

**ISO/IEC JTC 1/SC 29/WG 11**

**Coding of moving pictures and audio**

**Convenorship: UNI (Italy)**

**ISO/IEC JTC 1/SC 29/WG 11 N18438**

**Document type: Approved WG 11 document**

**Title: Thoughts on Immersive Media Decoding Interface for VVC**

**Status: Approved**

**Date of document: 2019-05-18**

**Source: System**

**Expected action: Review and comment**

**No. of pages: 7**

**Email of convenor: leonardo@chiariglione.org**

**Committee URL: mpeg.chiariglione.org**

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/WG 11 N18438**

**Geneva, CH – March 2019**

|  |  |
| --- | --- |
| **Source:** | **System SubGroup** |
| **Title:** | **Thoughts on Immersive Media Decoding Interface for VVC** |

**Thoughts on Immersive Media Decoding Interface for VVC**

Contents

[1 Introduction 3](#_Toc8820200)

[2 Immersive Application Examples 3](#_Toc8820201)

[2.1 Viewport-dependent 360 Tiled Streaming 3](#_Toc8820202)

[2.2 Point Cloud Objects With 360 Videos 4](#_Toc8820203)

[3 Analysis on System Decoder Model 6](#_Toc8820204)

[3.1 Legacy System Decoder Model 6](#_Toc8820205)

[3.2 Problems of legacy system decoder model for the use cases 6](#_Toc8820206)

[3.3 Proposed System Decoder Model for Immersive Media 7](#_Toc8820207)

[4 Next steps 8](#_Toc8820208)

[4.1 Preliminary design requirements on NALU structure and decoder operation 8](#_Toc8820209)

[4.2 Topics to be studied by the Systems 8](#_Toc8820210)

[References 8](#_Toc8820211)

# Introduction

At MPEG #125, the AHG on Systems technologies for VVC was established. As of MPEG #126, the mandates of the AhG are:

1. To facilitate participation of relevant meetings and discussions which happens before the start of MPEG meeting
2. Contribute to development of VVC regarding following topics.
   1. Metadata equivalent to the scope of SEI messages from AVC and HEVC
   2. Metadata equivalent to the scope of VUI messages from AVC and HEVC
   3. High-level syntax required signaling for post decoder processing
   4. Sub-Bitstream extraction and merging (e.g. “high-level syntax related to tile based bitstream extraction and merging”)
   5. Byte stream format
3. Improve document on Immersive Media Decoding Interface for VVC (N18438) and share it with JVET

This document captures the outcome of the discussion that happened during MPEG #126 and especially joint discussion between MPEG experts an JVET experts that took place on Saturday 23rd March.

In particular, the goal of this document is to constitute a working document to be shared with JVET experts to guide future discussion on VVC design and its integration into immersive media applications. Note that the present document only addresses the video decoding interface of such applications. For a comprehensive view of the technical challenges, please refer to the document N18357 “Draft Requirements for Immersive Media Access and Delivery”[1] which covers aspects from media access and delivery to interfaces with GPU, buffer, etc.

# Immersive Application Examples

The following are example of applications bringing new challenges for video-based applications:

* **Tiled 360 videos** in very high resolution
* Large **Point** **Clouds** that can be navigated in 6 DoF
* **Light fields** with lots and lots of small tiles
* A complicated **Scene** **Graph** with many objects to traverse
* All of these can be available in **multiple quality/bitrate** variations
* All of those need to be **decoded and decrypted** with constrained devices at the receiver side.

## Viewport-dependent 360 Tiled Streaming

Omnidirectional content allows the viewers to look around in the content by viewing different projection of this content on the current user’s viewport. Typical applications involve a Head Mounted Display (HMD) that tracks the head position of the viewer based on which the application can dynamically adapt the displayed viewport to the viewer. However, flat devices can also be used too. In this case, the user can either physically move the device to use it as a “virtual window” into the scene or the application enables hand gesture to navigation.

Regardless of the consuming devices, the application only renders a portion of the entire content any given point in time. As a result, streaming techniques leveraging the current position of the user’s viewport emerged to optimize the bandwidth usage by delivery in high quality only the part viewed by the user. These techniques are commonly called viewport-dependent streaming.

For instance, Figure 1shows an example application where the omnidirectional content is made available in independent video streams. Then, the application selects the suitable set of these videos to render the viewport.

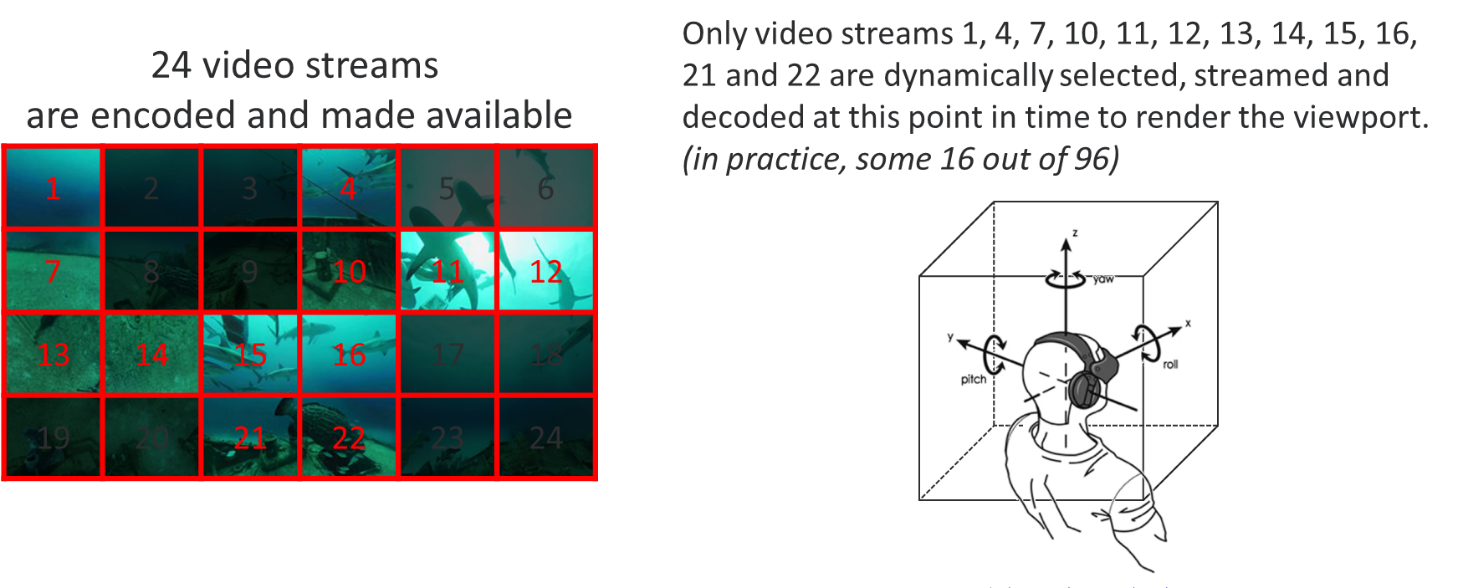


Figure 1 - Viewport-dependent 360 Tiled Streaming

From the system point of view, video elementary stream provides the following features;

* Each original elementary streams can have a different profile, level, tier definition.
* An independently decodable subset of each original elementary stream relevant to specific spatial region and temporal range is identifiable.
* A profile, level and tier of the elementary stream constructed by combining the subsets from several elementary streams is predictable from the information embedded in the original bitstream

In addition, decoder enables the followings;

* The interface between the decoder and the application to exchange information about the maximum profile, level and tier supported by the decoding resources is provided.

## Point Cloud Objects With 360 Videos

Some applications combine different type of immersive content, for instance 2D video and point cloud object. One example depicted in Figure 2 uses a 360-degree 2D video as a background and put two point cloud objects front.

Assuming that use has 6 DoF freedom to navigate inside the area surrounded by the background video at any given point in time, a user will see a combination of a portion of one or more point cloud object and a portion of 2D video which depends on the location, viewport, field of view, etc. For example, the user A will see a front of the solder point cloud object and a portion of 2D video around him, whereas the user B will see a right side of the woman point cloud object and a portion of 2D video around her.



Figure 2 - Point Cloud Objects with 360 videos

Due to the facts that the no user will see entire 360-dgree video and all the points of a point cloud object at once, the client can decode dynamically changing portion of encoded video bitstream for 360-degree video and the point cloud objects for efficient processing instead of decoding entire 360-dgree video and all points of the point cloud objects and rendering some of them. The dynamically changing combination of a fraction of an elementary stream for 360-dgree 2D video and some portions of each 2D video elementary streams carrying components of point cloud objects are extracted and then combined as a single elementary stream. As new elementary stream has been constructed by the combination of multiple independently encoded elementary streams profile/level of the newly constructed elementary stream could be different from that of original elementary streams.

From the system point of view, video elementary stream provides the following features;

* Each original elementary streams can have a different profile, level, tier definition.
* An independently decodable subset of each original elementary stream relevant to specific spatial region and temporal range is identifiable.
* A profile, level and tier of the elementary stream constructed by combining the subsets from several elementary streams is predictable from the information embedded in the original bitstream

In addition, decoder enables the followings;

* The interface between the decoder and the application to exchange information about the maximum profile, level and tier supported by the decoding resources is provided.

# Analysis on System Decoder Model

## Legacy System Decoder Model

Figure 3 and Figure 4 shows the system decoder model of MPEG-2 System (13818-1) and MPEG-4 System (14496-1), respectively. Even though the terms are different, basic architecture is same. A decoder receives an AU of a single elementary stream conform to a profile/level from the decoding buffer and generate a video frames then send them to the entity for display or further processing such as composition. The decoding buffer never underflow or overflow and decoder produces a picture fully occupied by the pixels generated from the received AUs.



Figure 3 - Transport Stream system target decoder notation (13818-1)



Figure 4 - Flow diagram for the systems (elementary) decoder model (14496-1)

## Problems of legacy system decoder model for the use cases

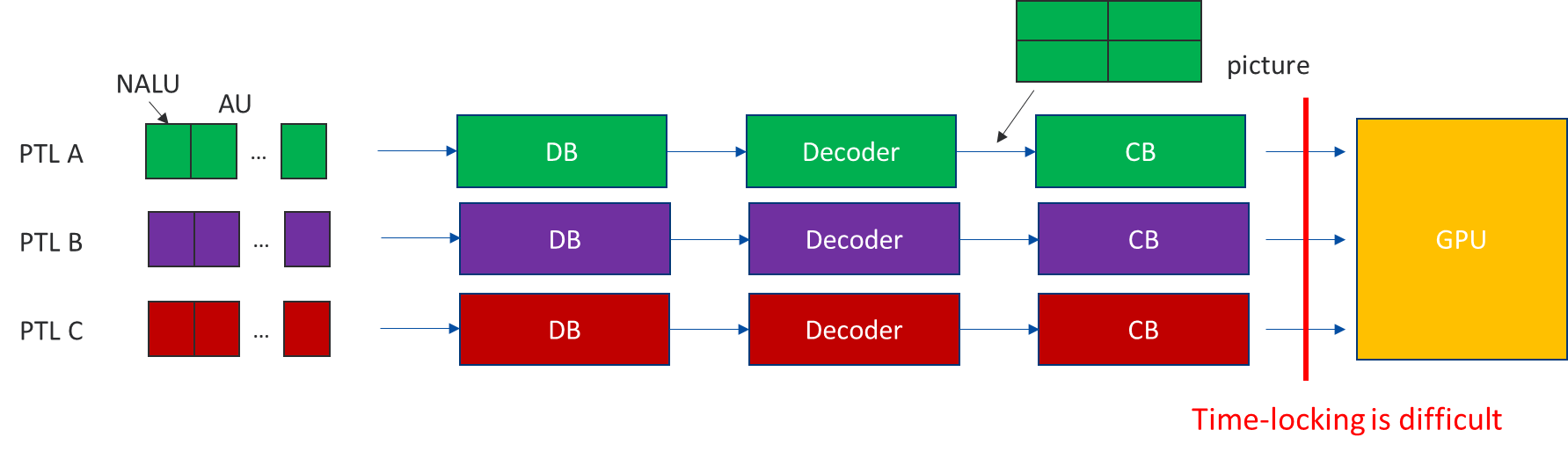
As the legacy system decoder models allow decoding of an elementary stream conforming to a single profile/level by a single decoder, multiple independent decoders need to be used to implement the uses cases described in the section 2 as shown in Figure 5. Such implementation has the following challenges:

Figure 5 - Legacy Decoder Model applied to Immersive Applications

* Decoding resources are wasted as not all the decoded pictures are actually rendered to the user
* No interoperability point between content and application as there are no way to indicate or predict the number of parallel decoding instances supported at the content creation time
* Variability of time consumed for frame decoding across decoder instances while the output pictures are intended to be composed into one picture for rendering.
* Frame accuracy is required for 360 video texture, some environment (web browsers) don’t allow it.

## Proposed System Decoder Model for Immersive Media

A system decoder model allowing efficient implementation of the use cases described in the section 2 is shown in Figure 6. In this model, a single decoder is used to decode an elementary stream comprised with some portions from the original elementary streams generated by the encoders. For example, three elementary streams conform to Profile/Tier/Level (PTL) combination A, B, C, respectively are generated by the encoders. Each elementary stream is composed of number of NALUs in which indication of independently decodable regions of a video frame is indicated. Figure 6 depicts an example that the video frames of three video are divided into four equal size regions. Each regions are encoded as independently decodable regions and carried by separate NALUs. A bitstream extractor and merger (BEAMer) exists before a decoding buffer in this model. There is an interface between a decoder and a BEAMer to indicate the PTL the decoder conform to. The BEAMer also receives information about the portion of elementary streams to extract. The BEAMer extract the portions of bitstreams from the three elementary streams according to the target information received and create a new elementary stream conform to PTL X and send it to the decoding buffer. The decoder process the elementary stream and generates the decoded picture. The decoded picture is a combination of independently decodable regions selected by the BEAMer. As the regions from multiple elementary streams results a single picture synchronization across them are natural in this model.



Figure 6 - New System Decoder Model for Immersive Media

# Next steps

## Preliminary design requirements on NALU structure and decoder operation

The proposed new system decoder model for immersive media described in the previous section introduced several new requirements to study as follows:

* Each source is encoded in different bitstreams conforming to different PTLs
* Each source may be in different format (e.g. resolution, framerate, bit depth, etc.)
* Bitstream extractor & merger operations are limited to copy/paste and simple modification operations
* Each AU may have NALUs of different random access type (e.g. IRAP and non IRAP)

## Topics to be studied by the Systems

To design the proposed new system decoder model for immersive media described in the previous section the Systems will study following topics

* How to specify normative bitstream extraction & merging operation?
* What information needs to be exposed in the NALU header?

# References

1. ISO/IEC JTC 1/SC 29/WG 11 N18357, “Draft Requirements for Immersive Media Access and Delivery”, Geneva, CH – March 2019