

Title: HEVC Profile & Level limits, for parallel partitions in particular.

Status: Input Document to JCT-VC

Purpose: Proposal for HEVC Profiles and Levels

Author(s) or Contact(s): Chad Fogg, Aaron Wells
 2975 San Ysidro Way
 Santa Clara, CA 95051
 USA

Tel: +1 408 734 8888
 Email: chadfogg@gmail.com,
awells@ambarella.com

Source: Ambarella

Abstract

This proposal suggests bitstream and decoder parallel partition (PP) tool limits for consideration in the upcoming HEVC profiles and levels discussions. Among the traditional bitstream slice restrictions, the most important in terms of quality impact is that the *lower limit of slice frequency* allow a single slice to cover an entire frame, while the most important slice decoder implementation cost factor is *the upper limit of slice frequency* over a window of time, in particular non-lightweight slices. In addition to slices, this proposal suggests restrictions for the new partition tools of HEVC --- Tiles, Wavefront, and Fine-Grain slices --- that meet the necessary goals motivating each of these schemes, while minimizing their implementation cost. For the Main Profile, it is desired that between 1 and 4 independently decodable tiles and slices be permitted per frame, and no more than 2 lightweight slices shall be coded per 32K coded pixels. Finally, the HEVC specification should explicitly state that no more than one wavefront substream shall be started per LCU row.

Introduction

The proposed Main Profile parallel partition limits are summarized in the table below.

Parallel partition scheme	Limit
Slices (regular / non-lightweight)	Between 1 and 4 per frame.
Lightweight slices	No more than one entropy (CABAC) restart per 32K coded pixels.
Dependent Tiles	Between 1 and 16 per frame
Independent Tiles (F335)	Between 1 and 4 per frame
Entropy Slices	Enabled only for partitions that are $\frac{1}{4}$ frame or smaller
Fine-grain slices (E483)	Same as regular slices
Wavefront (F274)	No more than one substream per row of LCUs within an independently decodable partition

Through the proposals are self-contained in the table above, further rationale and detailed restrictions are provided next. The implementation study by Broadcom (JCTVC-G110) also provides an excellent summary of single core issues.

1.1 Slices

The coding efficiency overhead of frequent slices has been well studied [6] in HEVC, but the implementation costs have not. On the decoder cost of slices, the CABAC reset at the start of each slice incurs a significant processing bubble in the hardware pipeline, raising the baseline clock rate needed to meet real-time decoding. Worse yet, the changes that can take place from one non-lightweight slice header to the next can cause massive state load bottlenecks.

If closely examined, AVC's slice frequency formula permits up to one slice per macroblock when a majority of HRD buffer bits are consumed on one frame, such as a bursty IDR. The formula appears to be aimed at pure-software decoders with complexity amortization windows over more than one frame period, rather than the slight fraction of a picture window of modern low-power hardware.

Several "professional" profiles of AVC limit the maximum slice size to $\frac{1}{4}$ of the coded picture. This slice structure was designed to quell concerns that mid-2000's "single core" hardware could not meet real time, while 4 concurrent decoders could. By the time the codec reached prime, mainstream deployment, this restriction was already unnecessary. Noting the impact on coding efficiency and visibility issues this forced slice structure creates, a group of three companies through the Canadian National Body at the Turin July 2011 MPEG meeting asked [1] that several of the "professional" profiles be amended to allow one slice per frame which already benefits Baseline, Main, and High Profiles in consumer class applications, particularly broadcasting.

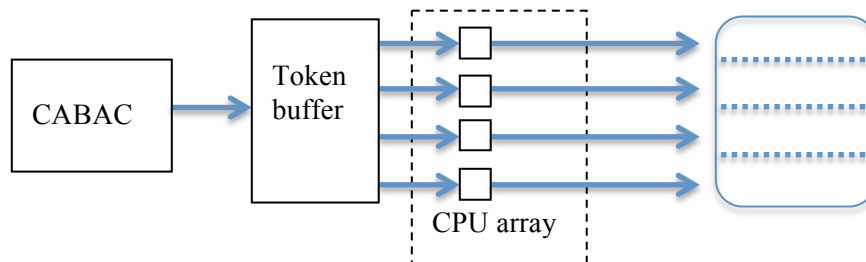
Similarly, MPEG-2 Main Profile had a restriction that a new slice had to commence at the beginning of each macroblock row, yet most first generation of consumer high-def decoders in the early 2000's were able to decode MP@HL (720p/1080i) in a single pipeline core.

1.2 Fine-grain slices

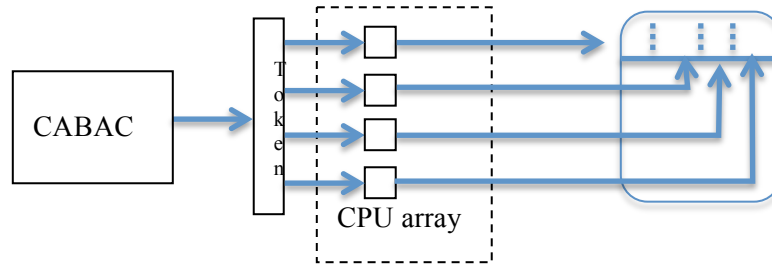
Fine grain slices [2] (F483) allow a slice to have CU granularity instead of LCU granularity for purposes of wasting fewer padding bits when aligning a slice into a network packet, commonly the 1500 byte MTU. Since FG slices have the same implementation impact as traditional slices, it is recommended that the Profile and Level limits for FG slices be the same as regular and lightweight slices.

1.3 Entropy slices

The basic concept of Entropy Slices (ES) is to introduce a large enough token buffer between entropy decoding and the reconstruction process such that multiple cores (such as in a GPU fabric) can operate concurrently once enough tokens have filled the buffer. A token buffer large enough to permit a worst case HDR buffer's worth of coded frame data to be broken into, for example, four equal areas would have a DRAM footprint larger than any reference buffer, and have a penalty of memory traffic on the order of the entire reference and reconstruction process.



It is desirable to limit the size of the token buffer such that it could optionally fit on-chip, in L2 cache or embedded DRAM for example. Therefore, it is suggested that the use of ES be allowed for independent partitions $\frac{1}{4}$ or smaller than the frame.



1.4 Wavefront

Wavefront Parallel Processing (WPP) described in [3] (F274) requires that the decoder maintain a table of bitstream pointers, with one entry for each substream, to indicate where each thread can begin parsing the substream within the greater linear bitstream. An additional CABAC state is required in order to allow the LCU row below to begin decoding as soon as the next LCU within the current row commences. Though the pointer table itself is not a large overhead, managing the table does incur some small cost. It is therefore suggested that a limit of no more than one substream per row of LCU's within an independently decodable partition, such as in combination with tiles [4], be permitted for the Main Profile. This is likely the intention of the proponents, but the pathological cases of several WPP substreams within the same LCU row should be avoided.

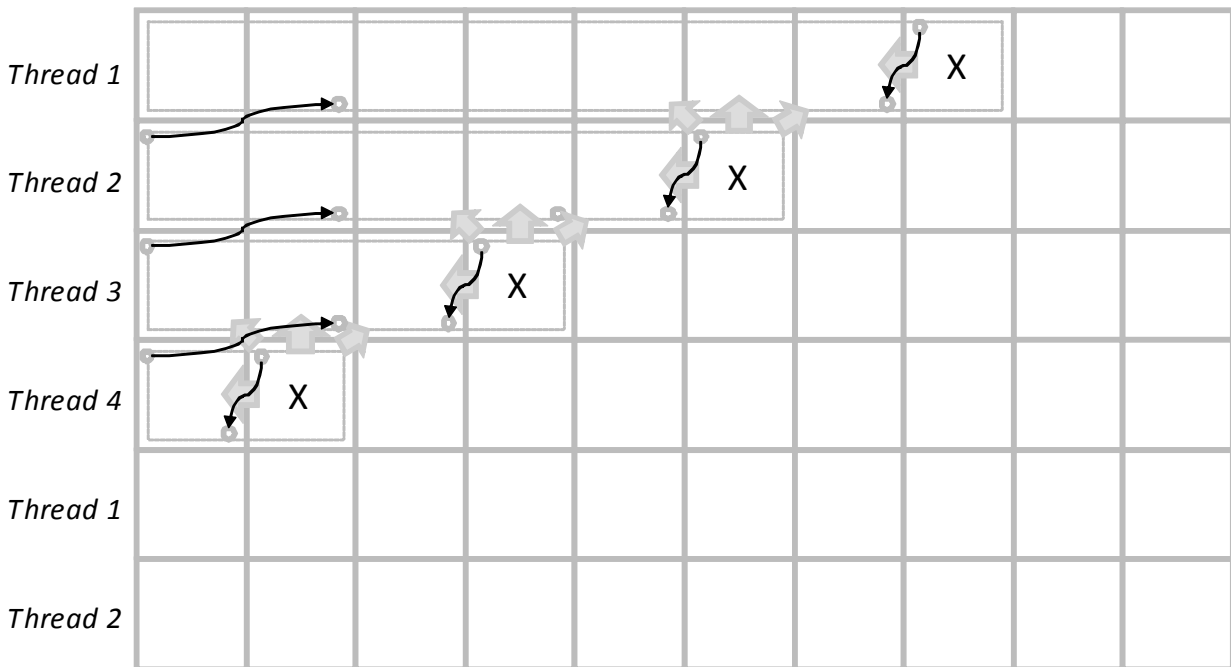


Figure 1 Modified from F274 illustrating 4 threads (one per LCU row) within an independently decodable partition. X is an LCU being concurrently decoded. Thin & long arrows show entropy dependencies, while thick & short arrows show spatial prediction dependencies.

1.5 Tiles

Though deblocking is enabled by default across independent tile boundaries [5], each independent tile potentially creates a visual discontinuity by breaking spatial prediction for the first LCU within an LCU row. It is therefore vital that encoders have the freedom to choose no independent tiles per frame.

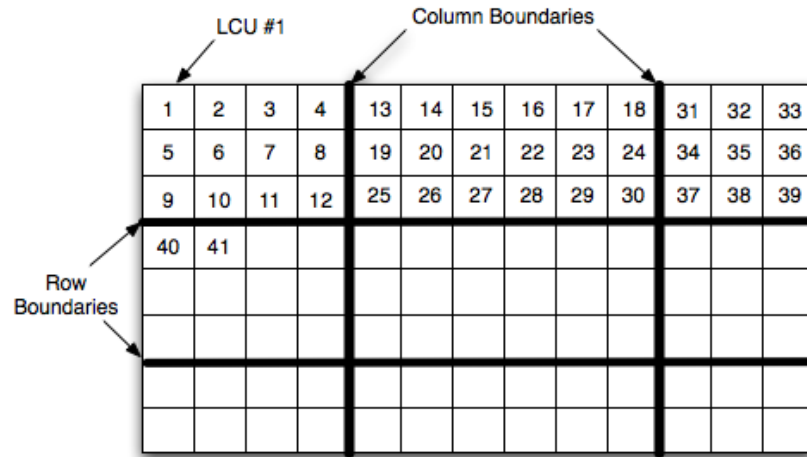


Figure 1. From [5] (F335). An example of Tiles partitioning using three columns and three rows.

2 References

- [1] [m20392](#), “CAN NB Contribution: Comment on ISO/IEC 14496-10, A.3.3(k), High 10, High 4:2:2”, 97th MPEG Meeting, Torino, July 2011.
- [2] [JCTVC-E483](#) , “BoG on fine granularity slices”, 5th JCT-VC Meeting, Geneva, March 2011.
- [3] [JCTVC-F274](#), “Wavefront Parallel Processing for HEVC Encoding and Decoding”, 6th JCT-VC Meeting, Torino, July 2011.
- [4] [JCTVC-F063](#), “Wavefront Parallel Processing with Tiles”, 6th JCT-VC Meeting, Torino, July 2011.
- [5] [JCTVC-F335](#) , “Tiles”, 6th JCT-VC Meeting, Torino, July 2011.
- [6] [JCTVC-F596](#) “The effect of LCU size on coding efficiency in the context of MTU size matching”, 6th JCT-VC Meeting, Torino, July 2011.

3 Patent rights declaration(s)

Ambarella does not have any current or pending patent rights relating to the specific proposals described in this contribution.