I. Introduction

4320p Ultra HDTV (Super Hi-Vision) offers a significantly enhanced visual experience relative to 1080p, and is on its way to being the next digital TV standard. This paper presents an H.264/AVC intra-frame encoder chip that supports real-time 7680x4320p 60fps coding. Its design involves two key challenges: 1) data dependencies of video coding algorithms restrict hardware utilization and degree of parallelism; 2) huge computational complexity results in circuit area and power dissipation issues.

Several new techniques are proposed to overcome these barriers. For intra prediction, coarse-to-fine mode decision (CFMD), interlaced block reordering (IBR) and probability-based reconstruction (PBR) alleviate the data dependencies and reduce hardware complexity by 74%. For CABAC, 31% performance gain is obtained from pre-normalization, hybrid path coverage (HPC) and bypass bin splitting (BPBS).

II. 1.991Mpixels/s Intra Prediction

High-throughput intra prediction involves extensive computation. In addition, hardware utilization is challenged by data dependencies of the reconstruction loop. To solve these problems, algorithm and architecture are co-optimized for the intra prediction engine supporting 8x8 and 16x16 modes.

Coarse-to-fine mode decision (CFMD) is first proposed. As shown in Fig. 1, coarse decision (CD) makes prediction block size selection, and sends to the fine decision (FD) stage four best candidate modes for each block for the 8x8 mode, or one best mode for 16x16. FD then refines the candidates to obtain the best result. With CD performed based on original reference pixels (excluding the upper MB boundary), FD is the only part relying on reconstructed pixels, which simplifies the reconstruction loop. Moreover, while FD uses SATD cost function for accurate decision, CD uses SAD as a simpler alternative. Area cost is thereby reduced by 35%.

Based on CFMD, interlaced block reordering (IBR) is proposed for the FD and reconstruction parts. As shown in Fig. 2 (b), 8x8-level scan is always performed in an MB-interlaced order between two neighboring MBs. IBR ensures no dependencies between two successively processed blocks. This allows FD and reconstruction to operate in parallel, reducing clock frequency required for 4320p60 coding by 45%.

Fig. 2 (a) shows the pipeline design targeted at processing each 8x8 block in 8 cycles. FD processes every mode in 6 cycles with 2-cycle latency. Reconstruction takes another 8 cycles. Owing to IBR, reconstructed pixels of block N is not needed until block N+2 starts. But this still leads to a 4-cycle idle for every 2 blocks. Probability based reconstruction (PBR) is proposed to shorten this idle. As shown in Fig. 2 (c), reconstruction starts right after FD finishes the first mode. If a subsequent mode turns out to have a lower cost, reconstruction is promptly restarted for the new mode. As a result, idle time for one block is 0, 0, 2 or 4 cycles when mode 0, 1, 2 or 3 is the best. To minimize the idle time, the modes are processed in an ascending order of SAD costs that have been calculated in CD, with the first two modes owning about 90% probability to cover the best. Finally, each MB can be processed in an average of 33 cycles, 18% reduced from the original 40.

Fig. 2 (d) shows the optimization results. Overall hardware complexity in terms of the product of area and frequency is reduced by 74%. Coding efficiency degradation from CFMD and IBR is 1.8% in terms of bit rate increase.

III. 1.41Gbins/s CABAC

Due to the critical bin-to-bin data dependencies, CABAC is difficult to be parallelized unless on a frame (or slice) level. However, frame parallelism has the disadvantages of memory bandwidth overhead [1], system latency and large area. In this design, we aim at fast CABAC without frame parallelism. The bottleneck is in arithmetic encoding (AE). In recent designs, pipelined multi-bin AE (Fig. 3 (c)) is commonly applied. Its critical path lies in stage 2, where multiple range update (RU) units are organized in serial. Fig. 3 (a) shows a single RU unit in detail. After a 4-1 LUT, the critical delay of RU is co-determined by a complex LPS renormalization containing several layers of selective shifts, and an addition followed by a simple MPS renormalization.

To optimize the RU delay, a pre-normalization (prenorm) architecture is first proposed. As shown in Fig. 3 (b), for LPS bins, a prenorm component added to AE stage 1 is used to renormalize all the 4 pre-calculated LUT candidates (RLPStatb) in advance. Consequently, LPS renormalization can be removed from the critical path, which improves operating frequency by 7%.

Based on prenorm, if an RU unit is only required to process LPS bins (Fig. 3 (d)), the delay of such an LPS-dedicated unit (LU) is significantly shorter than that of a full unit (FU). A hybrid path coverage (HPC) architecture is then proposed as shown in Fig. 4 (b). A 4-bin HPC consists of 4 paths with almost the same delay. Each path contains a specific permutation of 3 FUs and 1 LU. The HPC circuits can thereby process either 4 consecutive bins containing at least 1 LPS (or bypass bin), or the first 3 out of 4 consecutive MPS bins. Compared to the conventional 4-FU serial organization (Fig. 4(a)), HPC achieves 13% higher operating frequency with 5% less bins per cycle. Overall throughput is enhanced by 7%.

Finally, a bypass bin splitting (BPBS) scheme is proposed. As shown in Fig. 1, by utilizing the fact that bypass bins do not influence RU and the context model, they are split from the bin stream and forwarded through a dedicated channel until re-merged into the bin stream before the low update stage. With RU circuits focused on processing regular bins, average bins-per-cycle rate is increased to 4.27. This leads to a 14% performance improvement.

Fig. 4 (c) summarizes the 31% CABAC performance enhancement from the proposed techniques.
IV. Measurement Results

Fig. 5 shows the specification and die photo of the chip in 65nm CMOS. 678.8K logic gates, 27.1KB SRAM and two PLLs are integrated into a 2.07mm² core. At 1.2V nominal supply, the intra prediction engine delivers the maximum throughput of 1991Mpixels/s for 4320p60 video, at least 9.4 times better than previous designs [2-4], due to its high parallelism and significantly improved hardware efficiency. In the meantime, 1.41Gbins/s CABAC is achieved at 330MHz. The corresponding power dissipation is 106.7mW with an energy efficiency of 53.6pJ/pixel, reduced by 72% from the scaled results of [2]. At 0.8V, 1080p30 encoding dissipates 2mW with both engines operating at 9MHz. The best energy efficiency is 24.5pJ/pixel as measured at 0.8V and 33MHz.

Acknowledgements

This research was supported by Knowledge Cluster Initiative (2nd Stage) and Waseda University Ambient SoC GCOE Program of MEXT, Japan, and by JST CREST. The chip fabrication was provided by VDEC, the University of Tokyo in collaboration with STARC, e-Shuttle, Inc. and Fujitsu Ltd.

References


Fig. 1. Chip block diagram.

Fig. 2. Intra prediction optimization and results.

Fig. 3. Critical path optimization for CABAC.

Fig. 4. HPC and CABAC optimization results.

Fig. 5. Chip specification and micrograph.

Fig. 6. Design comparison.