processing 4 samples. The diagram is in the shape of butterfly diagram and is used for two types of transform. There are some multiplexers to select the shift factors (or scaled factors) in computations of each transform type. This diagram is a great inspiration to design the architecture of transform module.

2.2 Quantization

H.264/AVC standard defines a set of 52 values of quantization steps ($Qstep$). These values are indexed by QP and to be determined in the range of 0 to 51. As introduced above, the values of quantization parameters are associated with macroblocks (also blocks within a macroblock). Thanks to the wide range of QP, an encoder is able to accurately and flexibly control the trade-off between its bit-rate and quality [16].

Table I
Multiplication factor (MF)

QP%6	Positions (0,0),(2,0),(2,2),(0,2)	Positions (1,1),(1,3),(3,1),(3,3)	Other positions
0	13107	5243	8066
1	11916	4660	7490
2	10082	4194	6554
3	9362	3647	5825
4	8192	3355	5243
5	7282	2893	4559

Basically, forward quantization can be defined as follows:

$$Z_{ij} = round\left(\frac{Y_{ij}}{Qstep}\right), 0 \le i,j \le 3 \tag{3}$$

To avoid division operations, this equation can be represented in another way [9]:

$$Z_{ij} = round\left(W_{ij} \times \frac{MF}{2^{qbits}}\right), 0 \le i,j \le 3 \tag{4}$$

In consequence, the quantization can be computed as follows:

$$\begin{aligned} |Z_{ij}| &= \left(|W_{ij}| \times MF + f\right) \gg qbits, \\ sign\left(Z_{ij}\right) &= sign\left(W_{ij}\right), 0 \le i,j \le 3, \end{aligned} \tag{5}$$

where $qbits = 15 + floor(QP/6)$, MF is a multiplication factors matrix (see Table I) and f is an additional factor, $f = 2^{qbits}/3$ if the block is coded in Intra mode, and $f = 2^{qbits}/6$ if the block is coded in Inter mode. Especially, the quantization for a DC block is implemented as follows (it has already rescaled by 2 due to scaling by $1/2$ in transform):

$$\begin{aligned} |Z_{D(ij)}| &= \left(|Y_{D(ij)}| \times MF_{00} + 2f\right) \gg qbits, \\ sign\left(Z_{D(ij)}\right) &= sign\left(Y_{D(ij)}\right), 0 \le i,j \le 3, \end{aligned} \tag{6}$$

where MF_{00} is the multiplication factor at position $(0,0)$.

The innovation of quantization in H.264/AVC is the definition of $Qstep$. In where, $Qstep$ is non-uniform (or non-linear according to QP) and doubled in size if QP increases by 6. Therefore, whenever QP is changed by the encoder, the values of MF factor matrix is also changed as consequence, but it absolutely depends on the value of QP%6 (as shown in Table I). Besides, it does not require a lot of memory elements to store MF factors, only 18 values for full range of QP.

Similar to the transform part, the quantization has also been simplified to obtain low-complexity in a manner of avoiding division and floating point operations.

3 The Proposed Architecture

In this section, we present a hybrid architecture for DCT and Hadamard forward transforms and quantization of

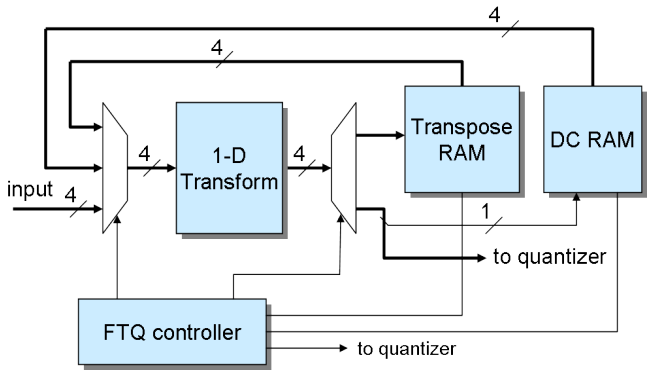

Figure 4. Architecture of forward transform.

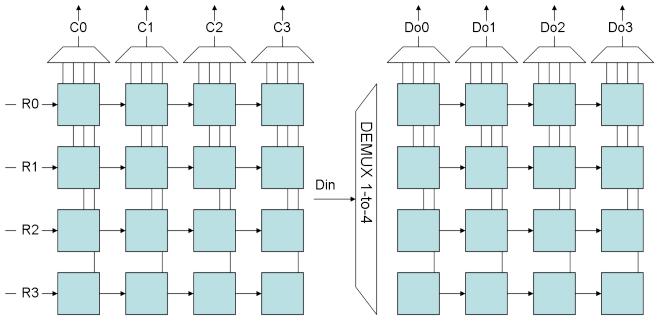

Figure 5. Transpose RAM (*left*) and DC RAM (*right*).

4×4 blocks. While the design of transform is only intended to area-efficiency by using 1-D transform module for all transformations of 4×4 blocks, the design of quantizer is carefully taken into account in order to improve on both performance and implementation area. This design is capable to process 4 samples at the same time. The details will be described as follows:

3.1 Transform Module

By sharing the only 1-D transform module, the second 1-D transform process could not be started until the first 1-D transform process had finished on the entire block. Therefore, it is necessary to have a memory buffer for storing and transposing the temporary data. Figure 4 shows the architecture of forward transform module with the sample-width of the datapath.

The architecture is simply composed of three main components: 1-D Transform module, Transpose RAM module, and DC RAM module. In addition, there are other components included in this architecture such as multiplexer and de-multiplexer for arbitrating the dataflow. The input data and output data of the transform module are 4 samples, equivalent to 64-bit (4×16-bit). To have better view, some detail control signals have been hidden.

The activity of the module can be easily realized through the list of all states (corresponding to the dataflow):

- *State_1: Input* $\Rightarrow$ *1-D Transform* $\Rightarrow$ *Transpose RAM*
- *State_2: Transpose RAM* $\Rightarrow$ *1-D Transform* $\Rightarrow$ *Output*
- *State_3: DC RAM* $\Rightarrow$ *1-D Transform* $\Rightarrow$ *Transpose RAM*

These states have a length of 4 clock cycles. By controlling the sequence of these states, a general block will be executed in the order of two states {*state_1*; *state_2*}, while a DC block will be executed in the order of two other states {*state_3*; *state_2*}.

1-D Transform: The 1-D transform module is a hybrid transform as illustrated in Figure 3. All multiplexers in this module are controlled by a selection signal which configures the module's activities as integer DCT-based transformation or Hadamard transformation. The responsibility of this module is one clock cycle and the throughput is therefore 4 *samples/clock cycle*. A higher throughput can be easily obtained by using several 1-D transforms in parallel.

Transpose RAM and DC RAM: The purpose of Transpose RAM is to store and transpose data for transformation processes. DC RAM is used to store luminance DC coefficients of a transformed macroblock. Basically, Transpose RAM and DC RAM are matrices of 4×4 16-bit registers (as shown in Figure 5). The input and output of Transpose RAM are 4 samples width for reading/writing accesses to a row/column data. Whereas the input of DC RAM is only one sample width for writing one DC coefficient at a time and the output is 4 samples width as Transpose RAMs.

The writing operations of Transpose RAM are enabled whenever valid data are ready at the output of 1-D transform module. These occur in 4 succesive clock cycles of the first 1-D transform process. The reading operations occur in next 4 clock cycles to get out the column-wise data for the second 1-D transform process. The registers in a row of Transpose RAM are connected in series. Thus, the data stored in a register will be automatically shifted into the back register in next clock. By this means, Transpose RAM will be filled up with new data in 4 clock cycles of the writing operations.

DC RAM is a bit different from Transpose RAM in their structures where the registers are independent from each other. This buffer is useful and will be active in 16×16 Intra prediction mode. In that case, the DC coefficients of any transformed luminance blocks are extracted and written into DC RAM. Therefore, the writing operations of DC RAM take place in only one clock cycle at the time where the earliest data is valid. The reading operations are enabled in 4 clock cycles when the last block of a luminance macroblock is completely transformed. Besides, the reading/writing address signals of DC RAM are directly controlled by FTQ controller.

Comparing with the cascading architecture using two separate 1-D transform modules [17], our architecture is required a bit challenging in designing the controller module but it has absolutely saved the hardware resource by the total cost of an 1-D transform module.

3.2 Quantizer Module

The quantizer can be easily realized from equations 5 and 6, as depicted in Figure 6. It consists of four quantization cores and some common parts: MF_ROM module, DIVIDER_BY_6 module, and F_CALC mod-

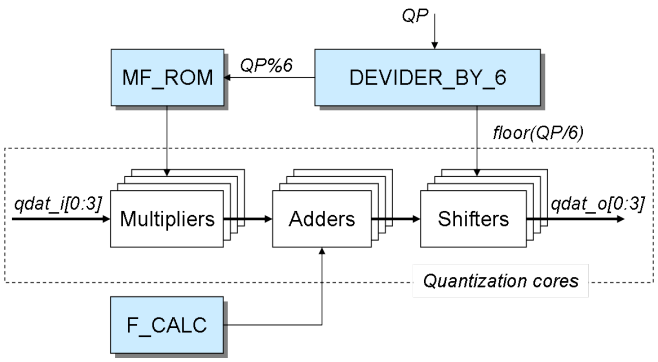

Figure 6. Quantizer architecture.

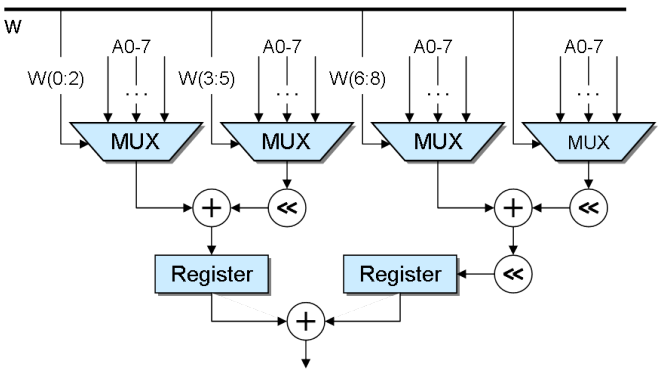

Figure 8. Multiplier for quantizer using the pre-multiplier element.

ule. These common modules are shared to the 4 quantization cores. Actually, DIVIDER_BY_6 module is possible to share with de-quantizer module.

DIVIDER_BY_6 module is a combinational block to calculate the value of $QP\%6$ and *floor* $(QP/6)$ as well. In some related works, it was designed as common look-up-tables (LUTs), such as [11]. This design may take lots of memory utilization because we have up to 52 values of QP. MF_ROM module is a ROM block for storing 18 constant values of MF factors. Accessing to a batch of MF factors is addressed by $QP\%6$ signal. F_CALC module is a combinational block to calculate the additional factor f based on the coded macroblock type (either Intra mode or Inter mode) and the block type (residual block or DC block).

Regarding to the multiplier design, when the size of multipliers are large (15-bit of $qdat\_i$ and 14-bit of MF), it can mostly impact to the performance of the quantizer as a result of large latency. For this reason, we have deeply investigated the design of multiplier, which will be presented in next paragraph to minimize the quantization latency.

A Fast and Highly Shared Multiplier: The fast multiplier that we proposed is a conditional multiplier. The idea is to build a basic element (called pre-multiplier) which is multiplier of MF factor with all possible 3-bit numbers (as shown in Figure 7), where $A_i = i * MF$, with $0 \leq i \leq 7$. In fact, we do not need to carry out A0, A1, A2, and A4 on this module when these signals can be directly driven from MF signal.

Consequently, the 15-bit multiplier using the pre-multiplier element is explored as Figure 8. Each group of 3-bit vector (so it has 5 groups) is multiplied with the multiplicand W by controlling a multiplexer to select the equivalent result from the pre-multiplier element. There are some registers are inserted at the adders' outputs to cut down the combinational paths of the multiplier. By this way, we have improved the performance of the multiplier.

Obviously, the pre-multiplier can be shared among 5 groups of 3-bit multipliers. Since the pre-multiplier takes 4 adder blocks, we save $(4 \times 4 = 16)$ adder blocks. In addition, from the MF factors table (Table I), $qdat\_i[0]$ and $qdat\_i[2]$ have the same MF factor, $qdat\_i[1]$ and $qdat\_i[3]$ have the same MF factor, therefore the pre-multiplier is also possible to share between two quantization multipliers which have the same MF factor. In consequence, our quantizer can be totally saved up to $2 \times (16 \times 2 + 4) = 72$ adder blocks. Certainly, several multiplexers will be required to replace these adders but the hardware utilization is significantly reduced.

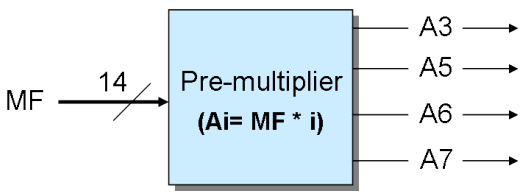

Figure 7. The pre-multiplier element.

3.3 Pipeline Operation

To achieve better performance, the proposed architecture is controlled to operate in pipeline mode. Figure 9 shows the timeline of the whole coding process.

The pipeline has three states as identified by S_1, S_2, and S_3. In where, S_1 (4 clock cycles) launches the first stage of 1-D transform on block B1 while previous block B0 is still quantized to the end. S_2 (4 clock cycles) launches the second stage of 1-D transform while starts quantizing block B1. S_3 (1 clock cycle) prepares loading new block B2 into transform module while block B1 is still quantized, valid data at the output are ready. Therefore, it normally takes 12 clock cycles to complete transform coding for a block.

Summary, by the pipelined schedule, our design will take 9 clock cycles on average to process a block and this is equivalent to 228 clock cycles to complete

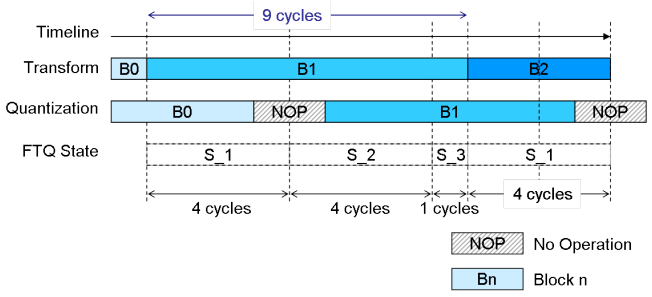

Figure 9. Three states of pipeline operation.

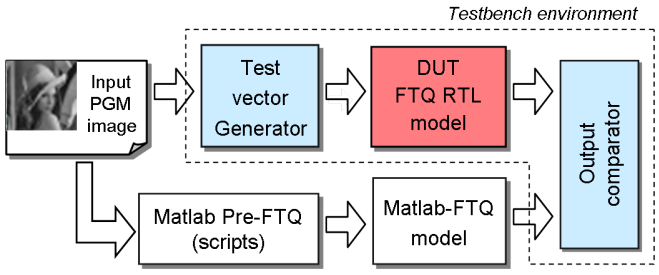

Figure 10. Verification model for the design.

the transform and quantization processes for all 4×4 blocks within a 4:2:0 macroblock.

4 Verification and Implementation

The architecture was modeled using VHDL language and simulated on Synopsys VCS tool. To verify the functionality of the design, we developed a simple simulation environment as described in Figure 10. Then, this environment is also used to verify the design before and after the layout implementation.

In this environment, we developed a software model of FTQ architecture on Matlab, which is used for testing purpose only. The input data used in this simulation is a PGM (portable graymap format) image. This image is provided to both Matlab model and hardware model. Then, the outputs of the both hardware and Matlab-based model will be compared to each other by using the developed testbench.

The proposed design was first prototyped on a Virtex-II Xilinx FPGA (XC2V1500-6) and then implemented using the $130nm$ CMOS technology from TSMC. The achieved operating frequency is $115MHz$ with the FPGA implementation, and $250MHz$ with the $130nm$ TSMC CMOS technology. In both cases, the design takes 228 clock cycles to complete the transform and quantization processes for the entire 4:2:0 macroblock. The transformation and quantization throughput is estimated of $204Msamples/s$ and $445Msamples/s$, respectively. The implementation area of the proposed design is $147755\mu m^2$ with the $130nm$ TSMC CMOS technology. Figure 11 presents the hardware implementation overhead of each component in the design, in where quantizer is the bigest component and occupies 56% area cost of the FTQ design.

Table II shows the implementation report of different designs on hardware overhead, operating frequency, and throughput. From this report, our design has a much higher throughput and a lower hardware overhead compared to the design presented in [15], while both the designs have the same data width (4-bit). The designs presented in [11, 14, 18] have higher throughputs than our design but the data width of these designs are 4 times (even 8 times) larger than the data width of our design (16-bit, 16-bit, and 32-bit, respectively). Certainly, these designs therefore occupy much more hardware implementation areas than our design. In particular, the design presented in [19] with 1-bit data width but the hardware overhead is higher than ours while the throughput is much lower than our proposal. With the achieved throughput, our design totally responses to the need of H.264/AVC encoders/decoders while it is very suitable for efficient hardware implementation.

5 Conclusion

Being equipped by many advanced coding tools and techniques, the H.264/AVC has become the most efficient video compression standard providing better video quality at a lower bit-rate than previous ones. However, it is very difficult and costly to implement this standard on hardware in order to meet the requirement of real-time applications because of its computational complexity. In recent years, several hardware architectures have been proposed for H.264/AVC codecs which focus only on optimizing and adapting prediction algorithms and/or entropy coding techniques.

We presented in this paper a cost-efficient and high-performance forward transform and quantization hardware implementation for real-time H.264/AVC encoders in mobile applications. To minimize the hardware overhead, the proposed design used only one unified architecture of 1-D transform engine to perform all required transform processes, including discrete cosin transform and Walsh Hadamard transform. In addition, the detail architecture is also investigated carefully to optimize as much as possible both area cost and performance. The proposed architecture is then implemented on both FPGA and ASIC technologies. It is experimentally verified to work at $115MHz$ on a Xilinx Virtex II FPGA and to work at $250MHz$ on a TSMC $130nm$ CMOS implementation. The area overhead of this design is very small, $147755\mu m^2$ (approximate $15KGates$) with ASIC implementation. In addition, the design is able to complete transform and quantization processes for a macroblock in 228 clock cycles. Consequently, the achieved throughput is $204Msamples/s$ and $445Msamples/s$ for FPGA and ASIC implementations, respectively, while the data width is 4-bit. The proposed FTQ architecture is therefore proved to achieve a high performance with a lower area cost than the previous works.

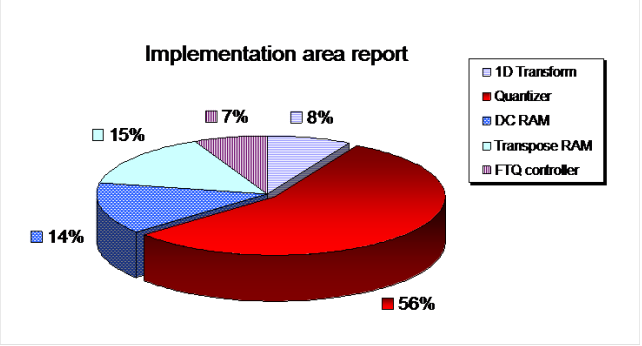

Figure 11. Hardware implementation overhead on different components of the FTQ architecture.

Table II
Implementation report on hardware overhead, operating frequency, and throughput between different designs

Design name	**Technology**	**Data width** (*bit*)	**Operating freq.** (*MHz*)	**HW overhead** (*gates*)	**Throughput** (*Msamples/s*)
Pastuszak [14]	TOWER 0.18μm	32	77	162,122	2,464
Chih-Peng Fan [11] (using CSD multipliers)	Xilinx XC2V1500	16	99.15	135,306	1,586
Kordasiewicz [18] (area-optimized)	TSMC 0.35μm	16	68	51.619	644
Heng-Yao Lin [15]	TSMC 0.35μm	4	32	30,785	273
Tasdizen [19]	ASIC 0.18μm	1	210	23,162	21.5
This work (FPGA)	Xilinx XC2V1500	4	115	24,052	204
This work (ASIC)	TSMC 0.13μm	4	250	15,033	445

In the framework of H.264 codec hardware architecture project, an inverse transforms and quantization (iTQ) architecture is being developed to complete the transforms and quantization part of H.264/AVC encoders. This work will be presented in an other report.

Acknowledgment

This work is partly supported by Vietnam National University, Hanoi (VNU) through research project No. QGDA.10.02 (VENGME). This work has been partly presented at the 2011 IEEE Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS) at Cottbus, Germany in April 2011 [20]. The authors would like to thank reviewers for their constructive and helpful comments.

References

[1] "ITU-T recommendation and international standard of joint video specification," *ITU-T Rec. H.264/ISO/IEC 14496-10 AVC*, March 2005.

[2] *Information Technology – Coding of Audio-Visual Objects – Part 2: Visual*, ITU-T Std., 1999.

[3] "Video coding for low bit rate communication," *ITU-T Recommendation H.263*, February 1998.

[4] *Information Technology – Generic Coding of Moving Pictures and Associated Audio Information: Video*, ITU-T Std., 1996.

[5] A. Joch, F. Kossentini, H. Schwarz, T. Wiegand, and G. J. Sullivan, "Performance comparison of video coding standards using Lagragian coder control," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2002, pp. 501–504.

[6] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, July 2003.

[7] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with h.264/avc: tools, performance, and complexity," *IEEE Circuits and Systems Magazine*, vol. 4, no. 1, pp. 7–28, 2004.

[8] D. Marpe, T. Wiegand, and G. J. Sullivan, "The H.264/MPEG4 advanced video coding standard and its applications," *IEEE Communications Magazine*, pp. 134–143, August 2006.

[9] H. Malvar, A. Hallapuro, M. Karczewicz, and L. Kerofsky, "Low-complexity transform and quantization in H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 598–603, 2003.

[10] J.-K. Lee and K.-D. Chung, "DCT block conversion for H.264/AVC video transcoding," in *Euro-Par 2005 Parallel Processing*, Lisbon, Portugal, September 2005, pp. 919–927.

[11] C.-P. Fan and Y.-L. Cheng, "FPGA implementations of low latency and high throughput 4x4 block texture coding processor for H.264/AVC," *Journal of the Chinese Institute of Engineers*, vol. 32, no. 1, pp. 33–44, 2009.

[12] Y.-T. Kuo, T.-J. Lin, C.-W. Liu, and C.-W. Jen, "Architecture for area-efficient 2-D transform in H.264/AVC," in *Proc. 2005 IEEE Int. Conference on Multimedia and Expo*, Amsterdam, Netherlands, July 2005.

[13] J. D. Bruguera and R. R. Osorio, "A unified architecture for H.264 multiple block-size DCT with fast and low," in *Proc. 9th EUROMICRO Conference on Digital System Design (DSD)*, Cavtat near Dubrovnik, Croatia, August 2006, pp. 407–414.

[14] G. Pastuszak, "Transforms and quantization in the high-throughput H.264/AVC encoder based on advanced mode selection," in *Proc. IEEE Computer Society Annual Symposium on VLSI*, Montpellier, France, April 2008, pp. 203–208.

[15] H.-Y. Lin, Y.-C. Chao, C.-H. Chen, B.-D. Liu, and J.-F. Yang, "Combined 2-D transform and quantization architectures for H.264 video coders," in *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, vol. 2, May 2005, pp. 1802–1805.

[16] I. Richardson, "H.264/MPEG-4 Part 10: Transform and quantization," *VCodex Ltd White Paper*, March 2003.

[17] T.-C. Wung, Y.-W. Huang, H.-C. Fang, and L.-G. Chen, "Parallel 4x4 2D transform and inverse transform architecture for MPEG-4 AVC/H.264," in *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, May 2003, pp. 800–803.

[18] R. C. Kordasiewicz and S. Shirani, "ASIC and FPGA implementations of H.264 DCT and quantization blocks," in *Proc. IEEE International Conference on Image Processing (ICIP)*, September 2005, pp. III–1020–3.

[19] O. Tasdizen and I. Hamzaoglu, "A high performance and low cost hardware architecture for H.264 transform and quantization algorithms," in *Proc. 13th European Signal Processing Conference*, September 2005, pp. 4–8.

[20] X.-T. Tran and V.-H. Tran, "Cost-efficient 130nm TSMC forward transform and quantization for H.264/AVC encoders," in *Proc. 14th IEEE Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, Cottbus, Germany, April 2011, pp. 47–52.

Xuan-Tu Tran was born in Nghe An, Vietnam, in 1977. He received a B.Sc. degree in 1999 from Hanoi University of Science and a M.Sc. degree in 2003 from Vietnam National University, Hanoi, all in Electronics Engineering and Communications; and a Ph.D. degree in 2008 from the CEA-LETI, MINATEC (in collaboration with Grenoble INP), France in Micro Nano Electronics.

Xuan-Tu Tran has worked as a lecturer at Vietnam National University, Hanoi (1999-2003), as a research engineer at the CEA-LETI, MINATEC, France (2003-2008), and then as an assistant professor at the University of Engineering and Technology (UET), VNU Hanoi (2008-recent). He was a visiting/invited professor at the University Paris-Sud 11, France (2009, 2010), visiting professor at Grenoble INP in 2011. He is currently deputy director of the key laboratory for Smart Integrated Systems and head of VLSI Systems Design group. He is in charge for CoMoSy, VENGME projects for embedded systems and multimedia applications. His research interests include design and test of systems-on-chips, networks-on-chips, design-for-testability, asynchronous/synchronous VLSI design, and hardware architecture for multimedia applications.

He is a member of the IEEE, IEEE CAS, and the Executive Board of the Radio Electronics Association of Vietnam (REV).

Van-Huan Tran was born in Vietnam, in 1985. He received the Bachelor of Science in Electronics and Telecommunication from University of Engineering and Technology, Vietnam National University, Hanoi in 2007. He is currently working as a researcher of the key laboratory for Smart Integrated Systems – VLSI Systems Design group, University of Engineering and Technology, VNU Hanoi.

His research interests include System-on-Chip design and verification, FPGA-based design, embedded systems, VLSI systems/circuits design for multimedia application.