A Highly Pipelined VLSI Architecture for All Modes and Block Sizes Intra Prediction in HEVC Encoder

Cong Liu[1]*, Weiwei Shen[1], Tianlong Ma[1], Yibo Fan[1]*, Xiaoyang Zeng[1]

[1] State Key Lab of ASIC and System, Fudan University, Shanghai, China
* Email:11212020033@fudan.edu.cn, fanyibo@fudan.edu.cn

Abstract

The adoption of 35 prediction modes and quad-tree structure in intra coding of High Efficiency Video Coding (HEVC) significantly improves the coding efficiency. In this paper, a highly pipelined 16-pixel parallel VLSI architecture of intra prediction in HEVC encoder is proposed, supporting all prediction modes and all block sizes. Original pixels are used to help to decide prediction mode and block partition in the premise of negligible PSNR degradation, and a universal predictor is presented. In order to reduce internal buffers when scanning full-mode and full-size predictions in encoder, post-order traversal is applied to the quad-tree structure blocks. It takes 8967 cycles to complete the intra prediction of a whole 32x32 treeblock, including prediction and the decision of mode and block partition. This design is synthesized with TSMC 65nm CMOS technology. It can run at 600 MHz, supporting real-time encoding of 1080P@30fps video sequence.

1. Introduction

High Efficiency Video Coding (HEVC) [1] is the next generation of video coding standard developed by Joint collaborative team on video coding (JCT-VC), which is formed by ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG). It aims to double the compression rates in comparison with the latest consolidated standard, H.264/AVC [2], under the same visual quality.

As H.264/AVC, HEVC still adopts the compressing methods based on block. However, while H.264/AVC has only two block size partitions (4x4 and 16x16), a quad-tree structure is introduced in HEVC. The largest block is 64x64, which then can be divided recursively into smaller ones, and the minimum can be 4x4 [1].

Intra prediction is an efficient method to reduce the spatial data redundancy. The accurate intra prediction helps a lot to improve compression performance. It uses the neighbor pixels as references to predict the current block. Often, the reference pixels must be the already coded ones, and this requirement leads to the data dependency which brings in much latency time and decreases the hardware utilization. This bottleneck also exists in H.264/AVC, and many researches have been done to overcome it, such as in [3] [4], the process order of 4x4 blocks is modified. But in HEVC encoder, it will cause many difficulties for controlling by this way, for block sizes are changing recursively and the scan order is not merely simple zig-zag. To reduce data dependency, a VLSI architecture, two prediction engines included, is proposed in this design. One is based on the original pixels, and the other refers to the reconstructed samples. Highly pipelined design is realized in this architecture to enhance the throughput.

To reach further improvement of throughput, in [4], two data paths (4x4 luma and 16x16 luma /8x8 chroma) are designed for intra prediction in H.264/AVC encoder. But as mentioned above, the block size in HEVC changes from 4x4 to 64x64. It's really not proposed to supply individual data path for each block size because it is really an unaffordable hardware cost. Utilizing the consistency across different block sizes, a universal and fully pipelined architecture is proposed in this paper, supporting the predictions of 32x32, 16x16, 8x8 and 4x4 blocks. Not like [5], including two data paths (above and left), this design has only one.

It is known that in encoder, intra prediction needs to scan all the prediction modes and then decides which the best one is. In HEVC encoder, the block partition must be considered, too. Therefore, big enough internal buffers are often needed to store the intermediate data which cannot be abandoned before the best mode and block sizes are decided. To reduce the buffers as far as possible, postorder traversal is applied to the quad-tree structure blocks in this design, instead of the preorder traversal in the HEVC Test Model HM encoder software.

The rest of this paper is organized as follows: Section II presents the algorithm of HEVC intra prediction; Section III describes our proposed VLSI architecture; Section IV shows the implementation results; and, finally, Section V concludes this paper.

2. Intra Prediction Algorithm in HEVC

Intra prediction in HEVC is an extension of H.264/AVC. As shown in Fig.1, the total number of prediction modes is increased to 35 (Planar, DC and 33 Angular modes), while in H.264/AVC there are only 9 modes. And samples in the left, above, above-left, above-right and

978-1-4673-6417-1/13/$31.00 ©2013 IEEE

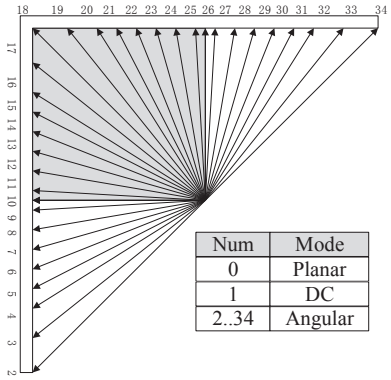

Figure 1. Intra prediction modes and reference pixels distribution in HEVC

below-left sides are all used as references for prediction.

The size of prediction unit (PU) ranges from 4x4 to 64x64. However, the 64x64 block is always divided into 32x32 or even smaller blocks to do the actual prediction. That means the effective intra prediction block sizes are only 32x32, 16x16, 8x8 and 4x4 [6], which are all supported in our proposed architecture.

In the algorithm of HEVC intra prediction, there are several innovated operations on the reference pixels before doing the actual prediction, such as substitution, filtering and projection between row and column.

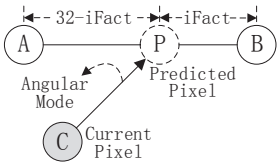

Figure 2. Angular prediction process

Among all modes, Planar and DC are mainly used for homogeneous regions while the Angular modes focus on directional texture, in which each predicted pixel is obtained by projecting its location to reference row or column as shown in Fig.2 and then using equation (1) to interpolate a value at 1/32 pixel accuracy.

$$P = \left((32 - \text{iFact}) \bullet A + \text{iFact} \bullet B + 16\right) \gg 5 \quad (1)$$

$$\text{iIdx} = \left((x+1) \bullet Angle\right) \gg 5 \quad (2)$$

$$\text{iFact} = \left((x+1) \bullet Angle\right) \,\&\, 31 \quad (3)$$

Which neighbor pixel is chosen as A or B is determined by iIdx and the location of current pixel. iIdx and iFact in modes 2 to 17 are calculated by equation (2) and (3), while in modes 18 to 34, x and y coordinates should be swapped. The parameter *Angle* is defined according to the angular modes.

3. Proposed VLSI Architecture

The proposed architecture aims to solve the problem caused by data dependency, support all prediction modes and block sizes and reduce the internal buffers. In the following parts of this section, details are shown.

3.1 Top-level Design

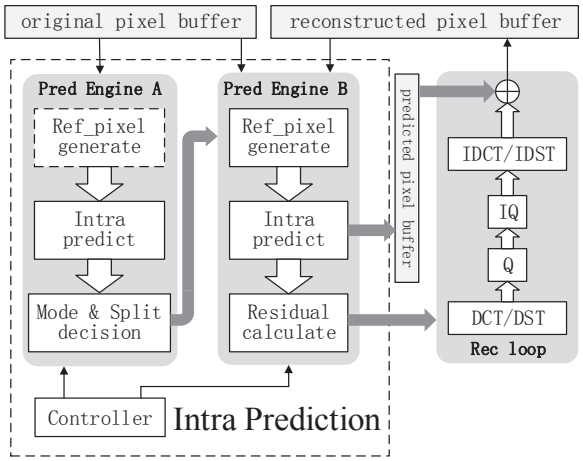

Figure 3. Top Architecture

During the process of the intra prediction in encoder, before the full-mode-prediction scan is performed on one block, the reconstruction of its neighbor block must have been done, for the reference samples are often already coded ones. This data dependency leads to the waiting time in the schedule as shown in Fig. 4(a). Furthermore, in HEVC, due to the transform of larger blocks after intra prediction, the reconstruction needs more cycles.

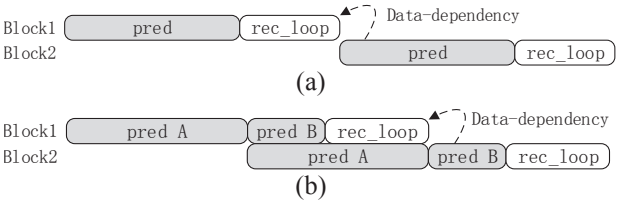

Figure 4. Timing diagram

Fig. 3 illustrates the proposed architecture, in which two prediction engines are used to reduce the waiting time. The engine A uses the original samples as the references, scans all the 35 prediction modes and also calculates the distortion to decide which mode and which block partition is the best. To imitate the normal intra prediction as much as possible, in engine A, the operations on references such as substitution, filtering and projection are still done, although the original pixels are always available. The engine B, which generates the residuals, refers to the reconstructed samples as usual, but only does the prediction of the best mode.

A test has been done in HM9.0 encoder software, keeping most processes unchanged, except that the original pixels, instead of reconstructed ones, are used as references when scanning the predictions to choose the best mode and the best block partition. The compared result is shown in Table 1, which indicates a negligible degradation of compress performance.

Table 1. Performance Comparison

Sequence	Δ PSNR (dB)			Δ BD-rate (%)		
	Y	U	V	Y	U	V
Class A	-0.0049	-0.0009	0.0005	0.1	0.0	-0.1
Class B	-0.0063	-0.0005	-0.0014	0.1	0.0	0.0
Class C	-0.0096	-0.0018	-0.0015	0.1	0.0	0.0
Class D	-0.0090	-0.0015	-0.0012	0.1	0.0	0.0
Class E	-0.0079	-0.0015	-0.0032	0.3	0.2	0.1
Class F	-0.0116	0.0018	-0.0043	0.1	0.0	0.0
Average	-0.0082	-0.0007	-0.0018	0.1	0.0	0.0

Because most of the predictions refer to the original pixels which can be used without waiting, the data dependency is alleviated to some extent. The schedule in the proposed architecture is shown in Fig. 4(b), and only the prediction in engine B depends on the result of the reconstruction loop.

Comparing Fig. 4(a) and Fig. 4(b), the reduction of the latency between two effective outputs (residuals) is determined by the cycles cost in prediction engine B. As described above, engine B only do the prediction using the best mode. There is no need to single out the best mode and the best block size, which means the complex process for getting the distortion is eliminated here. So the work in engine B is very simple and takes quite a few cycles, which helps a lot to achieve high throughput.

3.2 Universal Predictor for Intra Prediction

In order to improve the throughput more, the proposed architecture is 16-pixel parallel, and according to the consistency across prediction directions and block sizes in HEVC, a universal predictor for intra prediction, three-stage pipelined as shown in Fig. 5, is presented in this section, supporting all modes and all effective block sizes (32x32, 16x16, 8x8, 4x4). It is fully pipelined and easily controlled to complete the full-mode and full-size predictions scan in encoder.

There are three simple signals for controlling: "Mode" indicates which prediction mode is currently employed, "Cu_size" expresses the current block size, and "Blk4x4_num" illustrates the location of one 4x4 block including the 16 pixels which are being operated on simultaneously. These three control signals contain all the information needed for prediction.

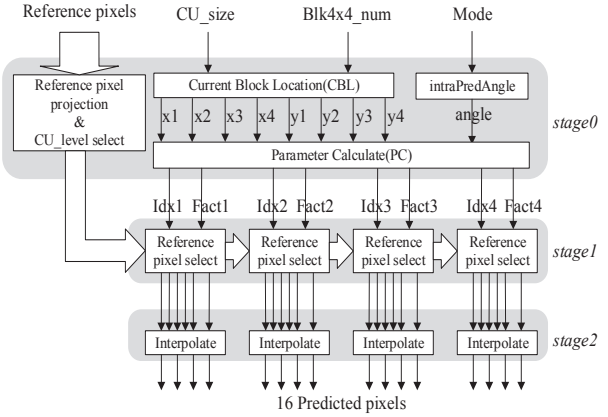

Figure 5. Intra Predict Data Path

The work in each stage is illustrated as follows:

Stage0: Calculate the parameters, iIdx and iFact, which remain constant when operating on the same row or column of samples according to equation (2) and (3). The projection of the reference pixels is also done in this stage.

Stage1: Using a register array similar to that in [7], which is developed to be appropriate for larger blocks in our proposed architecture, the reference pixels needed for prediction are selected.

Stage2: Do the interpolation as described in equation (1) to obtain the predicted pixels. The pixel equality based computation reduction (PECR) technique [8] is used here.

In addition, Planar and DC can be integrated in the three-stage architecture easily, only adding a few multiplexers and computing units like adders and shifters. This pipelined design makes high frequency possible.

3.3 Control Mechanism

Fig. 6(a) illustrates an example of the postorder traversal that applied to an incomplete quad-tree structure blocks in the proposed architecture, while the HEVC Test Model HM encoder software uses preorder traversal, as shown in Fig. 6(b). The numbers in the block indicate the scan order of all the blocks with different sizes.

Preorder traversal is efficient for the recursion in software but not suitable for hardware design, because it generates much intermediate data so that a lot of internal buffers are needed. For example, all the information of

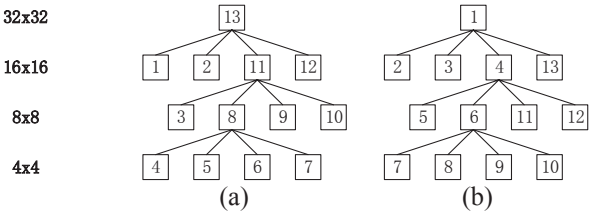

Figure 6. Scan order of quad-tree structure

the 32x32 block (1) (such as mode, distortion, predicted pixels, etc) must be remained until the process of all the 16x16, 8x8, 4x4 blocks (2 to 13) has been finished. This problem also exists in the 16x16 and 8x8 blocks.

However, in the postorder traversal, the work starts with the leaf nodes in the lowest layer. Whenever four leaf nodes are finished, their root node is processed. Then the better partition is chosen, the root or the leaves, and the worse is abandoned. In this way, too many internal buffers are reduced.

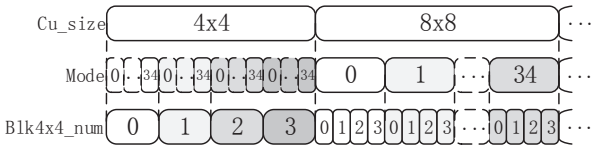

Figure 7. Partial Timing Diagram of Control Signals

Besides, because the process unit in this proposed architecture is 4x4 block, when scanning the full-mode prediction in a larger block, one mode is applied to every sub-block (4x4) first, and then turn to the next mode as shown in Fig. 7. Thus, when one mode is finished, the distortion value is replaced by the new or remains the previous. It reduces the buffers for storing the distortion of all modes and makes it convenient for mode decision.

In the proposed architecture, using this control mechanism shown in Fig. 7, it takes only 8967 cycles to complete the intra prediction of a whole 32x32 treeblock, scanning all modes and all block sizes.

4. Results

The prediction engine A in this architecture, which completes the full-mode and full-size predictions scan, is synthesized with TSMC 65nm CMOS technology. Table 2 presents the results.

Table 2. Synthesis Results

Work	This work		[5]	[7]
Technology	TSMC 65nm		IBM 65nm	TSMC 130nm
Gate Count	ref-generate	33K	37K	9K
	predict	31K		
	mode&split	13K		
	total	77K		
Frequency	600MHz		500MHz	150MHz
Cycles/32x32	8967		18130	--
Block Size	All		All	4x4

The reason why the logic gate count is more in this work is that the proposed architecture is 16-pixel parallel and includes the processes of the reference samples such as substitution and filtering. Furthermore, hadamard transform is used to calculate the distortion, instead of just calculating the SAD like in [5].

5. Conclusion

This paper proposes a highly pipelined VLSI architecture for intra prediction in HEVC encoder, supporting all modes and all block sizes. Data dependency is reduced by using two prediction engines: one based on original samples, and the other referring to the reconstructed. A universal predictor, 3-stage fully pipelined, is presented in this design. Under an applicable control mechanism, the internal buffers are decreased and the intra prediction of a whole 32x32 treeblock can be completed within 8967 cycles. Considering the frequency up to 600MHz, this designed architecture supports the real-time encoding of 1080P@30fps video sequence.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (61306023),Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP, 20120071120021),National High Technology Research and Development Program (863, 2012AA012001), State Key Lab of ASIC & System Project (11MS004).

References

[1] B.Bross, W.-J. Han, J.-R. Ohm, G.j.Sullivan, Y.-K. Wang, T.Wiegand, "High Efficiency Video Coding (HEVC) text specification draft 10 ", JCTVC-L1003, Jan.2013.

[2] Draft ITU-T recommendation and final draft international standard of joint video specification, ITU-T Rec. H.264/AVC/ISO/IEC 14496-10 AVC, JCT-G050, 2003.

[3] S.Smaoui, H.Loukil, A.Ben Atitallah, N.Masmoudi, "An efficient pipeline execution of H.264/AVC intra 4×4 frame design," Systems Signals and Devices (SSD), June 2010.

[4] Huailu Ren, Yibo Fan, Xinhua Chen, Xianyang Zeng, "A 16-pixel Parallel Architecture with Block-level/ Mode-level Co-reordering Approach for Intra Prediction in 4kx2k H.264/AVC Video Encoder", Design Automation Conference (ASP-DAC), 2012.

[5] D.Palomino, F.Sampaio, L.Agostini, S.Bampi, A.Susin, "A memory aware and multiplierless VLSI architecture for the complete Intra Prediction of the HEVC emerging standard", Image Processing (ICIP), 2012.

[6] J.Lainema, F.Bossen, Woo-Jin Han, Junghye Min, K.Ugur, "Intra Coding of the HEVC Standard", Circuits and Systems for Video Technology, 2012.

[7] Fu Li, Guangming Shi, Feng Wu, "An Efficient VLSI Architecture for 4x4 Intra Prediction in The High Efficiency Video Coding (HEVC) Standard", Image Processing (ICIP), 2011.

[8] Ercan Kalali, Yusuf Adibelli, Ilker Hamzaoglu, "A High Performance and Low Energy Intra Prediction Hardware for High Efficiency Video Coding", Field Programmable Logic and Applications (FPL), 2012.