Hardware-Software Collaborative Complexity Reduction Scheme for the Emerging HEVC Intra Encoder

Muhammad Usman Karim Khan, Muhammad Shafique, Mateus Grellert, Jörg Henkel
Chair for Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Germany
{muhammad.khan, muhammad.shafique, henkel}@kit.edu

***Abstract*—High Efficiency Video Coding (HEVC/H.265) is an emerging standard for video compression that provides almost double compression efficiency at the cost of major computational complexity increase as compared to current industry-standard Advanced Video Coding (AVC/H.264). This work proposes a collaborative hardware and software scheme for complexity reduction in an HEVC Intra encoding system, with run-time adaptivity. Our scheme leverages video content properties which drive the complexity management layer (software) to generate a highly probable coding configuration. The intra prediction size and direction are estimated for the prediction unit which provides reduced computational-complexity. At the hardware layer, specialized coprocessors with enhanced reusability are employed as accelerators. Additionally, depending upon the video properties, the software layer administers the energy management of the hardware coprocessors. Experimental results show that a complexity reduction of up to 60 % and the energy reduction up to 42 % are achieved.**

I. Introduction and Motivation

Digital video compression is a fundamental requisite of many day-to-day applications, like video conferencing, security and entertainment. Due to the ever increasing trend of video resolutions (from Full HD 1920×1080 to Quad Full HD 4096×2048 and Ultra HD 7680×4320) and frame rates (30 FPS to 60/120 FPS), the Joint Collaborative Team on Video Coding (JCT-VC) have recently developed the next generation video coding standard, called the High Efficiency Video Coding (HEVC, also termed as H.265) [1]. The goal of HEVC is to increase the compression efficiency by 50% as compared to that of the H.264. This coding efficiency is achieved by introducing additional coding tools and accompanies a tremendous increase in the computational complexity.

Unlike the H.264's concept of a Macroblock (MB, 16×16 region of the video frame used as a primary compression unit), HEVC implements a Quad Tree Coding structure (see Fig. 1), called the Coding Tree Blocks (CTB). The concept of MBs is replaced by the Largest Coding Unit (LCU) which can be recursively divided into 4 Coding Units (CU) of size 2N×2N. The LCU is subdivided into every possible block partition size (CU size) and the best combination of CU sizes is selected, by comparing the Rate-Distortion (RD) cost of one combination to others (the process is termed as RD Optimization (RDO)). A CU can be further subdivided into Prediction Units (PU) (of size 2N×2N or N×N) and Transform Units (TU).

Intra-video encoders exploit redundancies of video sequence only in the spatial domain. These encoders are well-suited to low latency applications like automotive, and high quality archiving solutions to remove motion artifacts. For HEVC Intra-encoding, a PU defines the basic entity for intra prediction, confining itself to the available many angular directions, DC and planar modes [1]. The PU partition for a CU and the best prediction mode are collectively called the *coding configuration* of the CU.

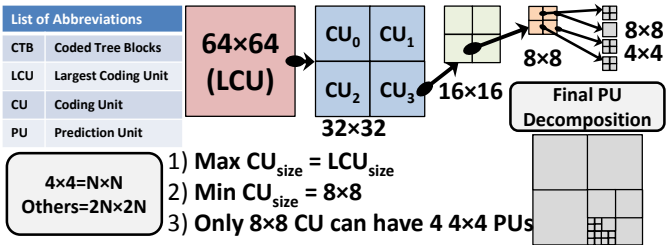

Fig. 1: One of the possible CU decomposition in HEVC where a CU is recursively converted into sub-CUs and PUs

Analysis and Problem: This enormous decision space for selecting a RDO coding is required for increased compression efficiency. However, the iterative and recursive behavior of RDO optimization incurs significant complexity overhead, even for intra-only encoders, because the RDO decision has to recursively check each possible PU and intra mode combination. It is noteworthy that the total number of mode combinations in HEVC is ~42.4× more than that in H.264.

Our experiments in Fig. 2 show that the computational complexity of the complete Intra-only HEVC has increased by a factor of ~1.4× for a compression efficiency increase of around 35% as compared to Intra-only H.264. A similar analysis can be found in [6]. Note, for a Full HD (1920×1080) video, it took approximately 83 seconds to encode one intra-frame on an Intel Core-2-Duo processor with 4 GB RAM which illustrates a significant challenge towards fast HEVC encoders. Therefore, it is vital to develop complexity reduction algorithms to realize real-world applications based on the HEVC intra encoders.

The coding complexity illustrates that hardware solutions are required in embedded video coding systems to fulfill the real-time encoding demands for HEVC. But a hardware-only solution of HEVC will have long time-to-market due to the time consuming full custom design cycle. The development of software-only solution for HEVC encoding is fast and flexible, but its throughput is low. Recently, a number of state-of-the-art HEVC intra encoders have been proposed, e.g. [7]. In [3], the authors proposed an HEVC Intra prediction HW for only 4×4 blocks. The work in [4] presents a gradient based fast intra mode decision for a given PU size. In [5], authors have also presented a fast partition size selection algorithm for inter-frames exploiting temporal correlations for frame compression. These methods try to alleviate pressure off the encoding modules by performing sub-optimal encoding and using hardware-only solutions, thus limiting the flexibility of the architectures and resulting in larger energy, area and memory overhead.

Our Novel Contributions: To satisfy the real-time throughput constraints of HEVC intra-encoding and to reduce energy

978-3-9815370-0-0/DATE13/©2013 EDAA

consumption while achieving enhanced adaptivity, we propose a system-level scheme with collaborative hardware-software modules. Our scheme leverages the advantages of high flexibility of software along with the high computational throughput and reusability of hardware coprocessors, while exploiting the video properties. The novel features for our scheme are:

- **Adaptive Complexity Management** using hardware accelerators, controlled by programmable modules and video properties. The programming modules specify *mode exclusion* and *adaptive hardware-energy control.* The hardware architecture is composed of *video feature extractors* and *reusable prediction accelerators.*
- **Content Driven Encoding adaptation** with respect to video properties by exercising only the high probability partition size and constraining the search for the best prediction mode.

Up to the best of authors' knowledge, this is the first work for collaborative hardware-software optimizations for intra-only HEVC encoding. Though, there are previous works that target joint hardware and software optimizations for H.264 [9][10] and Multi-View Coding (MVC)[11], these scheme for H.264 and MVC cannot be readily reproduced for HEVC, due to the novel CTB model and a different set of coding tools.

Paper Organization: In Section II, we present our analysis of HEVC coding and prediction units along with their relationship to the video properties. Using this relationship, we develop our system in Section III and discuss various aspects of the software modules and hardware architecture. In Section IV, we present the results of our system and conclude the paper in Section V.

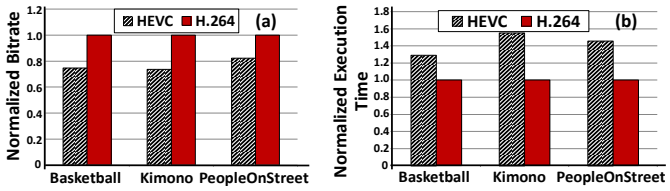

Fig. 2: Comparing H.264-Intra to HEVC-Intra (a) Bitrate (b) Execution Time

II. Our Analysis of HEVC/H.265

For this analysis, HEVC reference software is extended and various video sequences are tested. Fig. 3 shows the PU size encoding statistics of two test sequences[1] with diverse video content, (see experimental setup in Section IV) for quantizer settings QP=22 (high quality) and QP=37 (low quality).

Observation-I: The distribution of PU sizes for different videos with diverse characteristics is not uniform and the selection of PU size depends upon the properties of the video.

Exploitation: Therefore, complexity reduction algorithms need to exploit this statistical information for high performance.

In Fig. 4, a color-coded intra angular direction map is overlaid on top of the frame. This color map is shown in block D of the figure, with each intra prediction octant (each octant has 8 angular directions) having an associated color.

Observation-II: Fig. 4 shows that the intra angular direction depends upon the gradient direction[2] (blocks A, B and C). The gradient direction is perpendicular to one of the 4 octants of intra prediction direction. With high probability, the intra prediction mode lies in that octant.

Exploitation: Depending upon the gradient of the video frame block, a set of highly probable intra-prediction modes can be tested, excluding the unlikely modes. This provides a great potential for complexity reduction.

Summarizing the Analysis: We design our system to incorporate the findings of the above-mentioned analysis. The key challenges for fast energy-efficient HEVC intra encoding are:

- *Complexity reduction* by reducing RDO decision tree
- *Energy savings* by utilizing the statistical information of the video and utilizing the HW only when required
- *Minimum quality degradation* by selecting the most probable partition size and prediction mode selection
- *Leveraging video properties* (variance, gradients) for reduced encoding complexity.

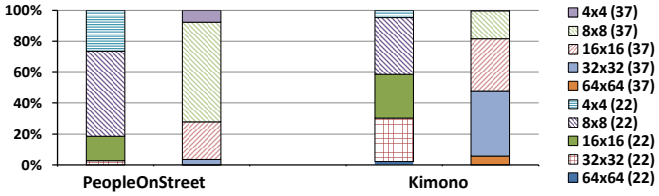

Fig. 3: PU size occurrences for HEVC Intra-Only HEVC

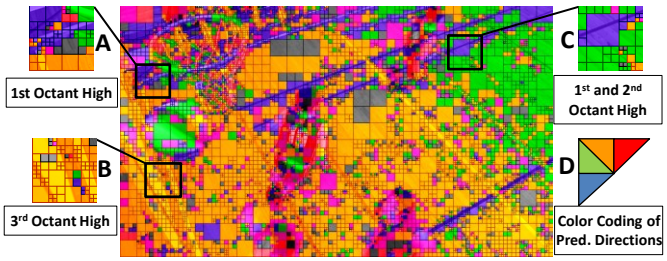

Fig. 4: Color-coded intra angular directions per PU on 2nd frame of "BasketBallDrill" with max CU size of 64×64

III. Our Fast Intra Encoding System

Fig. 5 shows the block diagram of our system. A structure comprising of hardware accelerators and programmable modules with their connections is shown. We can see that hardware modules act as special accelerators for the algorithms. The hardware modules are reusable, meaning that a single hardware module can be used to compute different intra predictions (depending upon the PU size). Our scheme works in the following three major steps, which are explained in the subsequent sections.

1. The best PU sizes in the LCU are estimated for the full LCU using the variance of the LCU, generating a PU map.
2. The best intra prediction modes are estimated by computing the gradients of the PU pixels.
3. Depending upon the PU size and its location in the LCU, prediction hardware modules are turned ON/OFF using clock gating and prediction is generated, 1 PU row at a time.

A. Complexity Reduction Scheme

This scheme performs the selection of the best PU/CU sizes, early mode elimination, fast intra mode selection, hardware configuration and clock gating of hardware modules to save

[1] The recommended test-sequences by JCT-VC were used for this analysis.
[2] Computed using the technique of [2][4].

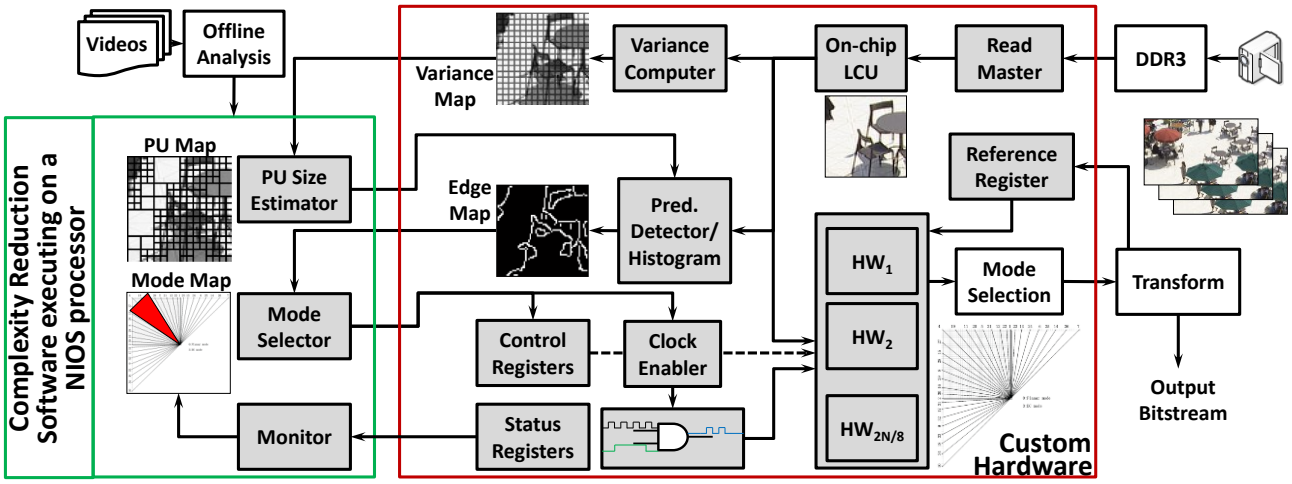

Fig. 5 Our HW-SW collaborative system overview diagram with our contributions marked as shaded blocks

energy. Main features of the programmable intra encoding complexity reduction scheme are discussed below.

1. **Early PU size estimation:** Fig. 6 illustrates that PU partition size selection problem can be reduced to a set of most probable partitions, depending upon the texture of the video frames. A large block with high variance must be sub-divided to decrease the individual block variances and smaller partitions with low variances can be combined to form larger partitions. We note that highly-detailed regions (blocks B and C having high spatial activity) are encoded using smaller PU sizes and less-detailed regions using larger block sizes (block A).

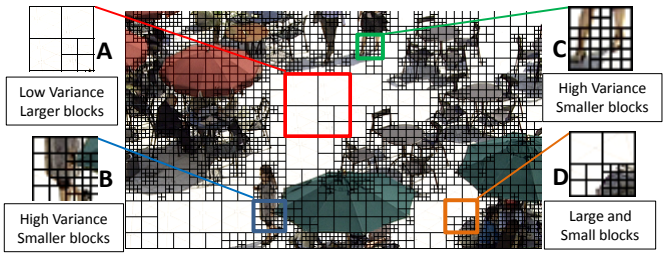

Fig. 6: PU borders on the 31st frame of BQSquare with max CU of 64×64

Fast_Intra_Mode_Sel ( ): ***Input****:* PU data *PU*, PU size PU_{size};
Output: Intra Modes array *IMA*;

1. $\forall_{\text{Pixel i}\in\text{PU}}$, **do** {
1. $G_x, G_y \leftarrow \text{Edge}(PU_i)$;
2. $\theta \leftarrow \text{ClosestIntraAngle}(G_x, G_y)$;
3. $\text{HashTable}(\theta) \leftarrow \text{HashTable}(\theta) + \|(G_x, G_y)\|$; }
4. $\text{IMA} \leftarrow \text{DesendingSort}(\text{HashTable})$;
5. **return** (IMA);

Fig. 7: Fast Intra mode selection algorithm

2. **Early mode exclusion and Fast mode selection:** Once a suitable PU size is selected, we can eliminate less-probable intra prediction angular modes. Selection of the most probable modes is mainly based upon the gradient based scheme and it is inspired from [2][4]. Based upon the gradient value in *x* and *y* directions, a histogram is generated and the most probable modes are executed. This algorithm is explained in Fig. 7. For each PU pixel, gradients and their angles are computed (lines 1-2) and they are used to fill a Hash table (serving as a histogram, line 3), which is then sorted to generated the best prediction modes, stored in Intra Mode Array (IMA, line 4).

3. **Clock Manager:** The hardware control and status registers are used to configure the distributed intra prediction blocks, according to the size of the PU. Additionally the clock manager reads the status register and gates the clock to the available HW modules.

B. Architecture for HEVC Intra Prediction

The programmable and hardware modules closely interact with each other. Hardware modules are mainly used as accelerators where as the programmable modules are used for configuration and complexity controlling. A read-master is responsible for fetching a full LCU from the memory. This read-master starts prefetching the LCUs once the programmable modules are ready and DDR3 is calibrated. We discuss main blocks of the architecture briefly in the following.

1. **Variance Generator:** Our hardware scheme generates the variance of a full LCU and stores them in a RAM, accessible to the complexity manager. The programmable modules query this RAM and generate the PU map.

2. **Edge Detector and Histogram Parser:** After the PU map generation, each PU is predicted one at a time. The location and width of a single PU from PU map is provided to the Edge detector [2], which computes a histogram of the intra prediction modes to predict highly probable intra directional modes.

3. **Intra Prediction Blocks:** Once the best modes are selected, intra prediction of the PU is carried out. Basically, one row at a time of the prediction is generated and it is used to compute the residue. For this purpose, we have designed a distributed prediction unit. The prediction unit consists of smaller blocks and each intra prediction block is capable of producing 8 pixels-per-cycle prediction. As seen from Fig. 8, the full row of the LCU (and the reference samples) is distributed to the HW blocks, specifically, 8 pixels per intra prediction block. The value 8 is chosen because the analysis in Section II suggests that 8×8 mode of a PU is highly likely. Note, a PU cannot not at the 8×8 boundary, because the minimum CU size in HEVC is 8×8.

4. **Clock Generator:** Eight pixels of a full LCU row are associated with each intra prediction block. When a PU is processed, it is possible that some of the intra prediction blocks

are not required. Therefore, the SW clock gates them to save energy. The gating circuit is controlled by the control register set by the programming module, depending upon the PU size and its location in the LCU.

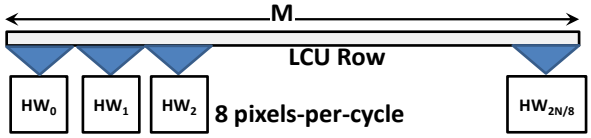

Fig. 8: Intra prediction HW blocks associated with an LCU row

IV. Experimental Results and Discussion

A. Experimental Setup and Prototype

The shaded blocks in Fig. 5 are developed using C++ (programmable modules for complexity reduction) and VHDL (for accelerators and custom logic) and prototyped on the Altera EP2AGX260FF3513 FPGA. NIOS II embedded processor is programmed to perform the software tasks and accelerators on the FPGA are connected to the NIOS processor using *Custom Instruction* (CI) interface [8].

B. Results

Fig. 9 shows the quality comparison using Rate-Distortion (RD) curves of our scheme as compared to the full-RDO decision, implemented on the HM-7.2 reference software for HEVC encoding and H.264. These curves indicate a small quality loss of the proposed scheme against HEVC, however, still superior to the H.264.

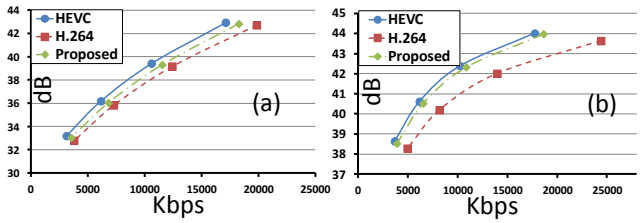

Fig. 9: Rate-distortion curves for HEVC, H.264 and proposed scheme for sequences: (a) "RachHorses" (b) "Kimono"

Table I presents the area and frequency results of individual hardware modules, along with their maximum frequency. EP2AGX260FF3513 is a mid-range FPGA, and hence, more improvement in the throughput and area savings is possible by using a complete custom design.

TABLE I: TOTAL HARDWARE UTILIZATION FOR LCU SIZE OF 64×64 AND PIXEL OF 8-BITS (1 PLL, ~205K ALUTS, ~205K REGISTERS, 736 DSP BLOCKS)

Hardware block	Frequency (MHz)	ALUTs; Registers	Memory (bits)	DSP blocks	Logic
Variance	167.14	253; 386	0	7 (<1%)	< 1 %
Sobel Histogram	243.72	77; 86	0	1 (< 1%)	< 1 %
Intra Pred. Blocks	243.61	203; 377	0	8 (1%)	< 1%
Total	162.79	13409; 6934	43008	72 (10%)	10 %

In Fig. 10, the energy consumption comparison between (a) no clock gating with single prediction hardware; and (b) proposed scheme is performed for 1 frame (with percentage energy savings on top of the bars). The stimuli data generated by the HM-7.2 reference software and ModelSim is provided to the Altera's Powerplay Power Analyzer tool (for determining signals' static probabilities and transition densities) and the energy numbers are reported.

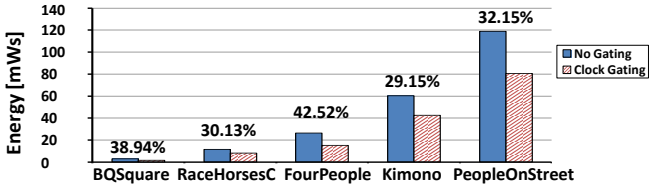

Fig. 10: Energy consumption for 1 frame with no clock gating and distributed intra prediction block against the proposed scheme

V. Conclusion

This paper presents a hardware-software collaborative scheme for HEVC intra-encoding engines to increase the throughput and energy saving. By exploiting video texture properties, our scheme limits the best partition search to a single PU-size and highly probable intra mode search is limited to a set of estimated modes. Moreover, texture properties are also leveraged to control the clocks of the hardware modules and to achieve high energy savings. Our scheme achieves 35 % energy reduction and 2.2x improving performance. Experimental results demonstrate promising impacts, denoting the proposed scheme as a powerful enabler for real-time multimedia platforms that will encode video streams via HEVC in the near future.

References

[1] B. Bross, W. J. Han, J. R. Ohm, G. J. Sullivan, T. Wiegand, "High Efficiency Video Coding (HEVC) text specification draft 7", 2012.

[2] M. Shafique, B. Molkenthin, J. Henkel, "An HVS-based Adaptive Computational Complexity Reduction Scheme for H.264/AVC video encoder using Prognostic Early Mode Exclusion", IEEE DATE, pp.1713-1718, 2010.

[3] F. Li, G. Shi, F. Wu, "An efficient VLSI architecture for 4×4 intra prediction in the High Efficiency Video Coding (HEVC) standard", IEEE ICIP, pp. 373-376, 2011.

[4] W. Jiang, H. Ma, Y. Chen, "Gradient based fast mode decision algorithm for intra prediction in HEVC", CECNet, pp. 1836–1840, 2012.

[5] M. B. Cassa, M. Naccari, F. Pereira, "Fast Rate Distortion Optimization for the Emerging HEVC Standard," in PCS, pp. 493–496, 2012

[6] T. Nguyen, D. Marpe, "Performance analysis of HEVC-based intra coding for still image compression", PCS, pp.233-236, 2012.

[7] G. V. Wallendael, S. V. Leuven, J. D. Cock, P. Lambert, R. V. de Walle, J. Barbarien, A. Munteanu, "Improved intra mode signaling for HEVC", IEEE ICME, 2011.

[8] Altera NIOS II Processor Custom Instruction, Online: http://www.altera.com/literature/ug/ug_nios2_custom_instruction.pdf

[9] M. Shafique, L. Bauer, J. Henkel, "Optimizing the H.264/AVC Video Encoder Application Structure for Reconfigurable and Application-Specific Platforms", JSPS, vol. 60, no. 2, pp. 183-210, 2010.

[10] T. Dias, N. Roma, L. Sousa, "Hardware/software co-design of H.264/AVC encoders for multi-core embedded systems", DASIP, pp.242-249, 2010.

[11] B. Zatt, M. Shafique, S. Bampi, J. Henkel, "Multi-level pipelined parallel hardware architecture for high throughput motion and disparity estimation in Multiview Video Coding", IEEE DATE, pp.1-6, 2011.

[12] M. Shafique et al., "Adaptive Power Management of On-Chip Video Memory for Multiview Video Coding", DAC, pp. 866-875, 2012.

[13] E. Kalali, Y. Adibelli, I. Hamzaoglu, "A high performance and low energy intra prediction hardware for High Efficiency Video Coding". FPL, pp. 719-722, 2012.