18.3 A 59.5mW Scalable/Multi-View Video Decoder Chip for Quad/3D Full HDTV and Video Streaming Applications

Tzu-Der Chuang, Pei-Kuei Tsung, Pin-Chih Lin, Lo-Mei Chang,
Tsung-Chuan Ma, Yi-Hau Chen, Liang-Gee Chen

National Taiwan Univeristy, Taipei, Taiwan

With advances in video coding technology, two main streams of multimedia applications are emerging. The first involves more vivid perceptual experience and is leading to the next generation of TV specifications – Quad Full HD (QFHD, 4096×2160p) and 3D/multi-view TV. The second is the scalable broadcasting and streaming of video.

To support these applications, the worldwide first video decoder that supports three most advanced video coding standards – H.264/AVC High Profile, multi-view video coding (MVC), and scalable video coding (SVC), is realized. Three main design challenges are encountered. 1) Conventional block-pipelining scheduling and architecture cannot efficiently support diverse SVC decoding schemes. 2) Up to 300Mbps high data-rate entropy decoding is required for QFHD, which is 2.3× that achieved in current state-of-the-art design [1]. 3) An external memory bandwidth (BW) of over 7.9GB/s is required for motion compensation (MC) in QFHD. Moreover, additional BW is required for inter-layer prediction (ILP) in SVC. High external memory BW results in high DRAM power consumption. Current state-of-the-art approaches [2-5] cannot resolve these issues effectively.

Three key design techniques of this chip are summarized as follows. 1) Frame-to-MB-level ILP scheduling optimization for SVC spatial scalability decoding, and a layer-interleaving decoding scheme for SVC quality scalability decoding. 2) Branch selection multi-symbol high-throughput context-based adaptive binary arithmetic decoder (CABAD) for high data-rate decoding. 3) Cache-based MC architecture and dedicated DRAM controller with DRAM data access optimization to reduce the external memory BW.

Figure 18.3.1 shows the system architecture and the proposed three-stage asynchronous macroblock (MB) pipeline scheduling. The three MB pipeline stages contain an entropy decoder, texture decoder, prediction engine, and deblocking filter. Three new coding tools, which are upsample, padding and MVC engine, are developed for SVC and MVC. The three MB pipeline stages start asynchronously as long as their inputs are ready. It eliminates idle cycles and reduces up to 18% of processing cycles. Moreover, a dedicated DRAM controller is embedded in our design to reduce DRAM access latency by optimizing the DRAM command order and data mapping.

Figure 18.3.2 shows the decoding scheduling optimization for SVC spatial scalability (SS) and quality scalability (QS). In SS, the ILP data for the enhancement-layer (EL) decoding are generated after the base-layer (BL) decoding followed by frame-level processes of padding and upsampling. Significant BW is required to arrange the intermediary frame-level data. Since the padding and upsampling can be performed on-the-fly with BL and EL decoding respectively, the padding and upsampling are merged into the three-stage MB pipeline after frame-level re-scheduling. 34% of processing cycles and 72~82% of ILP BW are saved. For SVC QS, the conventional layer-by-layer decoding schedule requires over 41% of the total decoding memory BW for ILP data. A layer-interleaving decoding scheme that decodes data of each layer in one MB in an interleaved manner is proposed. This scheme eliminates the external memory BW of ILP and reduces the total decoding BW by 41~51%.

Figure 18.3.3 shows the proposed branch selection multi-symbol CABAD architecture. The context model (CM) cache is adopted to reduce CM memory access latency and improve decoding throughput. A branch selection scheme is proposed to compute all possible CMs of the next two decoding bins and fetch them in advance. The branch selection scheme with CM cache boosts the throughput to 1.95 bins/cycle. Thus, it suffices for decoding a bitstream of 300Mbps, the maximum bitrate defined in H.264. In SVC QS, the layer-interleaving decoding scheme results in frequent CM memory access because the CM cache must be reloaded once the quality layer is changed. An adaptive write-back control is adopted to update the modified CM groups in each layer. An extra CM cache is added for quality enhancement layer texture coding which is the most frequently used CM type in QS decoding. These two schemes reduce 86% of CM memory access and 23% of processing cycles in SVC QS decoding.

Figure 18.3.4 shows the cache-based MC architecture and DRAM data access optimization approaches. A 2D-mapped two-way associative cache system is adopted to reuse the loaded reference frame pixels. Every reference frame pixel is 2D-mapped into one of the 64 cache banks. Each bank contains two cache lines. In the proposed cache system, luminance and chrominance data share the same tags. Each cache line contains 8×2 Y pixels and the corresponding 4×1 Cb and Cr pixels. The tag sharing strategy eliminates cache checking cycles and tag storage of chrominance data. The cache-based MC reduces data BW by 41% and DRAM latency BW by 30% compared with variable block size MC (VBSMC) [2, 4]. Access pattern reordering is employed to load pixels within the same DRAM bank together. It eliminates 67% of DRAM PRECHARGE/ACTIVE (P/A) activity. To further reduce access latency, the DRAM controller issues P/A commands out-of-order. The above two DRAM access approaches contribute to a DRAM latency BW reduction of 77%. In the proposed system, the DRAM bank mapping differs between the forward- and backward-list reference frames. This bank interleaving pattern can mitigate the problem of bank conflict between two adjacent accesses in B-frame MC. It can further reduce 66% of DRAM latency BW. The total MC BW is 76% less than that of the VBSMC only scheme [2].

The detailed chip features are summarized in Fig. 18.3.5. The core size is 3.88mm^2, which includes 414.3K gates and 9.0 KB on-chip SRAM. The chip micrograph is shown in Fig. 18.3.7. This chip can decode H.264/AVC High-Profile, SVC High Profile, and MVC High Profile with only 11% logic gates and 25% on-chip SRAM overhead. The power consumption is first reduced by architecture-level optimization: A highly parallel architecture for MC is developed to reuse accessed reference pixels and interpolated data for memory and computation power reduction; cache system is applied in MC/ED/DB/UP for less SRAM access. Operation rejection of DB processing and residual SRAM accessing is achieved by detecting zero DB boundary strength and all-zero/DC blocks. Finally, operand isolation and hierarchical clock gating are employed to reduce the power of inactive circuits. A total power reduction of 69% is achieved.

Figure 18.3.6 shows the performance comparison. With the proposed high throughput entropy decoder, cache-based MC with DRAM access optimization and low-power techniques, the proposed design improves the throughput by 1.71× to 3.41× with 47% less power consumption compared with the previous works. Additionally, with the ability to decode SVC, it supports spatial scalability from QCIF to 1080HD, and quality scalability to provide various bitrate-quality-power decoding trade-off points. View scalability for 3D and multi-view applications is also provided with MVC decoding. The proposed decoder can support applications from low-power portable devices to high-end QFHD and 3DTV.

Acknowledgements:
The authors thank UMC University Program for process support and chip fabrication. We also thank National Chip Implementation Center for chip testing. This work is funded by National Science Council under the grant number NSC 97-2221-E-002-191-MY3.

References:

[1] Y. C. Yang, et al., “High-Throughput H.264/AVC High Profile CABAC Decoder for HDTV Applications”, *IEEE Trans. CSVT*, vol. 19, pp. 1395-1399, Sept., 2009.
[2] C. C. Lin, et al., “A 160kGate 4.5kB SRAM H.264 Video Decoder for HDTV Applications,” *ISSCC Dig. Tech. Papers*, pp. 406-407, Feb., 2006.
[3] T. M. Liu, et al., “A 125 μW, Fully Scalable MPEG-2 and H.264/AVC Video Decoder for Mobile Applications,” *ISSCC Dig. Tech. Papers*, pp. 402-403, Feb., 2006.
[4] C. D. Chien, et al., “A 252kgate/71mW Multi-Standard Multi-Channel Video Decoder for High Definition Video Applications,” *ISSCC Dig. Tech. Papers*, pp. 282–283, Feb., 2007.
[5] D. Zhou, et al., “A 1080p@60fps multi-standard video decoder chip designed for power and cost efficiency in a system perspective,” *Symposium on VLSI Circuit*, pp. 262–263, June, 2009.

978-1-4244-6034-2/10/$26.00 ©2010 IEEE

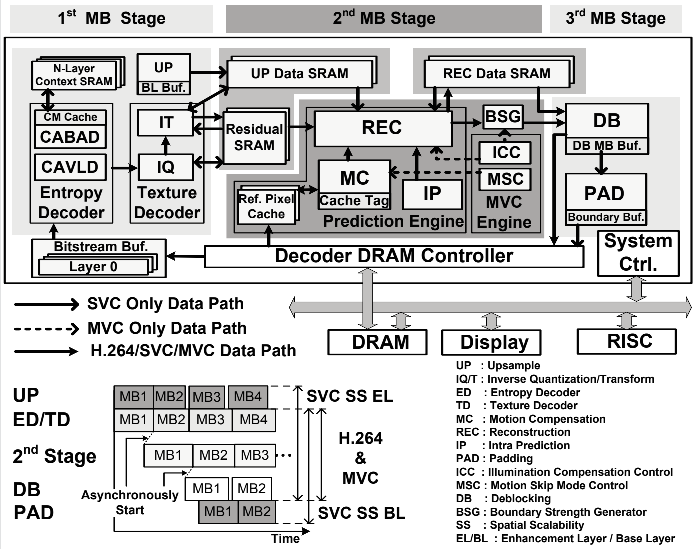


Figure 18.3.1: System architecture and asynchronous MB pipeline scheduling.

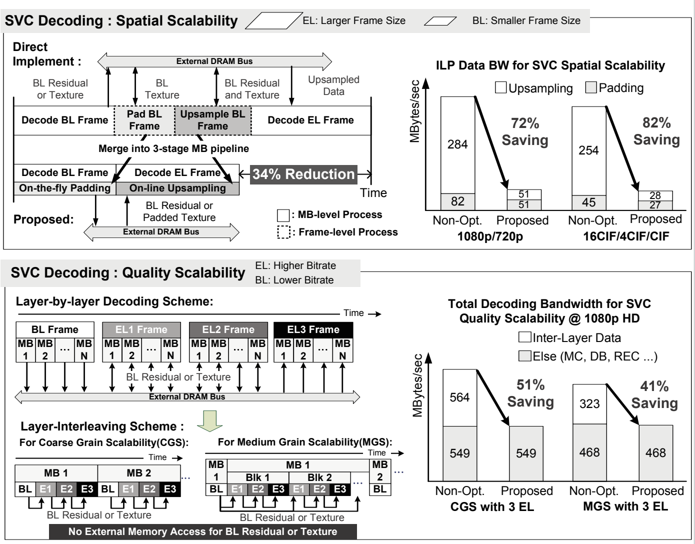


Figure 18.3.2: Proposed decoding scheduling for SVC SS and QS.

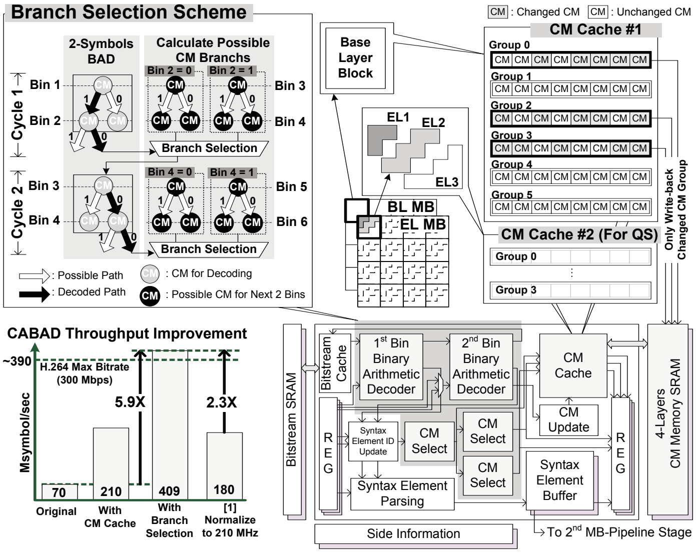


Figure 18.3.3: Branch selection CABAD architecture with CM caches.

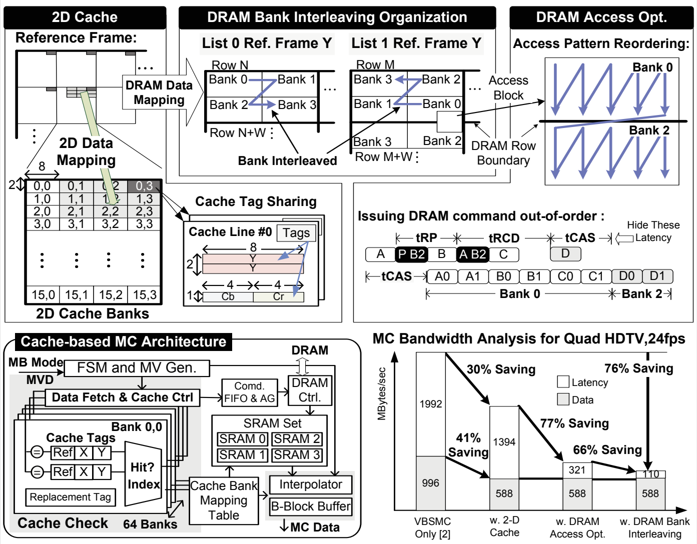


Figure 18.3.4: Cache-based MC architecture and DRAM access approaches.

Technology	UMC 90nm 1P9M Multi-Vth COMS		Supply Voltage	Core 1.0V, I/O 2.5V
Die Size	2.92×2.92mm^2		Core Size	1.97×1.97mm^2
Logic Gates	414.28K (2-input NAND gate)		On-chip SRAM	8.99KBytes
Decoding Feature	H.264 High Profile			
	SVC High Profile (Supporting ESS, Frame mode only)			
	MVC High Profile (Also Supporting IC, Motion Skip Mode)			
Core Power Consumption	H.264	59.52mW@210MHz, 4096×2160p/24fps/Single view		
		40.01mW@133MHz, 1920×1080p/60fps/Single view		
		20.81mW@ 67MHz, 1920×1080p/30fps/Single view		
		11.56mW@ 30MHz, 1280× 720p/30fps/Single view		
	SVC	26.33mW@100MHz, 1920×1080p/30fps/4 MGS Quality Layers		
		27.91mW@120MHz, 1920×1080p/30fps/4 CGS Quality Layers		
		25.14mW@ 98MHz, 1080pHD/720pHD/30fps/2 Spatial Layers		
		14.78mW@ 82MHz, 16CIF/4CIF/CIF/30fps/3 Spatial Layers		
	MVC	55.87mW@200MHz, 1920×1080p/30fps/Three views		
		53.92mW@180MHz, 1280× 720p/30fps/Six views		

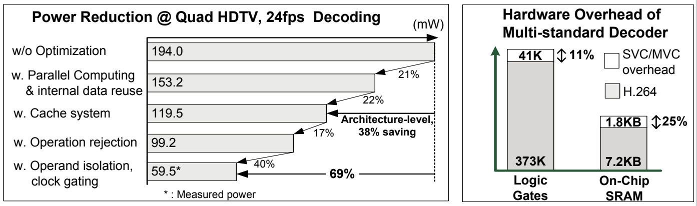
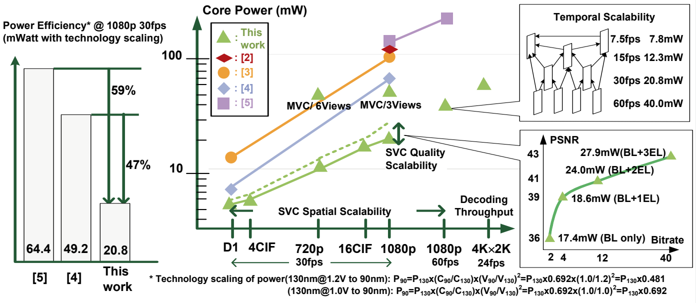


Figure 18.3.5: Chip features, low power schemes and hardware overhead.

	ISSCC 2006 [2]	ISSCC 2006 [3]	ISSCC 2007 [4]	VLSI Sym. 2009 [5]	This Work
Standard	H.264 Main Profile	H.264 Baseline MPEG-2 SP	H.264 Baseline MPEG-1/2/4 SP JPEG Baseline	H.264 High Profile MPEG-1/2/ MP AVS JP	H.264 High Profile SVC High Profile MVC High Profile
Max Specification	1920×1080@30fps	1920×1080@30fps	1920×1080@30fps	1920×1080@60fps	4096×2160@24fps
Technology	180nm(1.8V)	180nm(1.8V)	130nm(1.2V)	130nm(1.2V)	90nm(1.0V)
Logic Gate Count	160K	303K	252K	367K	414K
On-chip Memory	4.5KB	2.8KB	4.9KB	11.0KB	9.0KB
External Mem. I/F	Dual	Dual	Single(AHB)	Single(DRAM)	Single(DRAM)
Core Power for 1920×1080@30fps	108mW(Post-sim @ 130nm, 1.2V)	~110mW	71mW(1.0V)	134mW	20.8mW



* Technology scaling of power(130nm@1.2V to 90nm): $P_{90}=P_{130}x(C_{90}/C_{130})x(V_{90}/V_{130})^2=P_{130}x0.692x(1.0/1.2)^2=P_{130}x0.481$
(130nm@1.0V to 90nm): $P_{90}=P_{130}x(C_{90}/C_{130})x(V_{90}/V_{130})^2=P_{130}x0.692x(1.0/1.0)^2=P_{130}x0.692$

Figure 18.3.6: Comparison with state-of-the-art decoder chips.

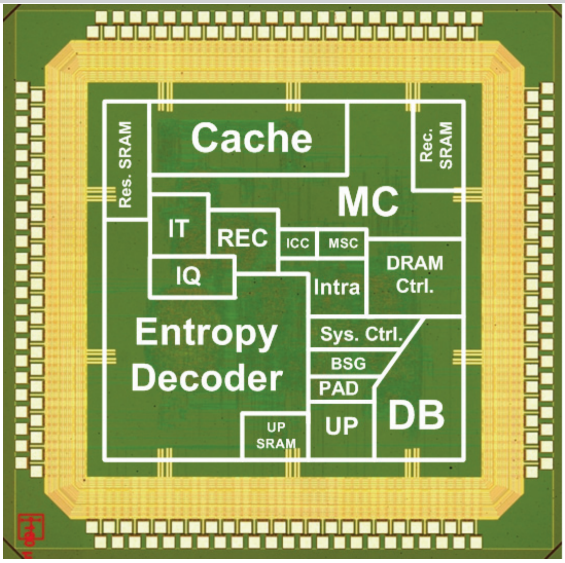


Figure 18.3.7: Chip micrograph.