早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論 文 題 目

Decoding Motion Vector based on Block Merging and Motion Compensation with Distance-biased Cache for Energy-efficient VLSI Architectures of HEVC

申 請 者
Shihao WANG

情報生産システム工学専攻
最適化技術研究

2017 年 5 月

Recently, 4K/8K ultra-high definition television (UHDTV) has been regarded as the next frontier of video formats. The increased resolutions have stimulated a new video coding standard in 2013 called HEVC (High Efficiency Video Coding). Compared to H.264 standard, HEVC has announced its twice compression capacity to achieve the same video quality at the expense of increased complexity, which is challenging the VLSI implementation of real-time HEVC decoder system. For VLSI implementations, the increases of computational complexity, external memory bandwidth and on-chip memory requirement caused by HEVC are challenging us to explore energy efficient VLSI architectures for 4K/8K UHDTV video decoding hardware. In detail, to pursue better energy efficiency, it is critical to eliminate the redundant memory accesses and computations by designing customized architectures for the complicated HEVC decoding algorithms.

HEVC inherits the basic idea of previous video coding standards where each sample can be intra or inter predicted. Especially, decoding inter prediction samples contributes the major coding efficiency at the expense of more than 49% of decoding complexity and 60% of memory bandwidth requirement. Therefore, it is indispensable to design VLSI architectures for real-time decoding processing, which consists of decoding motion vectors (called PDec: Parameter Decoder) and motion compensation (MC). However, HEVC employs several new coding features for both PDec and MC, which result a heavier burden on energy efficiency. For example, the increased number and maximum size of coding units require more design efforts because architectures should handle all 24 possible cases, which are three times more than that of H.264. Another example is the longer taps of MC interpolator, which directly introduces more memory bandwidth requirements and computations. In total, it is desirable for new VLSI architectures which are energy efficient to handle these new HEVC features in both PDec and MC.

To remedy this situation, this dissertation presents the proposed energy efficient VLSI architectures for PDec and MC, respectively. These proposals are motivated from a viewpoint of data reuse to reduce both memory accesses and computations by new hardware architectures. The unexploited potentials for data reuse are discovered and utilized to form the main idea in each design. Specifically, in PDec, the potential is that the same result of motion vectors is computed repeatedly for several blocks. Therefore, these blocks are merged together (called “Block Merging”) to reuse this result among blocks for saving computing energy.

In MC, I discovered the potential that possibility of reuse is highly related to the distance between two frames. Thus, I applied this relationship to the cache design (called “Distance-biased Cache”) to improve cache memory energy efficiency. Besides the above proposals, the related memory hierarchy and hardware optimization techniques are also carefully designed so as to achieve better energy efficiency for both computing and memory parts of their VLSI architectures.

The following presents the summaries of four chapters and related evaluation results:

Chapter 1 [Introduction] gives the background and the new features of PDec and MC in HEVC decoder, followed by the discussions on the VLSI design challenges caused by these new features. The perspective viewpoint for solving the design challenges and the research target are then discussed. Meanwhile, the state-of-the-art VLSI architecture designs from both my research lab and other research groups are discussed. Finally, the scope of this dissertation will be given.

Chapter 2 [Block Merging based Unified HEVC Parameter Decoder Design] presents an energy efficient VLSI PDec architecture by block merging. The proposed idea of block merging bases on the discovery of unexploited reuse potentials between blocks in previous works. All these blocks share the same MV result while they are calculated for multiple times, once per block. This work merges these blocks together to calculate this MV result only once and reuse this result for all blocks to save computing energy. By doing this, CU-adaptive pipeline is proposed for block merging to save up to 95.5% redundant computations and memory accesses compared to previous works without block merging. Meanwhile, the idea of block merging can also reduce the number of block shapes from 24 to 4 types to reduce the hardware costs. Besides block merging, a unified architecture decoding not only motion vector but also boundary strength is proposed for memory sharing as well as enhancing data reuse between these two decoding processes. Additionally, the memory architecture and organization are proposed, and PU based coding scheme for co-located storage is introduced for further reducing 30-90% DRAM bandwidth requirement. Finally, the proposed area optimization techniques like index-mapping scheme save area costs by 43.2k logic gates. In total, the proposed PDec design supports real-time 7680x4320@60fps video decoding at 249MHz and it’s the world’s first design for the new HEVC standard. Besides the reduced DRAM memory accesses, we also achieve 36% logic gate reduction by the proposed block merging and

index-mapping scheme compared to the state-of-the-art works. The demerit is the more area consumption for implementing complicated hardware to support the diverse block types in HEVC.

Chapter 3 [Distance-biased Cache based HEVC Motion Compensation Architecture] presents a new cache memory design based on the discovery of the relationship between distance and the data reuse potentials for each reference frame. Instead of fixing all the cache sizes, this work can on-the-fly program the cache into multiple cache sets and determine each cache size based on distances. Based on this idea, a novel cache design with distance biased direct mapping scheme is first proposed. The size of each cache set is inversely proportional to the proposed distance, which is defined as the time interval between current and reference frames. This means a reference frame with small distance is given a large cache set for its good reuse possibility and vice versa. This can achieve a near-optimum hit rate while it involves significantly lower complexity by being direct mapping. Besides, eight-bank cache is organized differently at reading and writing interfaces to double the data delivery efficiency by removing the unused data fetching as many as possible for energy and area saving. Additionally, row based miss information compression is applied and mask-based block conflict check scheme also efficiently solve the potential pipeline hazards as well as reducing the design costs. The proposed architecture achieves 8x throughput enhancement to support 7680x4320@60fps video applications. Compared with the state-of-the-art works, this work shows 76%, 81% and 62% performance improvement in terms of normalized logic gate, memory and power metrics. The demerit is the increased logic gate costs for supporting the reconfigurability of cache.

Chapter 4 [Conclusion] summaries the contributions of this dissertation from the view of data reuse for energy efficiency. The solved and remaining problems are discussed with an expectation for future works.