**3GPP TSG-SA4 Meeting # 128 S4-241129**

**Jeju, KR, 20th - 24th May 2024**

**Source: Apple Inc.**

**Title: On support of carriage of metadata in CMAF**

**Agenda item: 9.5**

**Document for: Agreement**

# 1 Background and Overview

The current version of CMAF specification [1] allows for only a single track carrying media data to be present in the MovieBox, as noted in section 7.3.2.1. This section mentions that timed metadata tracks can be provided as separate CMAF tracks in a distinct selection set. Although the wording does not explicitly prohibit the presence of metadata tracks in the same CMAF header, Table 3 clearly specifies a constraint of exactly one TrackBox inside a MovieBox. This restriction precludes the packaging of metadata tracks alongside media data within a single CMAF track.

We believe this limitation is unnecessary for use cases where certain media data is closely associated with timed metadata. For example, spatial media tracks, such as stereoscopic video, could benefit from having associated timed metadata within the same file. Although other methods exist to embed metadata within media data samples (e.g., SEI, T.35), metadata tracks offer significant advantages for dynamic metadata that is synchronized and closely associated with media data. Additionally, both the QuickTime File Format [2] and ISOBMFF [3] define multiplexed metadata track format ('mebx') capable of carrying multiple metadata items over a time range. Metadata item keys need not relate to other item keys, providing a flexible means to signal various structural or descriptive information.

Using separate switching sets in CMAF to provide necessary metadata for such use cases is impractical, particularly given the volume of data involved. The primary purpose of CMAF switching sets is to enable flexible late binding [1]. However, the metadata in question is not subject to late binding and maintains a clear one-to-one dependency with the video content. Thus, such metadata is not selected or “late bound” for specific content, as is the case for content in different switching sets. Rather, playing a specific video CMAF track requires the player to fetch the corresponding metadata track, which can be resource-intensive and complicate player logic.

# 2 Relevant deployments

There are existing products that already utilize this concept [4][5][6][7]. For instance, Apple employs a type of timed metadata payload to describe the parallax of decoded stereoscopic video frames. This metadata, known as **caption parallax** timed metadata items, addresses the **risk of depth collision in stereoscopic video**, where captions might intersect with stereoscopic elements having a parallax less than the screen plane. **Adjusting the parallax of captions** to a more negative value than the video prevents such collisions, **avoiding viewer discomfort**.

The document “Video Contour Map Payload Metadata within the QuickTime Movie File Format— Format Additions” [4] specifies a metadata payload structure for describing parallax values associated with 2D areas of a stereoscopic video frame, which is specific to the time-aligned video frame. This payload is carried as metadata items within samples in the QuickTime File Format [2] timed metadata or ISOBMFF [3] multiplexed metadata, using the 'mebx' format type of the 'meta' track handler type. Furthermore, Apple's HTTP Live Streaming tools support segmenting content with the 'mebx' track in a single file closely associated with the media data, with authoring recommendations for stereo video including a should statement on the carriage of parallax metadata if the content has subtitles.

# 3 Alternative methods

There are also other ways on how metadata can be added to CMAF tracks. For example, most codecs support the carriage of metadata in dedicated elementary stream metadata units such as SEI NAL units. In addition to that, an ITU-T T.35 message registration mechanism exists that allows to identify a message of a certain terminal provider. However, these methods while being useful in some cases, are also quite limited in other cases.

For example, each individual video sample will need to contain these units. That leads to additional overhead in samples especially if the same metadata applies to a range of media samples. There are ways to mitigate these issues by defining a persistence model but these are very complex to handle especially when performing operations like seeking, splicing, etc. In addition to that, most APIs will return the entire sample data of a particular sample, when the reader is interested in some metadata it will require more processing as a large chunk of data will be extracted (e.g. including video sample data) and an additional elementary stream parser will need to be invoked to find the necessary information.

CMAF also has support for the Event Message Box from DASH ('emsg') that provides signalling for generic events related to the media presentation time. The primary use-case for it is however the signalling of sparse metadata associated to events on the media timeline such as advertisement opportunities, ratings, content replacements, etc.

Another alternative would be the use of the sample groups these also allow carriage mechanisms for metadata and their associations to samples. However, sample groups have their limitations such as sample level granularity of metadata, increased overhead when metadata changes are frequent, etc.

# 2 Proposal

It is important to note that adding frame accurate metadata to CMAF tracks is a generic problem and not only a problem that occurs in layered video use-cases. We think that VOPS as a good starting point to define requirements that 3GPP can forward to MPEG where the actual normative work on CMAF standardization will happen.

Allowing the multiplexing of this small amount of metadata will result in a simple and efficient implementation, as demonstrated by existing products in mobile ecosystems. We believe this approach is also highly desirable for VOPS and ask SA4 to support the following requirements for consideration by MPEG:

* Specification amendments: Define and standardize the inclusion of metadata tracks in the same CMAF container, detailing technical parameters and constraints.
* Implementation Guidance: Provide clear guidelines for multiplexing metadata with media tracks, suggesting best practices and preferred techniques to ensure broad compatibility and ease of implementation.
* Clarification of benefits: Emphasize the practical benefits such as reduced complexity in player design and enhanced performance from a unified data stream.

# References

1.
2.
3.
4.
5. Apple Inc. "Apple HEVC Stereo Video Interoperability Profile (Beta)," Version 0.9 (Beta) June 21, 2023, [online]: <https://developer.apple.com/av-foundation/HEVC-Stereo-Video-Profile.pdf>
6. Apple Inc. "ISO Base Media File Format and Apple HEVC Stereo Video Format additions," [online]: <https://developer.apple.com/av-foundation/Stereo-Video-ISOBMFF-Extensions.pdf>
7. Apple Inc. "HTTP Live Streaming (HLS) authoring specification for Apple devices," [online]: <https://developer.apple.com/documentation/http-live-streaming/hls-authoring-specification-for-apple-devices>