

9.5 A 249Mpixel/s HEVC Video-Decoder Chip for Quad Full HD Applications

Chao-Tsung Huang¹, Mehul Tikekar¹, Chiraag Juvekar¹, Vivienne Sze², Anantha Chandrakasan¹

¹Massachusetts Institute of Technology, Cambridge, MA, ²Texas Instruments, Dallas, TX

The latest video coding standard High Efficiency Video Coding (HEVC) [1] provides 50% improvement in coding efficiency compared to H.264/AVC, to meet the rising demand for video streaming, better video quality and higher resolutions. The coding gain is achieved using more complex tools such as larger and variable-size coding units (CU) in a hierarchical structure, larger transforms and longer interpolation filters. This paper presents the first integrated circuit which supports Quad Full HD (QFHD, 3840x2160) video decoding for HEVC draft standard. It addresses new design challenges for HEVC ("H.265") with three primary contributions: 1) a system pipelining scheme which adapts to the variable-size largest coding unit (LCU) and provides a two-stage sub-pipeline for memory optimization; 2) unified processing engines to address hierarchical coding structure and many prediction and transform block sizes in area-efficient ways; 3) a motion compensation (MC) cache which reduces DRAM bandwidth for the large LCU and meets the high throughput requirement due to long filters.

Figure 9.5.1 shows the system block diagram and pipelining scheme. In contrast to the fixed 16x16 macroblock (MB) in H.264/AVC, the LCU size in HEVC can be one of 64x64, 32x32 or 16x16 in a video sequence. Our proposed approach is to pipeline the system in variable-size pipeline blocks (VPB) which can be 64x64, 64x32 or 64x16 respectively. The VPB sizes are chosen to unify hardware processing flow and enable a memory-efficient sub-pipeline in the prediction engine. The VPB pipeline buffer size is proportional to 64x64 which is 16 times larger than a MB, which significantly increases memory requirements. The pipeline is broken into two groups as shown in Figure 9.5.1 for two improvements. The first is the reduction of 22KB SRAM for coefficients by replacing the VPB pipeline between entropy decoder and transform engine with a smaller 2KB transform unit (TU) FIFO. The second is accommodating the MC cache that has uncertain DRAM access latency to the VPB pipeline. The MC cache receives access requests from the dispatch engine, and its outputs are stored in the reference pixel buffer for the prediction engine. For saving DRAM bandwidth, two line buffers are also used to store top row pixels and coding information.

Figure 9.5.2 shows the unified prediction engine. An LCU has a hierarchical coding structure in the form of a CU tree with mixed intra and inter CUs. To support all intra/inter combinations using the same pipeline, we unify intra and inter prediction into one control flow. Their throughputs are aligned (four samples in one cycle) such that they can share the same reconstruction core of adding residues and updating pixels for intra prediction. A two-stage sub-pipelining scheme is devised for the control flow to optimize memory usage. One 64x64 VPB is partitioned into four 32x32 prediction pipeline blocks (PPB) to reduce the subsampled LMChroma buffer from 32x32 to 16x16 pixels. If the MC cache outputs and the prediction engine were pipelined in PPB, the reference pixel buffer size would be 44KB. Instead, one PPB is further split into six sub-PPBs and only 8KB is required. The proposed pipeline also saves 9% DRAM bandwidth for LCU 16x16 compared to a direct LCU pipeline. It only switches Y and U/V once in a 64x16 block instead of four times and every switch may cause a DRAM row change which increases DRAM cycles.

The HEVC draft standard uses 8-bit integer transform for 4 to 32-pt Inverse DCT (IDCT) and 4-pt Inverse DST (IDST). The largest TU is 32x32 with seven smaller TUs – three square (4x4, 8x8, 16x16) and four non-square (4x16, 8x32, 16x4, 32x8). Compared to H.264/AVC, this is an 8x increase in 1-D transform logic and a 16x increase in transpose memory for the largest TU. To reduce logic area, only four row or column pixels in a TU are transformed every cycle by the variable-size partial 1-D transform block shown in Figure 9.5.3. We extend the typical recursive decomposition of a 2N-pt IDCT into N-pt IDCT and NxN matrix multiplication to our partial architecture with even-odd index sorting. The NxN matrix has only N unique elements differing only in sign. The 4x4 matrix, for example, has only four elements – 89, 75, 50, 18. We exploit this

property to use Multiple Constant Multiplication (MCM) [6] instead of N full multipliers. This reduces the variable-size partial transform area from 96K to 71K gates. Two MCM planes are needed for partial 4x4, 8x8 and 16x16 matrix multiplications to meet the required throughput. The transpose memory would need a 125K gate register array. Instead, it is implemented using four single-port SRAMs for 4 pixel/cycle read or write that is fully pipelined with the transform computation. Some pixels are stored in a register-based row cache to avoid a stall when switching from column to row transform.

Figure 9.5.4 shows the proposed MC cache and DRAM mapping for reference frames. HEVC uses an 8-tap interpolation filter compared to the 6-tap H.264/AVC filter. For small prediction blocks, this dramatically increases the required bandwidth. For example, 4x4 blocks require 11x11 reference regions, a 49% increase over 9x9 regions for H.264/AVC. For large LCUs (up to 64x64), achieving a high hit rate is a challenge. We propose a four-parallel four-way set associative 16KB cache to meet the requirements placed by both small prediction blocks and large LCUs. The cache line size is 32-byte and corresponds to an 8x4 block in DRAM. Bi-prediction for 4x4 blocks may require up to 8 rows of these cache lines within 8 cycles which is achieved by the four-parallel architecture. Also, this cache can store up to four 64x64 blocks and maintain an average hit rate of 61% despite the large LCU sizes. In addition, we propose a twisted 2-D tiling in the DRAM to increase horizontal separation between two DRAM rows for a given bank. This reduces the probability of changing rows in a bank to save bandwidth wasted on precharge/activate (ACT) cycles. Compared to a raster scan of 8x4 cache lines, 70% DRAM ACT bandwidth is saved. When benchmarked against sharing fetched pixels in only one sub-PPB, our cache design saves 67% DRAM bandwidth for MC.

The chip specifications are summarized in Figure 9.5.5. The core size is 1.77mm² in 40nm CMOS, which consists of 715K logic gates and 124KB on-chip SRAM. It is compliant to HEVC Test Model (HM) 4.0, and the supported decoding tools in HEVC Working Draft (WD) 4 [1] are also listed. This chip achieves 249Mpixels/s decoding throughput for QFHD videos at 200MHz with the target DDR3 SDRAM operating at 400MHz. The core power is measured for six different configurations as shown in Figure 9.5.5. The average core power consumption for QFHD decoding at 30fps is 76mW at 0.9V. The chip micrograph is shown in Figure 9.5.7.

Figure 9.5.6 shows the comparison with state-of-the-art video decoders [2-4]. This work supports HEVC draft standard which is the successor of H.264/AVC and has higher complexity and larger CUs for 2x compression capability. More SRAM is used than [2-4] because the pipeline buffers and cache SRAM are much larger for LCU 64x64 and [3-4] do not have line buffers. The DRAM power is 219mW (modeled by [5]) for QFHD decoding at 30fps, for which the proposed MC cache saves 122mW. Despite the increased complexity, this work demonstrates the lowest normalized system power, which facilitates the use of HEVC on low-power portable devices for QFHD applications.

Acknowledgements:

This work was funded by Texas Instruments. The authors thank the TSMC University Shuttle Program for chip fabrication.

References:

- [1] B. Bross, et al., "WD4: Working Draft of High-Efficiency Video Coding," ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11 Doc. JCTVC-F803, July 2011.
- [2] D. Zhou, et al., "A 2Gpixel/s H.264/AVC HP/MVC Video Decoder Chip for Super Hi-Vision and 3DTV/FTV Applications," *ISSCC Dig. Tech. Papers*, pp. 224-225, Feb. 2012.
- [3] T.-D. Chuang, et al., "A 59.5mW Scalable/Multi-View Video Decoder Chip for Quad/3D Full HDTV and Video Streaming Applications," *ISSCC Dig. Tech. Papers*, pp. 330-331, Feb. 2010.
- [4] C. C. Lin, et al., "A 160kGate 4.5kB SRAM H.264 Video Decoder for HDTV Applications," *ISSCC Dig. Tech. Papers*, pp. 1596-1605, Feb. 2006.
- [5] <http://www.micron.com/support/dram/power-calc>. DDR3 SDRAM System-Power Calculator.
- [6] M. Potkonjak, et al., "Multiple Constant Multiplications: Efficient and Versatile Framework and Algorithms for Exploring Common Subexpression Elimination," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 15, no. 2, pp. 151-165, Feb. 1996.

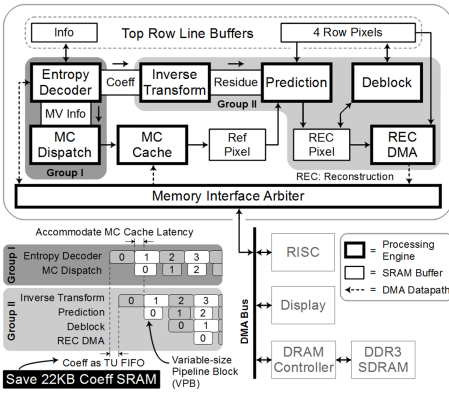


Figure 9.5.1: System block diagram and pipelining scheme in variable-size pipeline block (VPB).

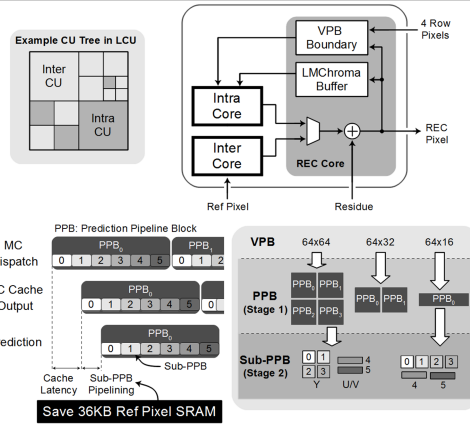


Figure 9.5.2: Unified prediction engine and memory-efficient two-stage sub-pipeline in PPB and sub-PPB.

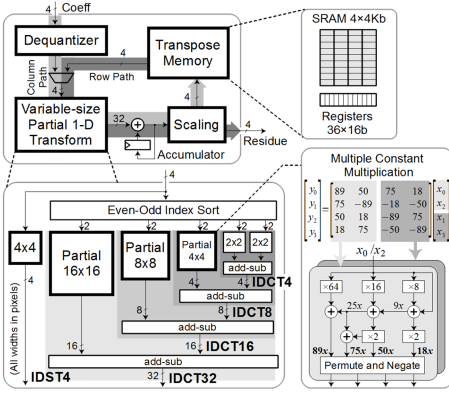


Figure 9.5.3: Unified 2-D inverse transform engine with variable-size partial transform for all square/non-square TUs with 4 to 32-pt IDCT and 4-pt IDST.

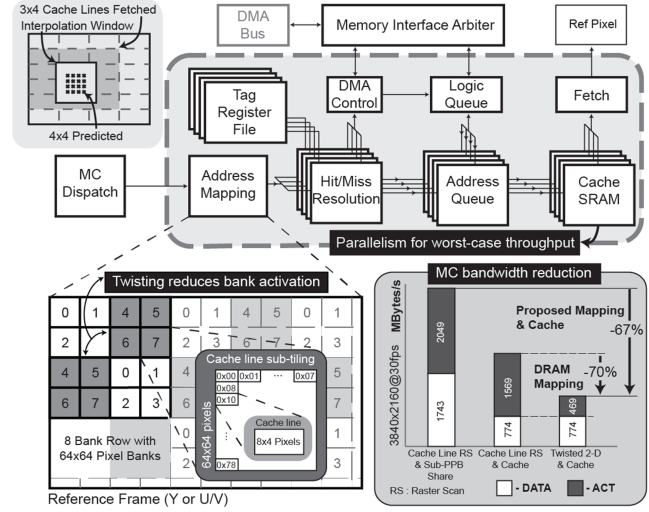


Figure 9.5.4: Four-parallel high throughput MC cache and twisted 2-D DRAM mapping.

Technology	TSMC 40nm CMOS
Supply Voltage	Core 0.9V, I/O 2.5V
Chip Size	2.18x2.18mm ²
Core Size	1.33x1.33mm ²
Gate Count	715K (2-input NAND)
On-Chip SRAM	124KB
Maximum Throughput	249Mpixels/s @ 200MHz
Decoding Tools	HEVC ("H.265") WD4 (HM4.0 low complexity w/o SAO) LCU size: 64x64/32x32/16x16 B-frame: Low delay (LD)/Random access (RA) All Motion Partitions (from 4x4 to 64x64, symmetric/asymmetric) All Transform Types (from 4x4 to 32x32, square/non-square) All Intra Modes (DC, Planar, 33 Angular, and LMChroma)
Measured Core Power Consumption*	76mW @ 200MHz, 3840x2160 @ 30fps (average) 51mW @ 100MHz, 1920x1080 @ 60fps (average) 31mW @ 25MHz, 1280x 720 @ 30fps (average)

* Measured at room temperature

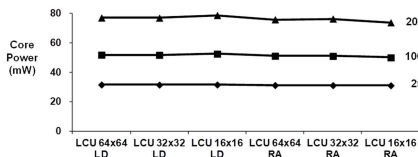


Figure 9.5.5: Chip specifications and measurement results.

	This Work	ISSCC'12 [2]	ISSCC'10 [3]	ISSCC'06 [4]
Standard	HEVC ("H.265") WD4	H.264/AVC HP/MVC	H.264/AVC HP/SVC/MVC	H.264/AVC MP
Max Specification	3840x2160 @30fps	7680x4320 @60fps	4096x2160 @24fps	1920x1080 @30fps
Gate Count	715K	1338K	414K	160K
On-Chip SRAM	124KB	80KB	9KB	5KB
Technology	40nm/0.9V	65nm/1.2V	90nm/1.0V	0.18µm/1.8V
Core Power*	76mW	410mW	60mW	320mW
DRAM Power*	219mW**	2.52W	NA	NA
Normalized Core Power	0.31nJ/pixel	0.21nJ/pixel	0.28nJ/pixel	5.11nJ/pixel
Normalized DRAM Power	0.88nJ/pixel	1.27nJ/pixel	NA	NA
Normalized System Power***	1.19nJ/pixel	1.48nJ/pixel	NA	NA
DRAM Configuration	32b DDR3	64b DDR2	NA	32b DDR + 32b SDR
LCU Size	64x64/32x32/16x16		16x16	
CU Size	64x64/32x32/16x16/8x8		16x16	
TU Size	32x32/16x16/8x8/4x4 32x8/8x32/16x4/4x16	8x8/4x4		4x4
Intra Prediction Block Size	32x32/16x16/8x8/4x4	16x16/8x8/4x4		16x16/4x4
Intra Prediction Mode	36 modes		9 modes	
Inter Prediction Block Size	64x64 to 4x4		16x16 to 4x4	
Inter Interpolation Filter	Y: 8-tap; UV: 4-tap	Y: 6-tap (half) + bi-linear (quarter); UV: 2-tap		

* Power for max. specification

** Modeled by [5]

*** System Power = Core Power + DRAM Power

Figure 9.5.6: Comparison with state-of-the-art video decoders.