

A High-efficiency HEVC Entropy Decoding Hardware Architecture

Hao Sun¹, Li Zhou^{1*}, Hongji Xu¹, Tao Sun², Yang Wang¹

¹School of Information Science and Engineering, Shandong University, No.27, South Shanda Road, Jinan 250100, China

²Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, P.R. China

sunhao1990@gmail.com, *zhou_li@sdu.edu.cn, Corresponding Author

Abstract— High Efficiency Video Coding (HEVC) is becoming more and more important in current consumer application platforms. Compared with H.264 standard, it can reach up to 8192x4320 resolutions at 120fps. To accelerate HEVC decoding processing, this paper presents an efficient hardware entropy decoding architecture, Entropy decoding includes Colomb and CABAC decoding. Hierarchy ring buffer is designed to storage input bit-stream; Stack based Quad-Tree decoding structure is proposed for bit-stream parsing. Since CABAC decoder is a well-known bottleneck of the decoding performance, a 5-stage pipelined CABAC decoding engine is designed to accelerate the serial decoding progress. The proposed entropy decoding hardware architecture is verified on Altera Stratix IV FPGA running at a 200MHz clock frequency, using 8404 slices, memory size is 204KB.

Keywords— HEVC, entropy, CABAC, architecture.

I. INTRODUCTION

High Efficiency Video Coding (HEVC), is developed by the Joint Collaborative Team for Video Coding (JCT-VC) [1]. It achieves a 2x higher compression ratio than H.264/AVC, and state-of-the-art video coding with HEVC provides the best ratio between bitrate and visual quality [2]. HEVC supports high resolutions such as 8192x4320. Instead of macro-block in H.264, HEVC uses more efficient and complex quad-tree to split each frame. Complex Coding Tree Unit (CTU) is split into small CUs and TUs. In order to achieve high coding efficiency, Context Adaptive Binary Arithmetic Coding (CABAC) is used to encode the CUs and TUs.

From H.264/AVC, entropy coding part is known as a critical throughput bottleneck in codecs, especially in decoding processing, which is mainly caused by its grievous serial dependencies caused by the feedback loops [3, 4]. Some entropy decoders try to increase the decode rate by decoding multi-bin in one cycle [5, 6]. That is with double or triple increased hardware resources and decreases of processing clock frequency. A prediction-scheme-based CABAC decoder is also proposed to be use in multi-bin case [7], but due to its delay of extra critical path and prediction miss penalty, it has lower the syntax element throughput. A pipeline based CABAC decoding architecture is proposed in [8] to accelerate its decoding performance. Pipeline technology is an efficient method with higher clock frequency [9], but need to carefully

consider the pipeline stage department and components. Otherwise, it would limit entropy coding throughput.

In hardware architecture of HEVC decoder, when unpredictable bit-stream is to be decoded to RBSP bit-stream, a large on-chip buffer is needed. When decoding the syntax elements, entropy engine analyses the bit-stream bit by bit, all the bits in buffered bit-stream should be addressable, only ring buffer structure can support this feature. HEVC quad-tree structure is another complex coding part. When parsing it, some structure may be useful to multiplex the quad-tree parser and decrease the resource consumption.

This paper proposes an architecture for HEVC entropy decoding with the goal to parse the bit-stream and decode all syntax element and variables with efficient hardware architecture and high throughput. The main contribution of this paper includes:

- In order to optimize the memory utilization, hierarchy ring buffer is designed to increase the peak input bandwidth and improve the flexibility of bit-stream buffer.
- As for the throughput bottleneck in codec, a 5-stage pipeline CABAC decoding engine is to improve the throughput by increasing the clock frequency and bypassing the dependencies between bins.

The rest of this paper is organized as follows. Section II introduces the whole architecture. Some novel structures of the circuit design are proposed: hierarchy ring buffer structure, stack-based quad-tree parser, clock-gated state machine, and 5-stage pipelined CABAC decoder. Section III presents the performance of hierarchy ring buffer and the FPGA resource area comparison between modules. Section IV gives conclusions and future research target.

II. PROPOSED HEVC ENTROPY DECODING ARCHITECTURE

Figure 1 illustrates the block diagram of the proposed HEVC entropy decoder architecture.

HEVC bit-stream entropy decoder is composed by four key components: Hierarchy ring buffer (HRB), Parser FSM (PF), Entropy Decoding Engines (EDE) and Direct memory access (DMA). The HRB is designed to detect and remove the emulation prevention three byte from the SODB (String Of Data Bits) to get RBSP (Raw Byte Sequence Payload) for other parsers and engines. PF is entropy decoding control

module. It traverses whole bit-stream tree in pre-order and offering the stepped control signal to all the other modules, including loads bins from the HRB, and a complex FSM control logic for whole entropy decoding process. It is also the synchronizer of the whole decoder, including the prediction and de-block modules. All the engines in the EDE are separately driven by the FSM and the outputs of them will be delivered to the temp RAM. EDE module contains: Fixed Length Engine (FLE), Colomb Engine (GE) and CABAC engine (CE). All the engines can separately decode syntax element in specific coding scheme. The DMA module and its arbiter is designed to share the external memory to other prediction module of the whole decoder.

In order to increase performance and power, a hierarchy ring buffer architecture is proposed to increase bit-stream parsing in HRB module. An efficient low power architecture is proposed for the complex parser FSM, and a stack-based quad-tree is proposed to decrease the area of the decoder.

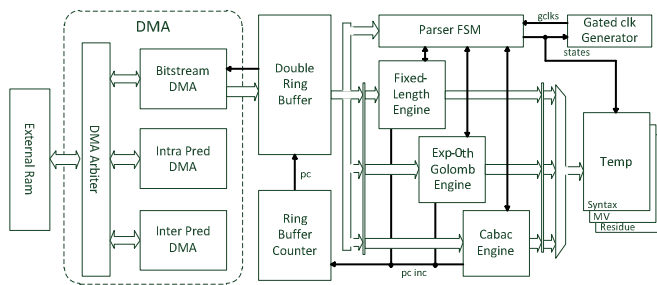


Figure 1. Overview of the architecture

A. Hierarchy Ring Buffer

Hierarchy Ring Buffer is proposed to achieve bit selection, so the data process logic can flexibly deal with the bit-stream. As shown in figure 2, the original bit-stream ring buffer is used to store SODB, the form of encapsulation of an SODB within an RBSP is mainly to prevent the emulation of start codes within NAL (Network Abstraction Layer) units. In SODB encapsulation, the emulation prevention three byte discarding logic is designed to detect start bits 0x000003 in SODB and remove 0x03, the resulting sequence of bytes is then prefixed with the NAL unit header.

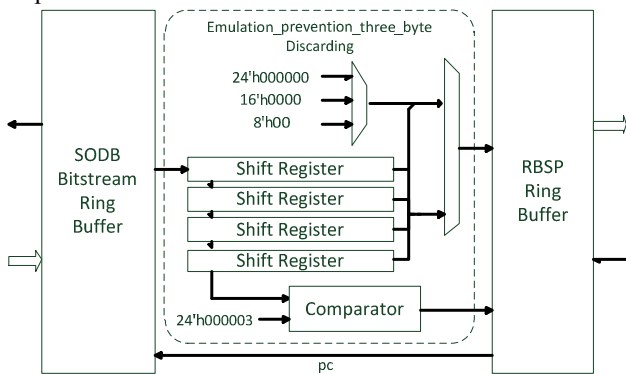


Figure 2. Hierarchy ring buffer

All the processed bit-stream is stored in RBSP hierarchy ring buffer, which can be directly used by other modules. When the parser parses the bit-stream, it does not need to care about the encapsulation, because this HRB can offer them the pure RBSP data.

B. Stack-Based Quad-tree Parser

The parser control module consists many sub state machine, all these state machine are organized as a complex FSM tree. NAL unit state machine is the root, when one state machine falls to a specific state, its child node state machine will start and the father one stops, and when the child one comes to the end, the father one continues. The FSM parsing progresses like pre-order traversing the state tree, and at some leaves, some state machine, for example, profile tier state machine is multiplexed.

But some state machines are multiplexed differently: they recursively call themselves. These kind of operation induce a problem that where to return. Because the state machine will call itself four times when the new decoded split flag is valid, and when it calls itself, the same multiplexed state machine will be restart with new inputs, but the old one should be kept in order to return to the same state when the recursive call finishes. Since its recursive call, a quad-tree is generated. A stack memory architecture is proposed to save the quad-tree split state.

A stack-based quad-tree architecture is proposed to solve the problem induced by the multiplexed state machine's recursive call, as shown in figure 3. When split is needed, current state will be pushed into the stack, and when recursive call is over, pop the state from the stack and to the next state.

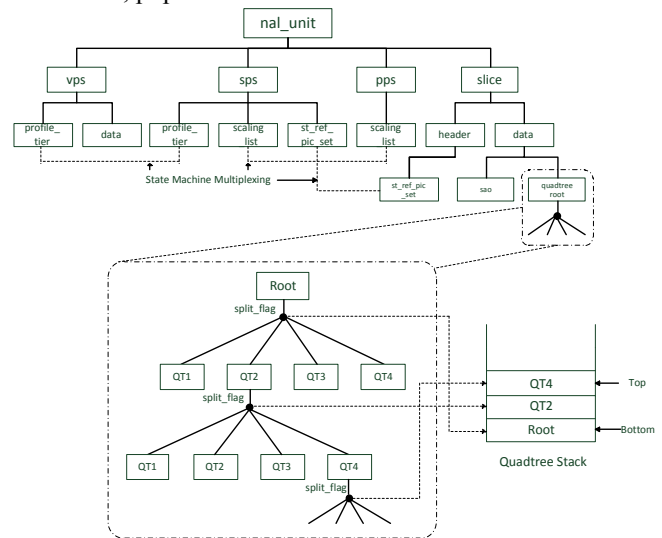


Figure 3. Stack-based quad-tree architecture

Figure 4 shows the parsing flow of the stack-based quad-tree. The quad-tree state machine's states are divided into two parts: split state module and decoding state module. The split flag's availability decides witch part to use. If the flag is valid, push the current node state to the stack and restart the state machine with left child node's parameter, then check its split flag; otherwise, start the decoding state machine and check

whether there's a brother node, then decide to pop to the father node state or restart with right brother node's parameter.

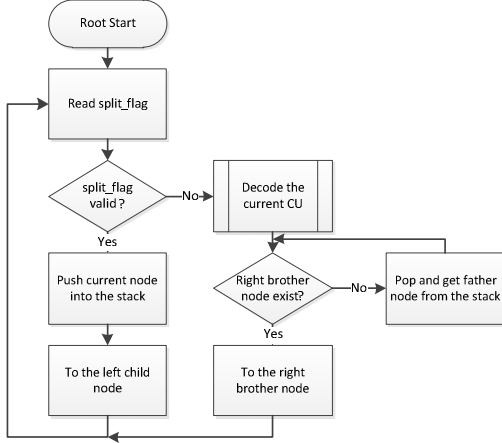


Figure 4. Stack-based quad-tree parser flow

C. Clock Gated FSM Controller

Since complex parser FSM need much processing power, which has to be considered during hardware design. An architecture of clock-gated state machine is proposed to decrease the power consumption.

The condition for state jumping is bring forward to the clock-gating enable logic, so when the jumping condition isn't met, the gated clock will be shut down, and all the state machines will stay at their states, so the dynamic power consumption will be much lower than the normal one. Figure 5 shows the difference between two architectures, the dashed box is the original architecture. Latch-and cell is the most widely used clock gating cell, it can resolve the normal problem in comb logic such as glitch and metastable clock signal.

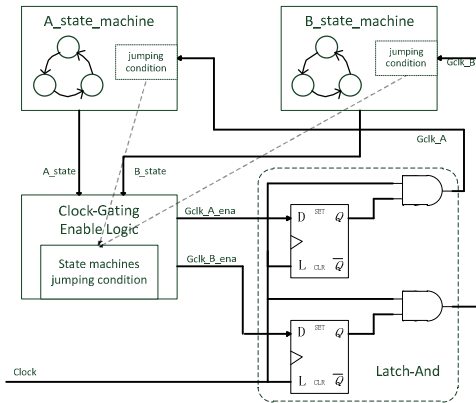


Figure 5. Latch-and-based state machine

Clock gating may influence the timing of the critical path, because all the gating combinational logic delay should be count into the clock tree delay and each path's timing check should consider about it. So the complexity of the enable signal should be decreased or the timing requirement is hard to meet.

D. 5-Stage Pipelined CABAC Decoding Engine

As the bottleneck of the whole decoder, CABAC decoding engine is the key engine of the EDE,. In order to increase the decoding speed, a 5-stage pipeline is designed in the engine, as shown in figure 6. Bit-stream comes into the arithmetic decoding engine and there are three sub decoding engine could be chosen according to the context: decode terminate engine, decode bypass engine and decode decision engine (DDE). With the required context values, the corresponding engine will be enabled to operating. Then the decoded bins are transmitting to the de-binarization module, the bins are de-binarized to the syntax element. One of context FSM and bypass FSM will decide which engine to use in next cycle according to the previously decoded syntax elements. If DDE engine is enabled, the context variable is required. At first, the context variables are initialized and stored into the Context Memory (CM), when one syntax element is decoded, the context variables will be updated. Context Selector (CS) will chooses the updated variables to write back to the CM for future access. Besides, CS and CM can only be controlled by Context FSM.

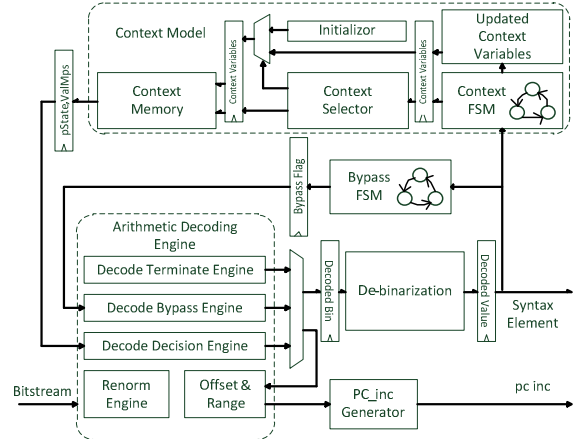


Figure 6. CABAC decoding engine architecture

Among the three sub decoding engine in arithmetic decoding engine, DDE contains the critical path and take up most of the decoding area. So, we use similar circuit of the arithmetic decoding architecture in [10] to optimize the DDE for an 11% reduction in delay.

III. EXPERIMENTAL RESULTS

In order to make sure that the HRB's input output bandwidth can well support the decoder, we make statistics of the random total number of input output bits in several clocks, the result is shown in TABLE 1. The parser and decoding engines' peak reading bits only happens in several clock, and because the ring buffer also buffers 512 bits data and start read in new bits when half is used, so, 44 bit/clock will never beyond the HRB's bandwidth. And the peak value in the average value is 4.67 bit/clock, which is much smaller than

the input rate 32 bit/clock. All that means the HRB is a bottleneck of the architecture.

We synthesis and implement the design of this architecture on FPGA Stratix IV EP4SE820F43C3, and TABLE 2 presents the resource usage of the design.

TABLE 1. TOTAL NUMBER OF INPUT OUTPUT BITS OF HRB

	bits	clk count	bits/clk
HRB output	44	1	44
	140	30	4.67
	239	68	3.51
	481	150	3.21
	512	325	1.57
	696	620	1.12
HRB input	32	1	32

TABLE 2. RESOURCE USAGE COMPARISON BETWEEN MODULES

	Resource usage		
	Combinational	Registers	Memory Bits
DMA	178	278	8192
HRB	159	1206	—
Parser	331	102	128
Clock	101	20	—
EDE	3243	1543	5840
Syntax Temp	1977	903	—
Variable Temp	1025	1287	977kb
MV Temp	1390	1134	640kb

IV. CONCLUSIONS

This paper proposes a high-efficiency HEVC entropy decoding architecture. Hierarchy ring buffer improves the peak band width of input path and the flexible of the architecture; Stack-based quad-tree decrease the resource consumed without sacrifice of parsing speed; The clock gated FSM control structure decrease the power consumption of the whole architecture; 5-stage pipelined CABAC engine architecture improves processing clock frequency and CABAC performance. This architecture is achieved on FPGA and run at a rate of 200MHz, using 8404 slices, memory size is 204KB.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of Shandong Province (ZR2013FQ006), Fund of Independent Innovation in Shandong Province (2013CXB3020), Foundation Grant of State Key Laboratory of ASIC & System (10KF011), State Key Laboratory of Digital Multimedia Technology (2013-1-2569), Shandong Post-Doctor Innovation Foundation Grant (201002029), Shandong University Innovation Foundation Grant (10000059614021), Grant of China Post-Doctor Innovation Foundation (20110491601) and University of Jinan Ph.D Foundation Grant (xbs1018). The

authors would like to thank all research partners for their significant contributions in this work. Thanks for the hardware platform supporting by University of Jinan.

REFERENCES

- [1] High efficiency video coding. ITU-T Recommendation H.265 and ISO/IEC 23008-2, April 2013.
- [2] J. Ohm, G.J. Sullivan, H. Schwarz, T.K. Tan, and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards—Including High Efficiency Video Coding (HEVC)," IEEE Transactions on Circuits and System for Video Technology, pp. 1669 – 1684, 2012.
- [3] Y.H. Chen, and V Sze, "A 2014 Mbin/s deeply pipelined CABAC decoder for HEVC," IEEE international conference on image processing (ICIP), pp. 2110 – 2114, 2014.
- [4] Y.H. Chen, and V Sze, "A Deeply Pipelined CABAC Decoder for HEVC Supporting Level 6.2 High-tier Applications," IEEE Transactions on Circuits and Systems for Video Technology, pp. 1. 2014.
- [5] J.W. Chen, and Y.L. Lin, "A high-performance hardwired CABAC decoder for ultra-high resolution video," IEEE Transactions on Consumer Electronics, pp. 1614 – 1622, 2009.
- [6] Y.J. Zhao, J.J. Zhou, and D.J. Zhou, "A 610 Mbin/s CABAC decoder for H.265/HEVC level 6.1 application," IEEE international conference on image processing (ICIP), pp. 1268 – 1272, 2014.
- [7] M.Y. Kuo, Y. Li, and C.Y. Lee, "An area-efficient high-accuracy prediction-based CABAC decoder architecture for H.264/AVC," IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1960 – 1963, 2011.
- [8] B.H. Bae, and J.H. Kong, "A design of pipelined-parallel CABAC decoder adaptive to HEVC syntax elements," The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014), pp. 1 - 2, 2014.
- [9] J. Hahlbeck, and B. Stabernack, "A 4k capable FPGA based high throughput binary arithmetic decoder for H.265/MPEG-HEVC," 2014 IEEE Fourth International Conference on Consumer Electronics (ICCE-Berlin), pp. 388 - 390, 2014.
- [10] V. Sze, and A.P. Chandrakasan, "A Highly Parallel and Scalable CABAC Decoder for Next Generation Video Coding," IEEE Journal of Solid-State Circuits, pp. 8 - 22, 2012.



Hao Sun, he was born on June 29, 1990. He received the BEng degree in Integrated Circuit Design and Integrated System from Shandong University, China, in 2013, and now he is pursuing the MEng degree in Engineering of Integrating Curcuit in Shandong University. His current research interests include VLSI design and HEVC codec, etc



Li Zhou, she received doctor degree on 2004 from Zhejiang University, China. On 2004, Dr.Zhou joined Freescale semiconductor R&D center as principle design architect till 2009. As an architect and core team member, Dr.Zhou participated in multiple National high-tech SoC and HDTV fundamental research projects, led multiple 65nm/90nm 10 million gate scale VLSI SoC chips, and joined many VLSI project and architecture researches. From 2009, Dr.Zhou is an assistant professor in Shandong University, China. Her current research interests include stereo vision system algorithm and hardware design, GPU architecture and VLSI design, high performance processor architecture and VLSI design, etc.



Hongji Xu, received the B.S. degree in electronic engineering from Shandong University of Technology, Jinan, China, in 1999, and received the M.S. degree in signal and information processing and the Ph.D. degree in communication and information system from Shandong University, Jinan, China, in 2001 and 2005, respectively. He is

currently an associate professor in the School of Information Science and Engineering at Shandong University. His research interests include wireless communications, ubiquitous computing, blind signal processing, and human-computer interaction.



Tao Sun, he received doctor degree on 2002 from Zhejiang University, China. Dr.Sun joined Suzhou China Core Co. Ltd as a researcher on CPU architecture and vice-general manager till 2007. From 2007 to 2009, Dr. Sun was a senior manager in Spansion China, led VLSI design projects. From 2009 till now, Dr.Sun is an associated professor in University of Jinan, China. His research interests include CPU/VLSI architecture, Solid storage architecture, etc.



Yang Wang, he was born on May 10, 1990. He received the BEng degree in Integrated Circuit Design and Integrated System from Shandong University, China, in 2013, and now he is pursuing the MEng degree in Circuit and System in Shandong University. His current research interests include 3D-GPU and Computer Architecture etc.