Text

Description automatically generatedISO/IEC JTC 1/SC 29/WG 03 N807

**ISO/IEC JTC 1/SC 29/WG 03  
MPEG Systems   
Convenorship: KATS (Korea, Republic of)**

**Document type:** Output Document

**Title:** Technologies under Consideration for ISO/IEC 14496-12

**Status:** Approved

**Date of document:** 2023-01-23

**Source:** ISO/IEC JTC 1/SC 29/WG 03

**No. of pages:** 33 (with cover page)

**Email of Convenor:** young.L @ samsung . com

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg3>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 03 MPEG SYSTEMS**

**ISO/IEC JTC 1/SC 29/WG 03 N807**

**January 2023, Virtual**

|  |  |
| --- | --- |
| **Title** | **Technologies under Consideration for ISO/IEC 14496-12** |
| **Source** | **WG 03, MPEG Systems** |
| **Status** | **Approved** |
| **Serial Number** | **22331** |

**Abstract**

The document contains following technologies under consideration for the ISO base media file format (ISO/IEC 14496-12):

[1 Handling lost or corrupted samples using a sample group for corrupted samples 2](#_Toc125363640)

[2 Edit lists and movie fragments 4](#_Toc125363641)

[3 Multiplexed timed metadata tracks 5](#_Toc125363642)

[4 Embedded Metadata Signaling 10](#_Toc125363643)

[5 On MIME type parameters 11](#_Toc125363644)

[6 On relation of entity groups and sample groups 12](#_Toc125363645)

[7 Stronger defaulting in Track Runs 13](#_Toc125363646)

[8 Sample reordering in Track Runs 21](#_Toc125363647)

[9 Segment Index and Level Assignment 25](#_Toc125363648)

[10 Generic sub-picture track grouping extensions 32](#_Toc125363649)

[11 Signaling of Multi-Layer Picture Compositing Information in the VVC File Format 35](#_Toc125363650)

[12 Integrating new codecs 41](#_Toc125363651)

[13 MovieFragmentHeaderBox update 43](#_Toc125363652)

[14 Sample Run Sample Group 43](#_Toc125363653)

[15 Improvements to movie fragments and dynamic tracks 46](#_Toc125363654)

[16 Support for Haptics 59](#_Toc125363655)

[17 Event-based media signaling 59](#_Toc125363656)

[18 Use of Preselection for signaling Picture in Picture in ISOBMFF 63](#_Toc125363657)

[19 Attaching collection timestamps and large metadata to every samples 68](#_Toc125363658)

[20 Carriage of video with depth (RGBD) 81](#_Toc125363659)

# Handling lost or corrupted samples using a sample group for corrupted samples

Issues for the original contributions: [m57362](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/66) and [m58084](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/74).

Updated at MPEG#140 with input contribution [m60772](https://dms.mpeg.expert/doc_end_user/current_document.php?id=84267).

*Define a new sample group ‘corr’*

class CorruptedSampleInfoEntry()  
extends SampleGroupDescriptionEntry ('corr')  
{  
 bit(2) corrupted;  
 bit(6) reserved;  
 if (corrupted==2)  
 bit(32) codec\_specific\_param;  
}

corrupted indicates the corruption state of the associated data. Value 0 means that the entire data is lost, and the associated data size (sample size, or NAL size) shall be 0. Value 1 means that the data is corrupted without any additional information on the corruption. Value 2 means that the data is corrupted with codec specific information on the corruption. Value 3 is reserved.

codec\_specific\_param indicates codec specific information on the corruption. The coding format is the one of the sample associated to this sample group description.

Note: codec\_specific\_param information being dependent on the coding format, file writers may need to add and associate a different CorruptedSampleInfoEntry() entry with a sample each time the coding format is changing across samples.

If a data is not associated with a CorruptedSampleInfoEntry or if a data is associated with a description\_group\_index = 0 by a sample group with the grouping\_type 'corr', this means the data is not corrupted.

The processing of a sample with corrupted equal to 1 or 2 is context and implementation specific.

For *NALUFF, state:*

For NALU based codecs, we propose the following semantics:

In the following, the term parameter-set-like NAL units refer to parameter set NAL units, DCI NAL unit, and OPI NAL units collectively.

For NALU-based video formats, the codec\_specific\_param field of the CorruptedSampleInfoEntry is defined as a bit mask, with most significant bit first, of the following flags:

* DecodingParameterSetCorruptedFlag (value 0x00000001): indicates that one or more parameter-set-like NAL unis in the associated data that are needed for decoding the bitstream are corrupted.
* NonDecodingParameterSetCorruptedFlag (value 0x00000002): indicates that one or more parameter-set-like NAL units in the associated data that are not needed for decoding the bitstream are corrupted.
* ConformanceSeiCorruptedFlag (value 0x00000004): indicates that one or more SEI NAL units in the associated data that contain SEI messages affecting the HRD conformance of the bitstream are corrupted.
* EssentialSeiCorruptedFlag (value 0x00000008): indicates that one or more SEI NAL units in the associated data that contain essential SEI messages not affecting the HRD conformance of the bitstream are corrupted.
* NonessentialSeiCorruptedFlag (value 0x00000010): indicates that one or more SEI NAL units in the associated data that contain non-essential SEI messages not affecting the HRD conformance of the bitstream are corrupted.
* VclHeaderCorruptedFlag (value 0x00000020): indicates that one or more NAL unit headers, slice headers or picture headers of the VCL NAL units in the associated data are corrupted.
* VclDataCorruptedFlag (value 0x00000040): indicates that VCL data of one or more slices in the associated data is corrupted, where VCL data refers to data in a VCL NAL unit excluding the NAL unit header, the slice header, and the picture header, if any.
* OtherNonVclNalCorruptedFlag (value 0x00000080): indicates that one or more non-VCL NAL units in the associated data that are not parameter-set-like NAL units and are not SEI NAL units are corrupted.
* RefPicCorruptedFlag (value 0x00000100): indicates that one or more of the reference pictures of the slices in the associated data are corrupted.
* RefDecParamSetCorruptedFlag (value 0x00000200): indicates that one or more of the parameter-set-like NAL units needed for decoding the slices in the associated data are corrupted.

A codec\_specific\_param with value 0 means no information is available for describing the corruption.

A CorruptedSampleInfoEntry may be used with a sample group of grouping\_type 'nalm' and a NALUMapEntry, using the grouping\_type\_parameter 'corr'. The groupID of the NALUMapEntry map entry indicates the index, starting from 1, in the sample group description of the CorruptedSampleInfoEntry. A groupID of 0 indicates that no entry is associated (the identified data is present and not corrupted).

# Edit lists and movie fragments

## Potential use cases

* Documenting the equivalent of the edit list but in fragments
* Enabling the fragmentation in the middle of a sample.

## Edit Adjustment

aligned(8) class EditAdjustmentBox extends FullBox('efrg', version, 0) {  
 unsigned int(6) reserved = 0;  
 unsigned int(1) initial\_adjust;  
 unsigned int(1) final\_adjust;  
 if initial\_adjust { unsigned int(32) initial\_delta; }  
 if final\_adjust { unsigned int(32) final\_delta; }  
}

EditAdjustmentBox should be placed near (after) the 'tfdt' box. If this box is absent, it is equivalent to it being present with initial\_adjust and final\_adjust set to false (0).

In an initial movie, if either there is no edit list (no 'edts' container box), or the initial track is empty (refers to no samples) and there is an edit of zero duration, it is equivalent to having final\_adjust set to false (0).

Derive the following two values:

1. insert\_start = earlest\_comp\_time + { initial\_adjust ? initial\_delta : 0}
2. insert\_end = latest\_comp\_time - { final\_adjust ? final\_delta : 0}

Then, on receipt of a fragment, if the previous fragment has no final\_adjust and this fragment has no initial\_adjust, extend the existing edit to insert\_end; otherwise, insert a new edit from insert\_start to insert\_end.

[Ed.: There is a small (well-known) problem here: the insert\_end is not quite right automatically, as it should include the composition duration of the latest sample; the snag is, we do not know it. There doesn't seem to be a good solution to this that doesn't also raise problems we are trying to avoid: in essence we need this 'edit adjustment' to talk about a 'little bit' of the future (the composition duration of the last frame) without straying into (e.g.) the next fragment. Fixes might include making the final\_adjust signed, explicitly including the frame duration, changing the editadjustment from having a final\_adjust to having a duration from the insert\_start, but none of these are ideal. Which is probably why this is still a discussion point.]

# Multiplexed timed metadata tracks

## Basic Design

The basic design is in the 7th edition Amendment 1. However, we have possible extensions on the table; inline associations, sample groups, and structurally-dependent metadata.

## Extensions

## Carrying inline associations

### General

In some circumstances, it can be useful to carry inline definitions. This allows for more flexibility, at the expense of risking that a reader will encounter metadata items it does not recognize and did not expect.

The support is fairly simple; a box in the sample entry to warm that inline carriage may occur, and then a value item box in the sample data that carries both a MetadataKeyBox (the declaration) and a value box.

Tracks without inline keys offer a few advantages:

* A client can determine the entire set of keys that may be present in the track allowing the client to ignore the entire track if no keys of interest to the client are present.
* Space is optimized as keys are carried once in the sample entry and values have only a box header to frame their data and associate them with their key. Inline key/value boxes carry a MetadataKeyBox so if multiple inline keys are present in the same access units, they do not share the key with sibling boxes or with other access units.
* If the sample group optimization technique described earlier is used, the client can determine which samples contain values for particular keys.

Tracks signaling the presence of inline keys offer a few advantages:

* If a new key is required (e.g., for new key, data type, locale, etc. combination), an inline key/value box can be written at any time. This contrasts to the non-inline key case where the set of keys must be known a priori.
* A client does not need to be able to enumerate all combinations of key properties (key, datatype, locale, etc.) to write the track.

With that said, it is recommended that tracks be created without using inline keys. Inline keys can be seen as an optional fallback facility to be used when needed.

A movie may contain a combination of metadata tracks, some using inline keys and some not using inline keys. Where the set of keys can be known a priori, non-inline key- bearing tracks can be used. Where that’s not possible, inline keys can be used.

### Sample entry

The optional MetadataInlineKeysPresentBox indicates if inline key/value boxes might occur in corresponding access units. If MetadataInlineKeysPresentBox is absent or present with a presence indicator of 0, no inline key/value boxes (value boxes with a local\_id of 0xFFFFFFFF) should occur in the access units.

The box is defined as:

aligned(8) class MetadataInlineKeysPresentBox extends Box('keyi') {   
 unsigned int(8) inlineKeyValueBoxesPresent;  
}

inlineKeyValueBoxesPresent is a Boolean integer that should be set to a non-zero value (the value 1 is preferred) if inline key/value boxes are known to be present or might be present in the access units associated with this sample entry.

If MetadataInlineKeysPresentBox is present but inlineKeyValueBoxesPresent is set to 0, access units should be treated as though no MetadataInlineKeysPresentBox is attached to the sample entry. Whether MetadataInlineKeysPresentBox is absent or inlineKeyValueBoxesPresent is set to 0, access unit metadata values having a local id of 0xFFFFFFFF should be ignored.

This approach allows a sample entry to reserve space for and include a MetadataInlineKeysPresentBox but to rewrite just the inlineKeyValueBoxesPresent field to 0 to signal there are no inline key/value boxes present.

If all sample values include inline keys, a MetadataKeyTableBox shall still be present although it may be empty (i.e., it contains no MetadataKeyBoxes). It is also possible to have a combination of some known keys signaled in the MetadataKeyTableBox and some inline key/values signaled with a MetadataInlineKeysPresentBox.

### Sample data item

If the access units associated with the BoxedMetadataSampleEntry contain inline key/value metadata, each inline item is carried in a box with a local\_key\_id of 0xFFFFFFFF and conforming to the type MetadataInlineKeyValueAUBox defined as:

aligned(8) class MetadataInlineKeyValueAUBox extends MetadataAUBox(0xFFFFFFFF) {  
 MetadataKeyBox inline\_key; // local\_key\_id of '1key'   
 MetadataAUBox inline\_value; // local\_key\_id of '1val'  
}

inline\_key is a MetadataKeyBox where local\_key\_id is set to ‘1key’ (for “one key”).

inline\_value is a MetadataAUBox where local\_key\_id is set to ‘1val’ (for “one value”).

The MetadataInlineKeyValueAUBox can be viewed as a MetadataAUBox with two differences:

* It is a container box carrying two boxes, one a MetadataKeyBox holding the key and the other a MetadataAUBox holding the value for the metadata item.
* It has a local\_key\_id (or box type) with the special value 0xFFFFFFFF. All inline key/value boxes share this special local\_key\_id of 0xFFFFFFFF regardless of the contained value’s key.

Because a MetadataInlineKeyValueAUBox carries both the key and the value using that key, this box alone is sufficient to carry what would otherwise require a MetadataAUBox and an associated BoxedMetadataSampleEntry with a MetadataKeyTableBox having the same local\_key\_id as the MetadataAUBox. This allows any non-inline key and associated value to be converted to an inline key/value box. The reverse transform (inline key/value box to non-inline value and sample entry) is possible, too.

While possible, writing a MetadataInlineKeyValueAUBox declaring a key that’s also declared within the MetadataKeyTableBox (i.e., it carries a duplicate MetadataKeyBox) is strongly discouraged. The presence of a MetadataInlineKeysPresentBox signaling the presence of inline keys defeats optimizations that are possible when all available keys are declared within the MetadataKeyTableBox and no inline keys are used. Using inline keys should be reserved for cases when the keys to be written cannot be known at the time the sample entry is constructed. Section xx also discusses the use of inline keys.

## Using sample groups to optimize key searches

### General

This section describes an optional mechanism to optimize searches for metadata track access units containing particular key/value pairs. While this mechanism’s support is not required, it allows a reader to locate access units with values for a key without having to read each access unit in the track and scan for the key’s value. This can be useful for some kinds of metadata (i.e., values that don’t occur in every metadata access unit) but for others (e.g., GPS metadata) it may not add much value. Different tracks in the same movie may choose to use or not to use this optimization.

If inline key/value boxes are used, the mechanism described here can be used to locate those access units with inline key/value boxes. This can be useful in limiting the scan for metadata items with keys only found in inline key/value boxes.

The following subsections describe the details.

### Sample group overview

A metadata track conforming to this specification may optionally make use of the SampleGroupDescriptionBox and SampleToGroupBox constructs to optimize searching for access units containing particular keys or inline keys. This can be characterized as having a ‘key search sample group.’

### Optimizing search with a sample group

In a metadata track containing one or more sample entries, the MetadataKeyTableBox() in the BoxedMetadataSampleEntry can be used to determine possible keys present in the track’s AUs. If a key is not present in the MetadataKeyTableBox(), it is known that the key doesn’t exist in any AUs. It doesn’t however indicate which samples have particular keys (and associated values). Therefore, to determine which metadata keys are present in the track requires an exhaustive search of AUs (associated with that sample entry) in the metadata track.

While it would be possible to create a track with sample entries for each combination of keys present in the track and only associate the samples with that combination with the particular sample entry, having many sample entries may not be ideal or easily done. An alternative (described here) is to define a new kind of sample group that indicates the keys present in each AU.

The new sample group consists of a SampleGroupDescriptionBox holding a new group description for each new combination of keys present in AUs. If all AUs consist of the same four keys, for example, there would be one group description with these four keys. If the set of keys varied, there need only be as many descriptions as there are different sets of keys present in AUs.

A client looking for AUs with a particular key (or keys) would first consult the sample entry (or sample entries if there are more than one) and determine if the key is present in the set of possible keys (via MetadataKeyTableBox()). If this succeeds, the client would check if the optional sample group exists, and finding this to be the case, the client would walk through the SampleToGroupBox checking if the corresponding sample group description contains the key. As these operations require only information present in the MovieBox(), direct reading and processing of AUs is unnecessary.

NOTE: While “key” is used here as being present in the sample group description, an equivalent, more compact identifier is used.

### The keysearch sample group

For this specification, an optional sample group known as a “key search sample group” is defined. It consists of SampleGroupDescriptionBox and SampleToGroupBox having the grouping type ‘keyp’.

The SampleGroupDescriptionBox will contain variable-sized SampleGroupDescriptionEntries, each of type MetadataKeySearchGroupEntry. MetadataKeySearchGroupEntry is defined as:

class MetadataKeySearchGroupEntry() extends SampleGroupDescriptionEntry ('keyp') {  
 unsigned int(32) entry\_count;  
 unsigned int(32) local\_key\_ids\_array[entry\_count];   
}

entry\_count is a 32-bit unsigned integer holding the number local key ids that follow in local\_key\_ids\_array[].

local\_key\_ids\_array is an array of 32-bit integers corresponding to the local\_key\_id field used in the associated MetadataKeyTableBox() and the local key ids used in associated metadata track access units. A value of 0 is reserved and can be used to mark an array entry as absent. A value of 0xFFFFFFFF indicates the associated access units all contain one or more inline key/value boxes.

Each sample group description entry signals the presence of one or more keys from the key table found in the sample entry associated with the sample(s). Access units associated with this sample group description shall have corresponding metadata values with these same keys.

Each key in use is signaled by using the 32-bit integer value of the local\_key\_id field associated with the MetadataKeyTableBox entry. This local key id is also used in access units as the type of Box holding the corresponding value.

If two samples differ in the keys present, they cannot share the same sample group description. A sample group description for each combination should be created.

NOTE 1 While not strictly required, it is recommended that the order of local\_key\_ids be the same as the order of local key ids in the MetadataKeyTableBox of the sample entry. This can be followed by 0xFFFFFFFF if there is an inline key present. This prevents group descriptions with the same set of keys but differing only in key order from creating multiple, trivially different sample group descriptions.

NOTE 2 There is no relationship between the order of keys in the MetadataKeySearchGroupEntry and the order of values for those keys in the associated access unit(s).

A version 0 SampleGroupDescriptionBox should not be used.

Finally, if a sample group spans multiple sample entries with different sets of keys, the local key ids present in the sample entries spanned must be compatible (i.e., the local\_key\_id must be present in each MetadataKeyTableBox and the corresponding key table entry must be the same). An easy way to accomplish this is not to have samples from different sample entries share the same MetadataKeySearchGroupEntry.

## Structurally dependent metadata

### General

If the metadata values in a metadata track are dependent upon another track in a way that a change in the other track may invalidate some or all metadata item values, the dependent metadata items and metadata track itself are termed structurally dependent upon the other track. Structurally dependent metadata tracks are linked to the track upon which they are structurally dependent using a track-reference of type ‘cdep’. The ‘cdep’ track-reference should be used in addition to the ‘cdsc’ track reference because the ‘cdep’ track-reference’s purpose is to indicate tracks needing attention if another track is transformed (e.g., a video track being scaled or cropped during a reencode to a new file where the metadata will also be copied). Tracks having a ‘cdsc’ track-reference without a ‘cdep’ track-reference can be passed through directly so long as there are no other conditions restricting pass-through. Metadata tracks having a ‘cdep’ track-reference may also need to have metadata items transformed or deleted due to the change in the other track (e.g., the video track).

### MetadataStructuralDependencyBox

The interpretation of a metadata value may depend upon a detail of another track. For example, a geometric value such as a region of interest may be interpreted in the coordinate system of a video track. If the other track is transformed (e.g., scaled or cropped in the case of video), the metadata value may become invalid. To signal which metadata items are structurally dependent, a MetadataStructuralDependencyBox() should be present in MetadataKeyBox() for such metadata items.

The metadata track should also have a ‘cdep’ track reference to the other track upon which values are structurally dependent. This reference is used to determine which metadata tracks might need attention if the target track is manipulated.

If the other track is transformed, currently one of three things can occur to the dependent metadata items:

* + If the values can be transformed in response to the change in the other track, metadata values can be read, transformed, and written in their corrected form. This will typically be limited to being performed in a process that reads the source movie and writes a new one. This requires that the code performing this transformation be able to understand the change in the other track and how to transform the dependent metadata values.
  + If the values cannot be transformed, the metadata item can be deleted by removing the MetadataKeyBox() from the MetadataKeyTableBox() (i.e., setting the local id to 0 and optionally removing the MetadataKeyBox()) and optionally removing metadata values from associated access units. Again, this will typically be limited to being performed in a process that reads the source movie and writes a new one. This removal should be avoided if possible but is available for cases where the transform is not understood, the metadata values are not understood, or the change cannot be applied to understood values.
  + Another option is to mark the structurally dependent value as invalid without rewriting access units or deleting the metdata item from the key table. This allows the presence of metadata items known to have become invalid. Clients reading structurally dependent but invalid items may choose to ignore these or do whatever they see fit. They should not treat them as valid.

The MetadataStructuralDependencyBox() is a Box with this definition:

aligned(8) class MetadataStructualDependencyBox extends Box('sdpd') {   
 MetadataStructuralDependencyInfoBox info;  
}

MetadataStructuralDependencyInfoBox is a FullBox having one currently defined flag value. This flag can be used to mark a key table entry as invalid under the structural dependency.

Other children boxes may be introduced in the future.

The MetadataStructuralDependencyInfoBox is a FullBox with this definition:

aligned(8) class MetadataStructualDependencyInfoBox   
 extends FullBox('sdpi', 0, flags) {  
};

flags can have the lowest order bit (i.e., 0x000001) set to indicate the structural dependency is invalid. If this bit is clear, the metadata item and associated values should be considered valid.

Newly written MetadataStructuralDependencyBox() should have this flag in their contained MetadataStructuralDependencyInfoBox() be clear. New flags may be introduced in the future.

NOTE: In the future, other children boxes of MetadataStructuralDependencyBox() may be introduced. The current thinking is that these will help clients understand under which kinds of transforms the values might remain valid. For example, a video scaling where metadata values use a normalized range of 0...1 might not need to be transformed. If the metadata values used pixels, they might however need to be transformed.

# Embedded Metadata Signaling

It would be beneficial to signal the presence of the embedded metadata at the file level.

We propose to introduce a new SampleGroupDescription box entry that signals the presence of metadata of a particular type. That sample group entry is then used with SampleToGroup boxes to signal the location of the samples that contain the related metadata. The type of the metadata is signalled using a URN scheme that uniquely identifies the metadata.

We define a new SampleGroupEntry with the following syntax and semantics:

class MetadataSampleGroupEntry ()   
 extends SampleGroupDescriptionEntry ('emmd')  
{  
 string metadata\_type;  
}

metadata\_type: a URN that uniquely identifies the type of metadata that is carried.

Additional metadata-specific information may also be provided in this sample group description entry.

[[ed: We could do with an example, to illustrate the usage. Would we allow further parameters, dependent on the URN, after the URN (whereupon we should box the URN)? There is a sample group rather like this in the timed metadata multiplex; should we harmonize? We need the 4CC defined (e.g. 'emmd' for embedded metadata?).]]

# On MIME type parameters

## Introduction

This section discusses issues and solutions for signalling of important video information, as part of the MIME type parameters, for HDR/WCG video, and video with display orientation changes. Hereafter, important video information refers to video information that may be used for content selection, e.g., selection of a video track or a part thereof for consumption.

### Video with display orientation changes

AVC and HEVC both support video content for which the decoder side should apply a transformation of rotation and/or flipping to the cropped decoded picture prior to display, indicated by the display orientation SEI message. Such video is also referred to as video with display orientation changes. Video with display orientation changes need special post-decoding rendering processing to generate desirable viewing experience.

### Signalling of HDR/WCG information in ISOBMFF

The HDR/WCG information can be signalled using the ColourInformationBox defined in clause 12.1.5 of the ISOBMFF specification, for example with the colour\_type equal to 'nclx', in which case the most important HDR/WCG information would be carried in the fields colour\_primaries, transfer\_characteristics, matrix\_coefficients, and full\_range\_flag.

In addition, certain HDR/WCG solutions make use of dynamic metadata conveyed in SEI messages.

### The restricted scheme design in ISOBMFF

The restricted scheme design in ISOBMFF is for handling of situations where the file author requires certain actions on the player or renderer, to enable players to simply inspect a file to find out such requirements for rendering a bitstream and stops legacy players from decoding and rendering files that require further processing. The mechanism applies to any type of video codec.

The mechanism is similar to the content protection transformation where sample entries are hidden behind generic sample entries, 'encv', 'enca', etc., indicating encrypted or encapsulated media. The analogous mechanism for restricted video uses a transformation with the generic sample entry 'resv'. The method may be applied when the content should only be decoded by players that present it correctly.

The restricted scheme is specified in clauses 8.15.1 to 8.15.3 of the ISOBMFF specification.

## Problems and discussions

[Ed. (FM): Issues and solutions in this section were initially introduced as Items 6 (first part) and 7 from m40373 (MPEG #118 meeting)]

The following problems related to the MIME type parameters and signalling of HDR/WCG video, and video with display orientation changes were observed:

1. For video with display orientation changes, a special restricted scheme is missing.
2. There lacks a mechanism to include important video information for HDR/WCG video as part of the MIME type parameters.

## Proposal

The following methods are proposed to solve problems:

1. To solve the first problem, a new restricted scheme type, 'vdoc', is defined, which, when used, indicates that the track carries video with display orientation changes. No further information is provided, and the SchemeInformationBox may be absent in the RestrictedSchemeInfoBox.
2. To solve the second problem, a new optional MIME type parameter 'hdrinfo' is defined to contain the important information of HDR/WCG video. The format of this optional MIME type parameters is a single value or a comma-separated list of values, where each value consists of one or more dot-separated elements. A value of the 'hdrinfo' parameter contains four fields, in the form of "elment1.elment2.elment3.elment4", where the four elements 1 to 4 are the hexadecimal representations of the fields colour\_primaries, transfer\_characteristics, matrix\_coeffs, and full\_range\_flag, respectively, as defined in clause 12.1.5 of the ISOBMFF specification.

# On relation of entity groups and sample groups

For tracks, ISOBMFF provides various tools at different levels either to group together tracks or samples. When a new grouping type (represented by a new 4CC) is defined, it is possible that some of the parameters characterizing this grouping type are common to all tracks or a subset of tracks in a group and some others can be changing or be overridden at sample level. Can ISOBMFF provide some guidance on how to combine entity group, track group and sample group (i.e. how they relate to each other) to declare such parameters?

The following paragraph provides guidance for the combination of entity group with sample group and could become part of a new subsection in Annex B ("guidance on deriving from this document"):

It is possible to specify an entity group and a sample group that use the same four-character code as their type and mutually contribute to indicate static and dynamic information for a specific purpose. Such a pairing of an entity group and a sample group is a documented characteristic of a specific four-character code. Common static information should be contained in an entity group with a particular group\_id value, and dynamic information should be contained in a sample group of the same type and with the value of grouping\_type\_parameter equal to the value of group\_id.

Similar guidance could be provided for combining a track group with a sample group when a "grouping" feature for a track requires simultaneously some parameters to be static and some other parameters to be changing over groups of samples (as already done for '2dsr' in OMAF for instance).

# Stronger defaulting in Track Runs

## Discussion

Many file writers operate by parsing the high-level syntax of a given input video bitstream and generating the file format metadata from the information of the bitstream. A backward-compatible approach could be achieved by letting advanced clients

* omit the downloading of MovieFragmentBoxes, and
* create the MovieFragmentBoxes in the client side by parsing the high-level syntax of the received MediaDataBoxes.

At the same time, the transmitted track run data is reduced to 0 bytes, i.e. ultimate compression of MovieFragmentBox metadata is achieved.

## Overview

A set of indications is proposed based on which a player can conclude that it is able to process the media data without the MovieFragmentBox. Thus, even though the MovieFragmentBox is available, the player does not need to fetch and process it. Hence, this is a backward compatible approach for avoiding the delivery of MovieFragmentBox.

The following indications are proposed:

* Either of the following indications is included in the ISOBMFF:
  + Indicate with a box flag in a data reference box that all MediaDataBoxes that are referenced through the data reference entry are "tigthly packed", i.e. contain samples for a single track only in decoding order without unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format. See Section 7.4.
  + Indicate with a box flag of the SegmentIndexBox that the media data box(es) carrying the data for the described segment are "tightly packed". See Section 7.6.
* The following indication is included in the NAL unit file format (ISO/IEC 14496-15):
  + Indicate with a box in the sample entry how picture timing can be derived. The box includes a multiplication factor that applies to convert picture order count differences to composition times. For a picture that starts a new coded video sequence, the box indicates the composition time difference from the last picture of the previous coded video sequence. See Section 7.5.
* For usage with DASH, either of the following approaches can be used:
  + Extensions of the SegmentIndexBox indicate the referenced metadata size (mainly MovieFragmentBoxes) or offset(s) to the media data and can be used to conclude the byte ranges of MediaDataBoxes. See alternative approaches in Sections 7.6 and 7.7.
  + The following indications are included in the DASH MPD: Indicate the byte ranges or URLs for requesting the MediaDataBoxes only with a new MPD attribute. See Section 7.8.

A player can operate as follows:

* Conclude from the flag indicating "tightly packed" media data boxes that reception of MovieFragmentBoxes is not necessary.
* Use the indicated byte ranges or URLs for requesting MediaDataBoxes only.
* Generate the MovieFragmentBoxes based on parsing the high-level syntax of the bitstream in the received MediaDataBoxes or directly decodes and plays the bitstream without deriving the file format metadata. In this operation, the information of the TrackRunBox is concluded based on the received MediaDataBox for a movie fragment as follows:
  + By carrying out the access unit boundary determination as specified in AVC or HEVC, the sample sizes and the sample count can be derived.
  + Picture composition timing may be provided in the bitstream (picture timing SEI message) or it is concluded that composition times are proportional to picture order count. Values of sample\_duration are derived accordingly.
  + VCL NAL unit type can be used to determine sample flags or the sample flags can be set to indicate an unknown status.

## Analysis

### Responses to comments at MPEG#127

In the following, we copy the comments from the File Format minutes of MPEG#127 (labeled Cx) and provide our answers to them (Ax).

C1. It seems that in the case that the URLs (e.g. in an MPD) resolve to "pure media data" one would need a new MIME type (not the one for an ISO segment). In that case, is it really in scope for the file format? (See bullet below).

A1. The scheme is primarily intended for on-demand streaming (e.g. ISO base media file format on-demand profile of DASH) where all SegmentIndexBox(es) are placed before any MovieFragmentBox. No changes in the segment formats are proposed and hence no new MIME types are needed either. The media data is selectively fetched using HTTP GET requests with byte ranges that are concluded from the SegmentIndexBox(es) with extensions specified in the proposed scheme.

C2. The 'pain' is not the file format overhead, but the implementation complexity (and edge cases). We need to evaluate this.

A2. We implemented options for providing media data offsets in MP4Box and tested the reader compatibility of the options with several readers. See Section 7.3.2 for details.

C3. This relies on getting somehow getting a segment index (either in-band, e.g. after the moov box, or out of band) at the client:

* because the representation is ftyp-moov-mdat-mdat-mdat…
* because the representation is ftyp-moov-moof+mdat-moof+mdat-moof+mdat… and the sidx tells you the byte-range requests in each segment to omit the moof box

A3. Right, we assume that sidx(es) are placed before any moof, which is required e.g. in ISO base media file format on-demand profile of DASH. Thus, the file structure would be ftyp-moov-sidx(es)-moof+mdat-moof+dat-moof+dat…

### Tests on reader compatibility on extended SegmentIndexBox and SegmentMediaOffsetBox

This section provides results of the reader compatibility tests of the options in the TuC.

Tests were carried out by segmenting a video clip with one AVC media track by differently modified MP4Box programs. Files were then served to players from an HTTP server.

Three different MP4Box versions were used:

* Unmodified MP4Box
* Modified MP4Box which adds SegmentMediaOffsetBox ('smof') after SegmentIndexBox
* Modified MP4Box which sets flags to 1 in SegmentIndexBox and adds media\_data\_offset fields to the end of the SegmentIndexBox.

As seen in Table 1, extended SegmentIndexBox did not introduce any degraded functionality in any of tested players, compared to the unmodified input files.

In browsers, Dash.js reference client relies on browser-side Media Source Extensions (MSE) to handle segment data parsing. Both Firefox and Chromium browsers use MSE implementations which stop segment processing with an error, if they encounter an unknown root-level box. This prevents playback of files which contain SegmentMediaOffsetBox.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Media player | Unmodified | Added 'smof' box | Extended 'sidx' box | Notes |
| Vlc (3.0.8) | OK | OK | OK |  |
| MP4Client (GPAC 0.8.0) | OK | OK | OK | Log messages about extra bytes in 'sidx' box. |
| ffplay (ffmpeg 4.1.3) | OK | OK | OK |  |
| Android Exoplayer (2.10.4) | OK | OK | OK |  |
| Dash.js reference client (3.0.0) running in Firefox browser (69.0.1) | OK | Not playing. Error because of unknown root-level box. | OK |  |
| Dash.js reference client (3.0.0) running in Chromium browser (76.0.3809.100) | OK | Not playing. Error because of unknown root-level box. | OK |  |
| Dash.js reference client (3.0.0) running in Safari browser (13.0.1) | OK | OK | OK |  |

Table 1: Playback test results

## Flag in data reference box

The following is proposed to be added into clause 8.7.2.1 of ISOBMFF:

When (flags & 0x000002) is greater than 0 in a data reference box, all MediaDataBoxes that are referenced through the data reference box are "tightly packed", i.e.:

* + MediaDataBoxes contain samples for a single track only.
  + The samples are in decoding order within a MediaDataBox.
  + MediaDataBoxes contain no unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format.

## Box in 14496-15

*Add the following clause 4.13:*

**4.13 Indicating composition times directly proportional to picture order counts**

**4.13.1 Definition**

Box Type: 'reti'  
Container: Sample Entry   
Mandatory: No  
Quantity: Zero or one

The presence of this box specifies that the composition time difference of any two consecutive pictures in output order in the same coded video sequence is directionally proportional their picture order count difference.

**4.13.2 Syntax**

class RelativeTimingBox extends FullBox ('reti', version=0, flags=0) {  
 unsigned int(32) poc\_unit\_duration;  
 unsigned int(32) cvs\_start\_interval;  
}

**4.13.3 Semantics**

poc\_unit\_duration specifies the composition time difference that corresponds to a picture order count difference equal to 1.

cvs\_start\_interval specifies the composition time difference of the first picture of each coded video sequence, in output order, relative to the last picture of the previous coded video sequence, in output order.

NOTE: In practice, cvs\_start\_interval is the sample duration of the last picture of each coded video sequence.

## Extension of the segment index box

### Overview

Figure 1 below illustrates new versions of the extended segment index box ‘sidx’. In these new versions of the segment index box, when indexing fragments (i.e. reference\_type=0), two indexes can be stored per fragment (instead of a single one currently: referenced\_size).



Figure : New version of ‘sidx’

As illustrated in the new ‘sidx’ syntax below, the first index is associated with the actual data of the considered fragment while the second index is associated with the metadata of this fragment.

*Add the following at the end of clause 8.16.3.1 (definition of SegmentIndexBox):*

The flags field has the following semantics:

(flags & 1) equal to 1 specifies that the referenced segments are constrained as follows:

* + There is a single referenced MediaDataBox or IdentifiedMediaDataBox for references with reference\_type equal to 0.
  + The referenced MediaDataBox or IdentifiedMediaDataBox contains samples for a single track only.
  + The samples are in decoding order within the referenced MediaDataBox or IdentifiedMediaDataBox.
  + The referenced MediaDataBox or IdentifiedMediaDataBox contains no unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format.

NOTE 1: Since encryption requires extra data to be stored with samples, it is not possible to use (flags & 1) equal to 1 with encrypted media data.

NOTE 2: When (flags & 1) is equal to 1, and media samples are either self-framing or of constant size (indicated by flags & 4), and the sample times are either predictable (indicated by flags & 2) or calculable, it is possible to process the referenced MediaDataBox or IdentifiedMediaDataBox without the MovieFragmentBox. Clause J.2.5 provides background and rationale for using flags and contains an example of a file structure.

(flags & 2) equal to 2 specifies that the referenced segments are constrained so that default\_sample\_duration of TrackExtendsBox applies to each sample and that sample\_composition\_time\_offset is equal to 0 for each sample.

(flags & 4) equal to 4 specifies that the referenced segments are constrained so that default\_sample\_size of TrackExtendsBox applies to each sample.

*Change the syntax of the* SegmentIndexBox *in clause 8.16.3.2 to the following (i.e., replacing* 0 *with* flags *in the box header, and adding the parts conditioned by the value of* flags*):*

aligned(8) class SegmentIndexBox extends FullBox('sidx', version, flags) {  
 unsigned int(32) reference\_ID;  
 unsigned int(32) timescale;  
 if (version==0) {  
 unsigned int(32) earliest\_presentation\_time;  
 unsigned int(32) first\_offset;  
 }  
 else {  
 unsigned int(64) earliest\_presentation\_time;  
 unsigned int(64) first\_offset;  
 }  
 unsigned int(16) reserved = 0;  
 unsigned int(16) reference\_count;  
 for(i=1; i <= reference\_count; i++)  
 {  
 bit (1) reference\_type;  
 unsigned int(31) referenced\_size;  
 unsigned int(32) subsegment\_duration;  
 bit(1) starts\_with\_SAP;  
 unsigned int(3) SAP\_type;  
 unsigned int(28) SAP\_delta\_time;  
 }  
 if (flags & 1)  
 for(i=1; i <= reference\_count; i++)  
 if (reference\_type == 0) // reference\_type of the same i value  
 unsigned int(32) media\_data\_offset;  
}

*Add the following to the end of clause 8.16.3.3:*

media\_data\_offset specifies the offset to the start of the referenced MediaDataBox or the IdentifiedMediaDataBox of a subsegment from the start of the subsegment.

## Alternative improvements to the SegmentIndexBox

### Option 1

To support the different indexing modes, the semantics of reference\_type is extended as follows (highlighted in yellow):

* when set to 1 indicates that the reference is to a SegmentIndexBox; otherwise the reference is to media content as follows:
* when set to 0 indicates content including both metadata and media data (e.g., in the case of files based on this document, to a MovieFragmentBox);
* when set to 2 indicates content including metadata only (e.g., in the case of files based on this document, one or more MovieFragmentBox);
* when set to 3 indicates content including media data only (e.g., in the case of files based on this document, one or more MediaDataBox or IdentifiedMediaDataBox);

if a separate index segment is used, then entries with reference type 1 or 2 are in the index segment, and entries with reference type 0 or 3 are in the media file;

In this option a new version of the segment index box requires two bits for the representation of the reference\_type as illustrated below. The referenced\_size field in the new version is interpreted according to the following values of the reference\_type:

* When set to 0, the referenced\_size is the distance in bytes from the first byte of the referenced index to the first byte of the next referenced index (moof) item.
* When set to 1, the referenced\_size is the distance in bytes from the first byte of the referenced index to the first byte of the next referenced index (sidx) item.
* When set to 2, referenced\_size is the distance in bytes from the first byte of the referenced metadata item to the first byte of the next referenced index metadata item, or in the case of the last entry, the end of the referenced index metadata item.
* When set to 3, referenced\_size is the distance in bytes from the first byte of the referenced data item to the first byte of the next referenced index data item, or in the case of the last entry, the end of the referenced index data item .

The value of subsegment\_duration of each entry with reference\_type equal to 2 or 3 corresponds to the duration of the indexed sub-segment. When the reference\_type is set to 1, the semantics of the subsegment\_duration is the same as in ISOBMFF Table J.3.

### Option 2

#### Syntax

aligned(8) class SegmentIndexBox extends FullBox('sidx', version, flags) {  
 unsigned int(32) reference\_ID;  
 unsigned int(32) timescale;  
 if (version==0 || new\_version) {  
 unsigned int(32) earliest\_presentation\_time;  
 unsigned int(32) first\_offset;  
 } else { //version =1 || new\_version  
 unsigned int(64) earliest\_presentation\_time;  
 unsigned int(64) first\_offset;  
 }  
 unsigned int(16) reserved = 0;  
 unsigned int(16) reference\_count;  
 if (new version)  
 unsigned int(16) subpart\_count;  
 for(i=1; i <= reference\_count; i++) {  
 bit (1) reference\_type;   
 unsigned int(31) referenced\_size;  
 if (new\_version) {  
 for (j=1; j<=subpart\_count;j++) {  
 unsigned int(32) data\_reference\_offset;   
 // may be controlled by a flags value   
 unsigned int(32) referenced\_data\_size;   
 }  
 }  
 unsigned int(32) subsegment\_duration;   
 bit(1) starts\_with\_SAP;  
 unsigned int(3) SAP\_type;  
 unsigned int(28) SAP\_delta\_time;  
 }  
}

#### Semantics (for the new fields, the other remaining unchanged)

Data\_reference\_offset indicates in bytes from where, in a file or in a segment file, the indexed data start. The offset corresponds to the first byte of the file or to the first byte of the considered segment file.

referenced\_data\_size indicates a size in bytes for a contiguous byte range in the data part of the referenced fragment

subpart\_count indicates the number of data blocks (for example tiles) for the current subsegment.

This extended '*sidx*' box can also be combined with '*sidx*' boxes of the current version, for example as in the hierarchical or daisy-chain schemes defined in ISO/IEC 14496-12

## Attributes in DASH MPD

The following is proposed to be added into the SegmentBase element (clause 5.3.9.2.2 of DASH).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | @mediaOnlyRange | O | specifies the byte range that consists only of the media data, such as MediaDataBox, applicable to all Media Segments of the Representation. When used with ISOBMFF Media Segments, the indicated byte range shall start with a box.  The byte range shall be expressed and formatted as a byte-range-spec as defined in RFC 7233, Clause 2.1.. It is restricted to a single expression identifying a contiguous range of bytes. |

The following is proposed to be added into the SegmentList.SegmentURL element (clause 5.3.9.3.2 of DASH).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | @mediaOnlyRange | O | specifies the byte range within the resource identified by the @media that consists only of the media data, such as MediaDataBox. When used with ISOBMFF Media Segments, the indicated byte range shall start with a box.  The byte range shall be expressed and formatted as a byte-range-spec as defined in RFC 7233, Clause 2.1.. It is restricted to a single expression identifying a contiguous range of bytes. |

The following is proposed to be added into the SegmentTemplate element (clause 5.4.9.4.2 of DASH).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | @mediaOnly | O | specifies the template to create the Media Segment List where Media Segments only consist of the media data, such as MediaDataBoxes. |

## Example usage of the segment index box

*Add the following clause J.2.5:*

**J.2.5 Simple one-level indexing of "tightly packed" media**

When the flags field of the SegmentIndexBox is set so that (flags & 1) is equal to 1, the media data is "tigthly packed", i.e. a single MediaDataBox or IdentifiedMediaDataBox contain samples for a single track only in decoding order without unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format. A legacy client simply omits the flags field and the SegmentIndexBox syntax conditional on the values of the flags field. A client taking advantage of the "tightly packed" media could operate as follows:

* Conclude that since the SegmentIndexBox has (flags & 1) equal to 1, the reception of MovieFragmentBoxes is not necessary.
* Omit the downloading of MovieFragmentBoxes and only download the media data by deriving a byte range from the media\_data\_offset given in the SegmentIndexBox. Thus, a bitrate saving in the transmitted data is achieved.
* Create the MovieFragmentBoxes in the client side either by parsing the high-level syntax of the received media data or, when (flags & 2) and (flags & 4) are set, use the default values given in of TrackExtendsBox. For example, in case of AVC or HEVC, the information in the TrackRunBox could be concluded as follows:
  + Deriving the sample sizes and the sample size by carrying out the access unit boundary determination as specified in AVC or HEVC.
  + Deriving sample composition times from picture timing SEI messages present in the bitstream or concluding that composition times are proportional to picture order count.
  + Deriving sample flags from the VCL NAL unit types or setting sample flags to indicate an unknown status.

Figure J.2 shows an example that is aligned with the structure presented in Figure J.1. All entries of the top level SegmentIndexBox point to segments comprising one or more movie fragments, i.e. reference\_type is equal to 0. The values of referenced\_size and subsegment\_duration of each entry are calculated as in Table J.1 above. The dashed double-ended arrows in Figure J.2 indicate the values of media\_data\_offset.



Figure J.2. Simple segment index including media\_data\_offset values (dashed arrows).

# Sample reordering in Track Runs

## Introduction

Samples in TRUN are contiguous in byte range and stored in decoding order. As discussed previously, we identified use cases where changing the sample ordering could be beneficial:

- Usage of partially received segments (HTTP streaming, ISOBMFF-based broadcast)

- More efficient unequal FEC protection of the segment, with moof and base sublayer in a single, more protected zone

In these use cases, the movie fragment is incomplete and some samples are not available. With the current design, it is very likely that samples from a temporal sublayer are received while depended-on samples of lower layers are not. This implies that additionally to the lost samples, we end up with unusable yet completely received samples. By ordering samples according to their temporal sublayers, the usage of partially received segments can be improved.

As discussed in m44768, there are several options to do this:

- Option1: split the temporal layers as one per track

This is straightforward and works for codecs having a multi-track ISOBMFF encapsulation defined (yet not all of them do). However, this is quite costly in terms of ISOBMFF structures, each sublayer requiring a dedicated track hence duplication of traf, tfhd and trun boxes for each track. The complexity is also high since samples have to be properly re-interleaved when "merging" these tracks. This also requires multi-track segments which is not very friendly in HAS delivery or ISOBMFF-based broadcast. It finally requires splitting sample grouping and CENC information per track, which is both complex and costly (additional sample to group boxes, subsamples boxes, sample group descriptions if not in moov, saio/saiz/senc boxes etc ...)

- Option 2: split the temporal layers as one per TRUN

Since samples have to be in decoding order within the fragment, the only possibility to store data per temporal sublayer is to have a new trun whenever a we have a change of temporal sublayer between non contiguous samples:

*Example:*

For a classical I0P0B1B2B3 layout at 25 fps, with BN having no dependencies on BN+1 a potential 1s GOP structure could be:

I10P250 B51 B32 B23 B43 B91 B72 B63 B83B131 B112 B103 B123 B171 B152 B143 B163 B211 B192 B183 B203 B232 B223 B243

Our expected layout would be:

I10P250 B51 B91 B131 B171 B211 B32 B72 B112 B152 B192 B232 B23 B43 B63 B83B103 B123 B143 B163 B183 B203 B223 B243

To achieve this with the current ‘trun’ design, we would need:

*TRUN* I10P250 B51 *TRUN* B32 *TRUN* B23 B43 *TRUN* B91 *TRUN* B72 *TRUN* B63 B83*TRUN* B131 *TRUN* B112 *TRUN* B103 B123 *TRUN* B171 *TRUN* B152 *TRUN* B143 B163 *TRUN* B211 *TRUN* B192 *TRUN* B183 B203 *TRUN* B232 *TRUN* B223 B243

Hence 17 trun instead of 1! With a base TRUN size of 20 (12 for full box + 8 for sample count+data offset) or 16 for ctrn (12 for full box + 8 for sample count+data offset assuming offset can be less than 65k), we end up with at least 320 bytes (trun) or 256 bytes (ctrn) of overhead.

This gets even worse if we start increasing the GOP size or the frame rate.

The proposal in m44768 to overcome this was to use a single trun with a sample layout index allowing custom sample layouts (more details below).

## Further discussion

### Context

We investigated how to reuse an existing ISOBMFF HAS packaging (single file or segmented) to provide a low frame rate version of the content without duplicating the files. Our ultimate goal is to have as few byte range requests to issue as possible for a given media segment, to reduce complexity.

The level assignment box ‘leva’ seemed to be designed for this, as illustrated in DASH 4th edition (Figure 6):

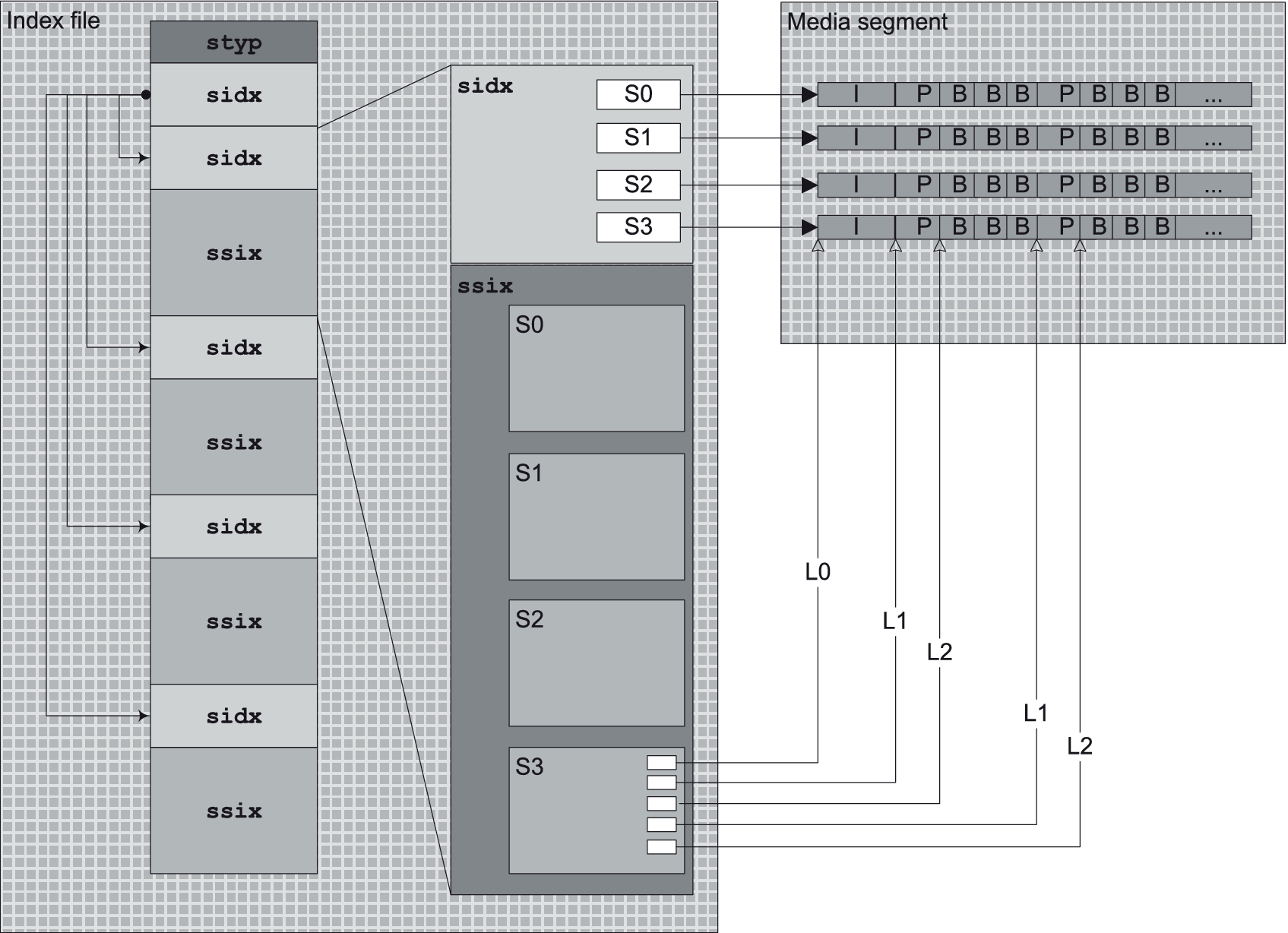


Figure : Mapping temporal sub-layers to Sub-Segment Index Box for trick mode in DASH

This box describes the assignment of one level per temporal sublayer and each level is further described in terms of byte range in an ‘ssix’ box.

### leva and ssix clarifications

See the Defect Report for 14496-12.

### Sample reordering using leva and ssix

Assuming our previous interpretation is correct, if we want to have samples organized per temporal dependencies, we have two possibilities:

1. Separate each sublayer in a dedicated traf, which we would want to avoid as we explained in section 2.
2. When a single track is used, the common (if not only) usage is to map the moof and first IDR (possibly the immediately following P frame if no leading pictures) in the first level, and the rest to a second level. Trying to map all P frames in the segment to the same level will not be possible (disjoint byte ranges), unless using the above trick (section 2) of splitting the ‘trun’ into multiple ‘trun’ to reorder the samples, with the size increase it induces.

One possibility would be to relax the leva/ssix constraint on byte range continuity per level, and allow multiple occurrences of a level in an ‘ssix’ box (see discussion in previous section). While this works, this has the following drawbacks:

- the ssix box becomes quite big: for our previous example, 17 entries instead of 4 (one per sublayer), each entry counting for 32 bits

- it seems to break the philosophy of ssix

- multiple byte ranges will be required for a given level

- samples are still in decoding order in the ‘mdat’ (not compatible with the identified use cases above)

We therefore propose to introduce a sample ordering index at the ‘trun’ level to enable:

- Single trun, single track

- Single byte-range request for a given sublayer (or level)

- Unmodified 'ssix' / 'leva', except specification clarifications as discussed in previous section

- Reusability of partially downloaded segments

- Friendliness for unequal FEC protection schemes

## Proposal

The proposed syntax below could be added either directly in the 'trun' or 'ctrn', or as a companion box (with mandatory processing) of the 'trun'. We propose here our preferred solution which is an extension of 'ctrn'.

### Sample interleave index in (Compact) Trun

The initial proposal in m44768 proposed to use the trun or compact trun currently under investigation to provide sample interleaving (or reordering) information. The principles of trun are kept (still describes contiguous set of samples in decoding order in file), only the locations of the samples in the mdat within the run change. This make sure that one reader fetching one trun has still all the data for these samples.

As noted in m44768, using a data offset per sample to provide the sample\_interleave\_index would not be very efficient since we already have the ‘trun’ base offset and each sample size. We therefore need to indicate the sample\_interleave\_index in the track run in a compact way.

The initial proposal from m44768 cost one index per sample, the sample\_interleave\_index using the same coding trick (1, 2 or 4 bytes) as the other fields in ‘ctrn’. We propose to deduce the number of bits to use for the sample\_interleave\_index from the one for sample count, since the sample\_interleave\_index shall be given for each sample.

We then need:

* 1 bit flags to indicate presence/absence of reordering/interleaving index
* 1 bit flags for the sample\_count\_index\_size

We propose to add the following flags value in ‘ctrn’ flags (see m50571 on ctrn tests):

0xTO\_BE\_DEFINED sample\_interleave\_bit: when set, indicates the samples in the trun may be in an order different from the decoding order, and that a map of sample index in decoding order will be given at the end of the trun.

Add in ctrn syntax section:

unsigned int(8) function indexToBitSize(sample\_count) {  
 if (sample\_count<256) return 8;  
 else return 16;  
}

Add at the end of ctrn box syntax:

if (tr\_flags & sample\_interleave\_bit) {  
 unsigned int(indexToBitSize(sample\_count))   
 sample\_interleave\_index [ sample\_count];  
}

Add to semantics:

sample\_interleave\_index: indicates the order of sample interleaving in the trun. A value of 0 indicates that the sample data start at the trun data offset. A value of K>0 indicates that the sample data start at the trun data offset plus the sum of the size of all samples with an interleaving index strictly less than K. The index shall range between 0 and sample\_count-1 inclusive. There shall not be two samples with the same interleaving index in the same trun.

We also propose to clarify the content of a compact track run in its definition section:

“A track run documents a contiguous set of samples in decoding order for a track. However, the actual storage of samples in mdat within the track run may be interleaved according to an optional index sample\_interleave\_index”

### Sample count in Compact Trun

We note that sample\_count is hardcoded to 16 bits in ctrn, but it is quite common for fragments to have less than 256 samples. Was this made on purpose?

We suggest using one remaining flag in ctrn to signal the size of the sample\_count field (see m50571 on ctrn tests results), as follows:

In Section 8.8.8.2.2 of AMD4, add the following:

0xTO\_BE\_DEFINED sample\_count16bits that when set indicates that sample\_count is coded on 16 bits. When not set, sample\_count is coded on 8 bits.

and in Section 8.8.8.2.2 of AMD4, replace in the syntax for CompactTrackRunBox:

unsigned int(16) sample\_count;

with

if (tr\_flags & sample\_count16bits)  
 unsigned int(16) sample\_count;  
 else  
 unsigned int(8) sample\_count;

# Segment Index and Level Assignment

Issues: [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/123*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/123) *and* [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/42*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/42)

See also [*https://github.com/MPEGGroup/FileFormat/issues/12*](https://github.com/MPEGGroup/FileFormat/issues/12)

## Discussion

*https://github.com/MPEGGroup/FileFormat/issues/12*

The proposal is based on the following observations:

1. signaling of IDR byte-range is very similar to the range concepts of ‘ssix’
2. avoid modifying ‘sidx’ box, both for backward compatibility issues (comment from TuC, section 15) and because we index subsegments
3. signaling multiple byte-ranges for the same level in ‘ssix’ seems reasonable (for example, two IDRs in a subsegment)
4. usage of ‘ssix’ with ‘leva’ is not always desirable, especially since:

* level assignment may be dependent on sample group description and sample to group mapping, which is not always available (‘moof’ not yet fetched).
* ‘leva’ requires level to be present in increasing order in the ‘mdat’, which does not allow multiple byte-ranges for a given level.
* ‘leva’ cannot be updated on the fly (present in 'mvex'), all levels to describe have to be known at the start of the session

We therefore would like to introduce a new design of ‘ssix’, fixing the above shortcomings. Moreover, while redesigning ‘ssix’, we also considered the use case of spatial indexing for tile tracks in a file.

## Proposal

The proposal defines a way to use multiple byte ranges per level in ‘ssix’/’leva’, and multiple byte ranges with ‘ssix’ without ‘leva’ through predefined level assignments.

Yellow-highlighted corresponds to text (Part-12) move

Green comes from above proposal

Blue are changes as proposed in TuC.

*In 8.8.13.1 replace*

"Within a fraction, data for each level shall appear contiguously. Data for levels within a fraction shall appear in increasing order of level value. All data in a fraction shall be assigned to levels. "

*with*

"When version 0 of the LevelAssignmentBox is used, within a fraction, data for each level shall appear contiguously, and data for levels shall appear in increasing order of level value. All data in a fraction shall be assigned to levels.

When version 1 or more of the LevelAssignmentBox is used, data for each level need not be stored contiguously and data for levels may be stored in random order of level value. Some data in a fraction may have no level assigned, in which case the level is unknow but is not a level from the levels defined by the LevelAssignmentBox."

*In 8.8.13.1 remove*

“

When padding\_flag is equal to 1 this indicates that a conforming fraction can be formed by concatenating any positive integer number of levels within a fraction and padding the last MediaDataBox by zero bytes up to the full size that is indicated in the header of the last MediaDataBox. The use of padding\_flag is deprecated.

”

*In 8.8.13.2 replace*

aligned(8) class LevelAssignmentBox extends FullBox('leva', 0, 0)

*with*

aligned(8) class LevelAssignmentBox extends FullBox('leva', version, 0)

*In 8.8.13.3 replace*

“padding\_flag equal to 1 indicates that a conforming fraction can be formed by concatenating any positive integer number of levels within a fraction and padding the last MediaDataBox by zero bytes up to the full size that is indicated in the header of the last MediaDataBox. When

padding\_flag is equal to 0 this is not assured.

”

with

“

padding\_flag deprecated, should be set to 0.

”

*Replace 8.16.4.1 with*

“

The SubsegmentIndexBox provides a mapping from levels (as specified by the LevelAssignmentBox or as indicated in the box itself) to byte ranges of the indexed subsegment. In other words, this box provides a compact index for how the data in a subsegment is ordered according to levels into partial subsegments. It enables a client to easily access data for partial subsegments by downloading ranges of data in the subsegment.

~~Each byte in the subsegment shall be explicitly assigned to a level, and hence the range count shall be 2 or greater.~~ ~~If the range is not associated with any information in the level assignment, then any level that is not included in the level assignment may be used.~~

There shall be 0 or 1 SubsegmentIndexBoxes per each SegmentIndexBox that indexes only leaf subsegments, i.e. that only indexes subsegments but no segment indexes. A SubsegmentIndexBox, if any, shall be the next box after the associated SegmentIndexBox. A SubsegmentIndexBox documents the subsegments that are indicated in the immediately preceding SegmentIndexBox.

In general, the media data constructed from the byte ranges is incomplete, i.e. it does not conform to the media format of the entire subsegment.

For leaf subsegments based on this document (i.e. based on movie sample tables and movie fragments):

* ~~Each level shall be assigned to exactly one partial subsegment, i.e. byte ranges for one level shall be contiguous.~~
* ~~Levels of partial subsegments shall be assigned by increasing numbers within a subsegment, i.e., samples of a partial subsegment may depend on any samples of preceding partial subsegments in the same subsegment, but not the other way around. For example, each partial subsegment contains samples having an identical temporal level and partial subsegments appear in increasing temporal level order within the subsegment.~~
* For version 0 of the box, each level shall be assigned to exactly one partial subsegment and in increasing order of level value, i.e. byte ranges for one level shall be contiguous and samples of a partial subsegment may depend on any samples of preceding partial subsegments in the same subsegment, but not the other way around. This implies that all data for a given level require a single range to be retrieved.
* For version 1 or more of the box, multiple byte ranges, possibly discontinuous, with the same level may be described. This implies that all data for a given level may require multiple byte ranges to be retrieved.

//editor's note: the next notes correspond to the previously existing last 2 bullets of the spec but they are informative or repetitions from leva.

Note 1: When a partial subsegment is accessed in this way, for any assignment\_type other than 3, the final MediaDataBox may be incomplete, that is, less data is accessed than the length indication of the MediaDataBox indicates is present. The length of the MediaDataBox may need adjusting, or padding used. ~~The padding\_flag in the LevelAssignmentBox indicates whether this missing data can be replaced by zeros. If not, the sample data for samples assigned to levels that are not accessed is not present, and care should be taken not to attempt to process such samples.~~

Note 2: The data ranges corresponding to partial subsegments include both MovieFragmentBoxes and MediaDataBoxes. The first partial subsegment, i.e. the lowest level, will correspond to a MovieFragmentBox as well as (parts of) MediaDataBox(es), whereas subsequent partial subsegments (higher levels) may correspond to (parts of) MediaDataBox(es) only.

For version 0 of this box, the presence of the LevelAssignmentBox in the movie is mandatory, and the LevelAssignmentBox shall have a version equal to 0.

*Editor's note:* the current v0 spec is unclear, it does not explicitly mandate leva with ssix, maybe we should keep this.

Note 3: assignment\_type equal to 0 (specified in the LevelAssignmentBox) can be used, for example, together with the temporal level sample grouping ('tele') when frames of a video bitstream are temporally ordered within subsegments; assignment\_type equal to 2 can be used, for example, when each view of a multiview video bitstream is contained in a separate track and the track fragments for all the views are contained in a single movie fragment. assignment\_type equal to 3 can be used, for example, when audio and video movie fragments (including the respective MediaDataBoxes) are interleaved. The first level can be specified to contain the audio movie fragments (including the respective MediaDataBoxes), whereas the second level can be specified to contain both audio and video movie fragments (including all MediaDataBoxes).

For version 1 of this box, the presence of the LevelAssignmentBox is only mandatory for level\_assignment\_type 0, in which case the LevelAssignmentBox shall have a version of 1.

“

*Replace 8.16.4.2 with*

1. aligned(8) class SubsegmentIndexBox extends FullBox('ssix', version, flags) {  
   if (version==0) {  
    unsigned int(32)subsegment\_count;  
    for( i=1; i <= subsegment\_count; i++) {  
    unsigned int(32)range\_count;  
    for ( j=1; j <= range\_count; j++) {  
    unsigned int(8) level;   
    unsigned int(24) range\_size;   
    }  
    }  
   } else {  
    unsigned int(16)subsegment\_count;  
    unsigned int(1)lsc;   
    unsigned int(1)incomplete;   
    unsigned int(2)lbs;   
    unsigned int(2)rbs;   
    unsigned int(2)reserved;   
    unsigned int(8)level\_assignment\_type;  
    for( i=1; i <= subsegment\_count; i++) {  
    unsigned int(lsc ? 32 : 16)range\_count;  
    for ( j=1; j <= range\_count; j++) {  
    unsigned int((lbs+1)\*8) level;   
    unsigned int((rbs+1)\*8) range\_size;   
    }  
    }  
   }  
   }

*Replace 8.16.4.3 with*

subsegment\_count is a positive integer specifying the number of subsegments for which partial subsegment information is specified in this box. subsegment\_count shall be equal to reference\_count (i.e., the number of movie fragment references) in the immediately preceding SegmentIndexBox.

lsc if set, indicates that the number of indexed ranges within a partial subsegment is coded on 32 bits, otherwise the number of indexed ranges within a partial subsegment is coded on 16 bits.

incomplete if set, indicates that the last range of a given subsegment may ~~not cover the entire~~ end before the last byte of the subsegment, in which case assignment of remaining bytes to level is unknown but the remaining bytes should not correspond to any level listed in the box.

lbs gives the number of bytes, minus 1, used in coding the level field.

rbs gives the number of bytes, minus 1, used in coding the range field.

level\_assignment\_type gives the associated semantics of the indicated level.

* 0: the level value corresponds to the level indicated in the leva box. If the range is not associated with any information in the level assignment, then any level that is not included in the level assignment may be used. This value shall only be used when the leva box version is 1 or more.
* 1: the level value corresponds to a dependency level.
  + Level 0 indicates that the byte range contains:
    - exactly one or more file-level boxes (e.g. MovieFragmentBox) other than a media data container box (e.g. MediaDataBox or IdentifiedMediaDataBox), and/or
    - zero or at most one box header (8 or 16 bytes) of a media data container box which shall correspond to the last 8 or 16 bytes of the byte range
  + Level 1 indicates same type of data as level 0 but having a dependency on the previous preceding byte range with level L0 (e.g. one single box header (8 or 16 bytes) of a media data container box, the media data container box containing data described by the preceding MovieFragmentBox)
  + Level 2 means the data is independently decodable (SAP 1, 2 or 3). Byte range with level 1 immediately preceding if specified and the first preceding byte range with level 0 are required to process the data.
  + Level N, with N>2, requires data from the preceding byte ranges with lower levels (level N-1 and below) to be processed, stopping at the previous preceding byte range with level 2 if specified, otherwise at the first byte range in the box. Each first byte range with level 0 or 1 preceding any required byte range (level 2 to N) is required to process the data.
* 2: the level value corresponds to a multitrack dependency level. In this mode, lbs shall be 1 or more (i.e., at least 16 bits to code the level). The first 8 bits of the level field give the dependency level value, with the same values and semantics as level\_assignment\_type 1. The remaining less significant bits of the level field give a trackID, which shall identify a track present in the indexed subsegment for level values other than 0 and 1, and shall be 0 if level value is 0 or 1. In this mode, each range with level N>1 consists only of data from the identified track~~, possibly with some meta-data boxes (movie fragments, etc...). The~~ and the level value only gives dependency information within the track. This allows cross-track indexation within a same level.
* 3->0xFF: ISO reserved

range\_count specifies the number of partial subsegment levels into which the media data is grouped. For version 0 of the box, this value shall be greater than or equal to 2 and each byte in the subsegment shall be explicitly assigned to a level. For version 1 or more of this box, this value may be 0 or more, and the described ranges may lead to a size smaller than the subsegment if and only if incomplete is set to 1.

range\_size indicates the size of the partial subsegment. This value shall not be 0, except for the last entry for which the value 0 may be used ~~in the last entry~~ to indicate the remaining bytes of the segment, to the end of the segment.

level specifies the level to which this partial subsegment is assigned.

For level\_assignment\_type 1 or above, byte ranges assigned to levels other than 0 or 1 shall not contain file-level box headers. Typically, the header of a media data container box (e.g. MediaDataBox or IdentifiedMediaDataBox) is in level 0 or 1 while data may be in level 2 or more.

NOTE For level\_assignment\_type 1, since level N depends only from level N-1 and below, a direct mapping of temporal sublayers to levels will not always be possible in case frames from one temporal sublayer depend on preceding frames from the same temporal sublayer in another byte range.

## Examples of use

### Single track indexing

A picture containing shape

Description automatically generated

Figure - level\_assignment\_type=1 with 2 IDRs and no explicit range for moof

A picture containing shape

Description automatically generated

Figure - level\_assignment\_type=1 with 2 IDRs and explicit range for moof

A picture containing text

Description automatically generated

Figure - level\_assignment\_type=1 with low latency DASH segment, 2 chunks and no explicit range for 'moof

### Multi-track indexing

A picture containing text

Description automatically generated

Figure - level\_assignment\_type=2, 2 moof+traf with 1 IDR each and no explicit range for moof

A picture containing shape, rectangle

Description automatically generated

Figure - level\_assignment\_type=2, 1 moof with 2 trafs, 1 IDR/traf and no explicit range for moof

A picture containing rectangle

Description automatically generated

Figure - level\_assignment\_type=2, 1 moof with 2 trafs, 1 IDR/traf and explicit explicit range for moof

# Generic sub-picture track grouping extensions

Issue:[*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/53*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/53)

*And updated proposal in http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/152*

The purpose of this proposal is to allow describing 2D spatial relationship between multiple video bitstreams that relate to a same source content (characterized by a source\_id) (for instance, multiple videos representing subparts of a large panorama). This is currently defined in OMAF, but the concept is generic-enough to apply to ISOBMFF.

## Updated semantics of track\_group\_type

The semantics of track\_group\_type of the TrackReferenceBox is changed from

track\_group\_type indicates the grouping\_type and shall be set to one of the following values, or a value registered, or a value from a derived specification or registration:

'msrc' indicates that this track belongs to a multi-source presentation. Specified in 8.3.4.4.1.

'ster' indicates that this track is either the left or right view of a stereo pair suitable for playback on a stereoscopic display. Specified in 8.3.4.4.2.

The pair of track\_group\_id and track\_group\_type identifies a track group within the file. The tracks that contain a particular TrackGroupTypeBox having the same value of track\_group\_id and track\_group\_type belong to the same track group.

to

track\_group\_type indicates the grouping\_type and shall be set to one of the following values, or a value registered, or a value from a derived specification or registration:

'msrc' indicates that this track belongs to a multi-source presentation. Specified in clause 8.3.4.3 of ISO/IEC 14496-12.

'ster' indicates that this track is either the left or right view of a stereo pair suitable for playback on a stereoscopic display. Specified in clause 8.3.4.4.2 of ISO/IEC 14496-12.

'2dsr' indicates that this track belongs to a group of tracks with two dimensional spatial relationships (e.g. corresponding to spatial parts of a video source). Specified in clause (TBD) of this document.

The pair of track\_group\_id and track\_group\_type identifies a track group within the file. The tracks that contain a particular TrackGroupTypeBox having the same value of track\_group\_id and track\_group\_type belong to the same track group.

Add a new section in 12.1 Video media:

## 12.1.10 Two dimensional spatial relationships

### 12.1.10.1 Definition

A SpatialRelationship2DDescriptionBox TrackGroupTypeBox indicates that this track belongs to a group of tracks with 2D spatial relationships (e.g. corresponding to planar spatial parts of a video source). A SpatialRelationship2DDescriptionBox TrackGroupTypeBox with a given track\_group\_id implicitly defines a coordinate system with an arbitrary origin (0,0) and a maximum size defined by total\_width and total\_height; the x-axis is oriented from left to right and the y-axis from top to bottom. The tracks that have the same value of source\_id within a SpatialRelationship2DDescriptionBox TrackGroupTypeBox are mapped as being originated from the same source and their associated coordinate systems share the same origin (0,0) and the orientation of their axes. For example, a very high resolution video could have been split into sub-picture tracks. Each sub-picture track then conveys its position and sizes in the source video.

Tracks in the same track group shall declare the same source\_id, total\_width, and total\_height. Track groups with different track\_group\_id values and the same source\_id represent the same source content, possibly at different resolutions (i.e. with different values of total\_width or total\_height).

NOTE: A source can be represented by different such track groups (for instance when the same source is available at different resolutions). Each of these track groups is identified by its own identifier track\_group\_id. Since all of these track groups originate from the same source, they share the same source\_id.

There shall be one or more associated SpatialRelationship2DGroupEntry(s) in the associated track (this track possibly has a dynamic size and/or position). When every sample maps to the default indicated in the SampleGroupDescriptionBox in the MovieBox; that SampleGroupDescriptionBox can indicate a default sample group or indicate that all samples are unmapped, depending on its version by setting the static\_group\_description and static\_mapping flags.

### 12.1.10.2 Syntax

aligned(8) class SpatialRelationship2DSourceBox  
 extends FullBox('2dss', 0, 0) {  
 unsigned int(32) total\_width;  
 unsigned int(32) total\_height;  
 unsigned int(32) source\_id;  
}

aligned(8) class SpatialRelationship2DDescriptionBox extends TrackGroupTypeBox('2dsr') {  
 // track\_group\_id is inherited from TrackGroupTypeBox;  
 SpatialRelationship2DSourceBox(); // mandatory, must be first  
 // other optional boxes  
}

### 12.1.10.3 Semantics

total\_width specifies the maximum width in the coordinate system of the SpatialRelationship2DDescriptionBox track group. The value of total\_width shall be the same in all instances of SpatialRelationship2DDescriptionBox with the same value of track\_group\_id.

total\_height specifies the maximum height in the coordinate system of the SpatialRelationship2DDescriptionBox track group. The value of total\_height shall be the same in all instances of SpatialRelationship2DDescriptionBox with the same value of track\_group\_id.

source\_id parameter provides a unique identifier for the source. It implicitly defines a coordinate system associated to this source.

## 12.1.10.4 Spatial relationship 2D sample group

### Definition

The '2dsr' grouping\_type for sample grouping declares the positions and sizes of the samples from a sub-picture track in a spatial relationship track group. Version 1 of the SampleToGroupBox shall be used when grouping\_type is equal to '2dsr'. The value of grouping\_type\_parameter shall be equal to track\_group\_id of the corresponding spatial relationship track group.

### Syntax

class SpatialRelationship2DGroupEntry extends VisualSampleGroupEntry('2dsr') {  
 unsigned int(16) object\_x;  
 unsigned int(16) object\_y;  
 unsigned int(16) object\_width;  
 unsigned int(16) object\_height;   
 /\* Application specific extension here \*/  
 unsigned int(32) app\_specific\_parameters;  
}

### Semantics

object\_x specifies the horizontal position of the top-left corner of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The position value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 0 to total\_width − 1, inclusive, where total\_width is included in the corresponding SpatialRelationship2DDescriptionBox.

object\_y specifies the vertical position of the top-left corner of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The position value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 0 to total\_height − 1, inclusive, where total\_height is included in the corresponding SpatialRelationship2DDescriptionBox.

object\_width specifies the width of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The width value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 1 to total\_width, inclusive.

object\_height specifies the height of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The height value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 1 to total\_height, inclusive.

app\_specific\_parameters is a parameter that provides an extension point to define codec specific parameters.

# Signaling of Multi-Layer Picture Compositing Information in the VVC File Format

## Introduction

From the minutes of File-format October 2020: We add to the TuC for ISOBMFF this question and problem area, with this solution as a strawman that is too VVC-specific, with notes etc. on the problems, and welcome further contributions on the subject to develop a generic solution suitable for the ISOBMFF.

VVC file format is being developed as part of the Amendment to ISO/IEC 14496-15 on Carriage of VVC and EVC in ISOBMFF [1].

Multi-layer support in the VVC file format is described in clause 11.3.4. It is desirable to provide a general and flexible solution to multi-layer picture compositing which supports bitstream extraction and merging (BEAM) functionalities, while minimizing bitrate overhead. With this motivation, two SEI messages were proposed for VVC in [2]-[3], which describe recommended composite pictures of decoded pictures from multiple layers. The layer composite SEI message is sent for each layer and contains parameters for the current layer’s decoded picture. The recommended composite layers info SEI message applies to multiple layers, and includes parameters for each output layer set (OLS).

In this contribution, we propose a system-layer alternative to the signaling of recommended multi-layer picture compositing information. In particular, two new entity groups for the VVC file format are proposed for this purpose.

The layer composite position info entity group applies to the decoded picture of the different layers of the VVC bitstream (each layer identified by nuh\_layer) and includes syntax elements to indicate the following for each layer:

* top left vertical and horizontal position of the decoded layer picture within a composite picture, in units
* height and width of the decoded layer picture within a composite picture, in units

The recommended composite layers info entity group signals parameters for a recommended composite pictures for each OLS, which apply to multiple layers of the VVC bitstream and contains syntax elements to indicate the following:

* flag to indicate if scaling of decoded layer pictures in the composite picture is enabled
* flag to indicate if decoded layer pictures may overlap in the composite picture
* number of OLSs
* For each OLS
  + size of the units in luma samples of the position and size syntax elements in the layer composite position info entity group
  + size of the composite picture
  + offset to be applied to all decoded layer pictures in the composite picture

The size of each layer’s decoded picture is not required to equal its recommended display size in the composite picture, unless the flag indicates that scaling is not enabled. If not equal, scaling is applied to the layer’s decoded picture when forming the composite picture. The scaling method to be used is not specified.

The decoded pictures may overlap one another in the composite picture, with the layer with the higher value of nuh\_layer\_id taking precedence, unless the flag indicates that overlap is not enabled. There is no requirement that the entire composite picture be covered by decoded layer pictures, Uncovered areas in the composite picture are undefined and left to the application to define.

The offset parameters are intended to be used when some layers are not included in the OLS, without needing to change the per layer parameters in each layer composite position info entity group.

## Proposed New Signaling for Amendment in [1]

### Layer Composite Position Info Entity Group

The layer composite position info Entity Group describes the recommended position and size of the decoded picture of each layer within a recommended composite picture comprised of decoded pictures from multiple layers.

**Syntax**

aligned(8) class LayerCompositePositionGroupBox extends   
 EntityToGroupBox('lcpg',0,0)  
{  
 unsigned int(16) num\_olss;  
 for (i=0; i<num\_olss; i++){   
 unsigned int(16) output\_layer\_set\_idx;  
 unsigned int(8) layer\_count;  
 for (j=0; j<layer\_count; j++) {  
 unsigned int(8) layer\_id;  
 unsigned int(12) lcpi\_param\_num\_bits\_minus1;  
 unsigned int(lcpi\_param\_num\_bits\_minus1+1)   
 lcpi\_top\_left\_ pos\_in\_units\_ver;  
 unsigned int(lcpi\_param\_num\_bits\_minus1+1)   
 lcpi\_top\_left\_pos\_in\_units\_hor;  
 unsigned int(lcpi\_param\_num\_bits\_minus1+1)  
 lcpi\_width\_in\_units;  
 unsigned int(lcpi\_param\_num\_bits\_minus1+1)  
 lcpi\_height\_in\_units;   
 }  
 }  
}

**Semantics**

num\_olss indicates the number of OLSs for which syntax elements are present in the entity group.

output\_layer\_set\_idx is the index of the output layer set. The mapping between output\_layer\_set\_idx and the layer\_id values shall be the same as specified by the VPS for an output layer set with index output\_layer\_set\_idx.

layer\_count: This field indicates the number of necessary layers, as defined ISO/IEC 23090-3, for this output layer set.

layer\_id: provides the nuh\_layer\_id values for the layers of the output layer set.

lcpi\_param\_num\_bits\_minus1 + 1 specifies the number of bits used to represent the lcpi\_top\_left\_pos\_ver, lcpi\_top\_left\_pos\_hor, lcpi\_width, and lcpi\_height syntax elements.

lcpi\_top\_left\_pos\_verandlcpi\_top\_left\_pos\_horindicate the recommended composite display vertical and horizontal positions, respectively, for the decoded picture of the current layer given by layer\_id. The number of bits to represent the syntax elements is lcpi\_param\_num\_bits\_minus1 + 1.

lcpi\_widthandlcpi\_heightindicate the recommended composite display width and height, respectively, for the decoded picture of the current layer given by layer\_id. The number of bits to represent the syntax elements is lcpi\_param\_num\_bits\_minus1 + 1.

### Recommended Composite Layers Info Entity Group

The recommended composite layers info entity group, together with the layer composite position information entity group, describe a layout of decoded pictures from the layers of an OLS within a recommended composite picture. If the recommended composite layer info entity group is present, then the layer composite position information entity group shall also be present.

**Syntax**

aligned(8) class RecommendedCompositeLayersInfoGroupBox extends   
 EntityToGroupBox('rclg',0,0)  
{  
 unsigned int(1) rcli\_layer\_scaling\_enabled\_flag;  
 unsigned int(1) rcli\_layer\_overlap\_enabled\_flag;  
 unsigned int(1) rcli\_unit\_size\_present\_flag;  
 unsigned int(1) rcli\_composite\_size\_present\_flag;  
 unsigned int(1) rcli\_offset\_present\_flag;  
 unsigned int(16) num\_olss;  
 for (i=0; i<num\_olss; i++){  
 unsigned int(16) output\_layer\_set\_idx;  
 if(rcli\_unit\_size\_present\_flag){  
 unsigned int(16) rcli\_unit\_size\_ver;  
 unsigned int(16) rcli\_unit\_size\_hor;  
 }  
 if(rcli\_composite\_size\_present\_flag){  
 unsigned int(16) rcli\_composite\_size\_ver;  
 unsigned int(16) rcli\_composite\_size\_hor;  
 }  
 if(rcli\_offset\_present\_flag){  
 signed int(16) rcli\_offset\_ver;  
 signed int(16) rcli\_offset\_hor;  
 }  
 }  
}

**Semantics**

rcli\_layer\_scaling\_enabled\_flag equal to 0 indicates that layerPicSizeInCompositeHeight[ i ] and layerPicSizeInCompositeWidth[ i ] derived in composition process in subclause 2.3 are equal to the width and height, respectively, of the coded picture with nuh\_layer\_id equal to i. rcli\_layer\_scaling\_enabled\_flag equal to 1 indicates that layerPicSizeInCompositeHeight[ i ] and layerPicSizeInCompositeWidth[ i ] may differ from the width and height, respectively, of the coded picture with nuh\_layer\_id equal to i.

rcli\_layer\_overlap\_enabled\_flag equal to 0 indicates that all values of Count[ y][ x ] shall be les than or equal to 1, as derived in subclause 2.3. rcli\_layer\_overlap\_enabled\_flag equal to 1 does not impose a restriction.

rcli\_unit\_size\_present\_flag equal to 1 specifies that the rcli\_unit\_size\_ver[ i ] and rcli\_ unit\_size\_hor[ i ] syntax elements are present. rcli\_unit\_size\_present\_flag equal to 0 specifies that the rcli\_ unit\_size\_ver[ i ] and rcli\_ unit\_size\_hor[ i ] syntax elements are not present.

rcli\_composite\_size\_present\_flag equal to 1 specifies that the rcli\_composite\_size\_ver[ i ] and rcli\_composite\_size\_hor[ i ] syntax elements are present. rcli\_composite\_size\_present\_flag equal to 0 specifies that the rcli\_composite\_size\_ver[ i ] and rcli\_composite\_size\_hor[ i ] syntax elements are not present.

rcli\_offset\_present\_flag equal to 1 specifies that the rcli\_offset\_ver[ i ] and rcli\_offset\_hor[ i ] syntax elements are present. rcli\_offset\_present\_flag equal to 0 specifies that the rcli\_offset\_ver[ i ] and rcli\_offset\_hor[ i ] syntax elements are not present.

num\_olss indicates the number of OLSs for which syntax elements are present in the entity group.

output\_layer\_set\_idx is the index of the output layer set. The mapping between output\_layer\_set\_idx and the layer\_id values shall be the same as specified by the VPS for an output layer set with index output\_layer\_set\_idx.

rcli\_unit\_size\_ver[ i ] and rcli\_unit\_size\_hor[ i ] indicate vertical and horizontal unit size parameters respectively, used in the composition process in subclause 2.3 for the i-th OLS. When not present, the values of rcli\_offest\_ver[ i ] and rcli\_offset\_hor[ i ] may be determined by external means.

rcli\_composite\_size\_ver[ i ] and rcli\_composite\_size\_hor[ i ] indicate the vertical and horizontal size, respectively, of the recommended composite picture in luma samples used in the composition process in subclause 2.3 for the i-th OLS. When not present, the values of rcli\_offest\_ver[ i ] and rcli\_offset\_hor[ i ] may be determined by external means.

rcli\_offset\_ver[ i ]andrcli\_offset\_hor[ i ]indicate vertical and horizontal offsets, respectively, of the positions of the decoded layer pictures used in the composition process in subclause 2.3 for the i-th OLS. When not present, the values of rcli\_offest\_ver[ i ] and rcli\_offset\_hor[ i ] are inferred to be equal to 0.

### Recommended composition process (from [2])

This subclause describes a composition process to derive sample values for a recommended composite picture, CompositePicture, for the i-th OLS.

CompositePicture[ 0]  is a 2-D sample array of size rcli\_composite\_size\_hor x rcli\_composite\_size\_ver, of the luma samples of CompositePicture.

CompositePicture[ cIdx ]  for cIdx in 1 .. 2 are 2-D sample arrays of size rcli\_composite\_size\_hor/ SubWidthC x rcli\_composite\_size\_ver/ SubHeightC.

For each layer j included in the i-th OLS, if a picture is present in the layer access unit with nuh\_layer\_id equal to j, the recommended size of the representation of the decoded picture in the composite picture, scaledLayerPic[ j ], is layerPicSizeInCompositeWidth[j ] x layerPicSizeInCompositeHeight[ j ], in luma samples, as derived below.

layerPicSizeInCompositeHeight[ j ] = lcpi\_height\_in\_units[j ] \* rcli\_unit\_size\_ver[ i ]

layerPicSizeInCompositeWidth[ j ] = lcpi\_width\_in\_units[ j ] \* rcli\_unit\_size\_hor[ i ]

When rcli\_layer\_scaling\_enabled\_flag equal to 0, scaledLayerPic[ j ] is set to the decoded picture.

Otherwise, scaledLayerPic[ j ] is derived by scaling the decoded picture.

scaledLayerPic[ j ] is a picture of size layerPicSizeInCompositeWidth[ j ] x layerPicSizeInCompositeHeight[ j ], in the luma samples .

The sample values of CompositePicture are derived as follows:

for (y = 0; y < rcli\_composite\_size\_ver[ i]; y++)  
 for (x = 0; x < rcli\_composite\_size\_hor[ i ]; x++)  
 Count[ y ][ x ] = 0  
for (j = 0 ; j< 64; j++)   
 if (j is in the i-th OLS && a picture is present in the AU with nuh\_layer\_id = j)  
 comp\_y = rcli\_offset\_ver[ i ] + lcpi\_top\_left\_pos\_in\_units\_ver[ j ] \* rcli\_unit\_size\_ver  
 comp\_x = rcli\_offset\_hor[ i ] + lcpi\_top\_left\_pos\_in\_units\_hor[ j ] \* rcli\_unit\_size\_hor  
 for (y = 0; y < layerPicSizeInCompositeHeight[ j ] \* ; y++)  
 for (x = 0; x < layerPicSizeInCompositeWidth[ j ]; x++)  
 CompositePicture[ 0 ] [ comp\_y + y ][ comp\_x + x ] = scaledLayerPic[ 0 ][ j ][ y ][ x ]  
 Count[[ comp\_y + y ][ comp\_x + x ]]++  
 for (y = 0; y < layerPicSizeInCompositeHeight[ j ]/SubWidth \* ; y++)  
 for (x = 0; x < layerPicSizeInCompositeWidth[ j ]/SubHeight; x++)  
CompositePicture[ 1 ][ comp\_y/SubHeight + y ][ comp\_x/SubWidth + x ] =  scaledLayerPic[ 1 ][ j ][ y ][ x ]  
CompositePicture[ 2 ][ comp\_y/SubHeight + y ][ comp\_x/SubWidth + x ] = scaledLayerPic[ 2 ][ j ][ y ][ x ]

The value of samples of CompositePicture[cIdx] not assigned above are undefined.

## Discussion

Figures 1, 2, and 3 illustrate example usages of multiple OLSs with different offset values, resulting in a different composite picture from the same bitstream for each OLS.

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 |
| 8 | | 10 | |
| 9 | |

**Figure 1. Example layout for OLS 0**

|  |  |
| --- | --- |
| 2 | 3 |
| 6 | 7 |
| 10 | |
|

**Figure 2. Example layout for OLS 1**

|  |  |
| --- | --- |
| 5 | 6 |

**Figure 3. Example layout for OLS 2**

In this example, 11 layers are present in the bitstream, with values of nuh\_layer\_id from 0 .. 10. The recommended display sizes of the pictures of layers 0 to 7 are the same, 480x240. The recommend display size of pictures of layers 8 and 9 are 960x240. The recommended display size of the picture of layer 10 is 960x480. In this example, the unit size can be 480x240 luma samples, as that is the smallest unit of granularity of sizes and positions needed.

The sizes and positions in units are signaled for each layer in its own layer composite position info entity group. The signaled size corresponds to the recommended display size and not the coded picture size. Scaling is recommended to be applied when the coded picture size differs from the display size, but a scaling method is not defined. Explicit signaling of the recommended display size enables use of reference picture resizing, where the coded picture size may vary but the recommended display size remains constant. Scaling of the pictures from some layers and not from others may also be desired for other applications.

**Table 1. Example syntax element values in layer composite position info entity group (common to all OLSs)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **nuh\_layer\_id** | **lcpi\_top\_left\_ pos**  **\_in\_units\_ver** | **lcpi\_top\_left\_pos**  **\_in\_units\_hor** | **lcpi\_height\_in\_units** | **lcpi\_width\_in\_units** |
| 0 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 2 | 0 | 2 | 1 | 1 |
| 3 | 0 | 3 | 1 | 1 |
| 4 | 1 | 0 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 |
| 6 | 1 | 2 | 1 | 1 |
| 7 | 1 | 3 | 1 | 1 |
| 8 | 2 | 0 | 1 | 2 |
| 9 | 3 | 0 | 1 | 2 |
| 10 | 2 | 2 | 2 | 2 |

**Table 2. Example syntax elements in multiple recommended composite layers info entity group**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **OLS idx** | **rcli\_composite\_size\_ver** | **rcli\_composite\_size\_hor** | **rcli\_offset\_ver** | **rcli\_offset\_hor** |
| 0 | 960 | 1920 | 0 | 0 |
| 1 | 960 | 960 | 0 | -960 |
| 2 | 480 | 960 | -240 | -480 |

## References

[1] w19454: Text of ISO/IEC 14496-15:2019 DAM 2 Carriage of VVC and EVC in ISOBMFF

[2] m54093: AHG9/AHG12: Recommended multi-layer composite picture SEI messages

[3] m54832/JVET-S2017: Technologies under consideration for VSEI

# Integrating new codecs

Issue: [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/95*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/95)

## Introduction

New codecs (audio, video, text, …) are regularly defined within MPEG or externally, and that need to be integrated into ISOBMFF. After 20 years of evolution, the best way to integrate a new codec is not necessarily the same way it was years ago. This section contains a discussion on this topic for the purpose of improving the documentation, for example Annex B.5 of ISOBMFF.

There are 2 high level questions one can ask when integrating a codec into ISOBMFF:

* Given a codec already specified, how to best integrate it into ISOBMFF? What tools should I use for this and that?
* Given a codec under development, how to structure its high level syntax such that it integrates well with ISOBMFF and what are the pitfalls to avoid?

These questions are discussed below.

## Integrating an existing codec

The following questions should be asked when new codecs are integrated in ISOBMFF:

* Which handler should be used? Should a new handler be defined?

Although it is not documented (yet?), the underlying assumptions of a track handler are the following:

* Codecs within a handler are interchangeable, in the sense that once track data is decoded, all codecs should produce the same type of output. In other words, theoretically within a track you can have multiple sample description entries each with a different codec for that handler. (But one should not do that!!!)
* Each handler comes with a base sample entry (vide -> VisualSampleEntry, soun -> AudioSampleEntry, …). If a base sample entry contains the necessary fields for a codec (sample rate or width/height), probably the codec should use the corresponding handler (if the base sample entry is not enough, extend it using the base sample entry as a parent class). Conversely, if there is no corresponding base sample entry (or if there is no such media type yet supported), a new handler should be created.
* How to design the codec-specific Sample Entry ?

**About Decoder initialization.** It has been assumed in the past that whatever is needed for decoder initialization (including opaque sequences of bytes) should be in the sample entry. Doing so should be considered carefully as it leads to the problematic dichotomy: live vs. ondemand (avc1 vs avc3, hvc1 vs. hev1, …). In live cases, all opaque sequences of bytes for the entire session are not necessarily known upfront, and creating a new sample entry on the fly is not (yet?) possible in ISOBMFF. Usually the concern with decoder initialization is initialization latency, but often latency is due to memory allocation which for example for video can be done by knowing width, heigh, and depth. MPEG welcomes feedback on how decoder initialization is done for various codecs. One can use this public issue https://github.com/MPEGGroup/FileFormat/issues/58.

**How to compute “codecs” parameter?** In segmented media, it should be possible to compute the “codecs” parameter without having to fetch any media segment, and only the initialization segment. This should be done primarily based on the information in the sample entry. Sample group information should not be used (because that would remove freedom to put sample group descriptions in the fragments).

**Consider content splicing complexity.** Content splicing (i.e. merging 2 single-track files with the same handler and codec into 1 single-track file) is a typical operation that can be performed with ISOBMFF files. It can always be done by using 2 sample description entries in the output track. However, using multiple sample entries makes implementation more complex (and the “codecs” parameter only describes the first one). A proposed rule is to minimize the amount of data in the sample description in order to maximize the opportunity for single sample description splicing. Processing sample entries when splicing should be as simple as "doing a binary comparison of the 2 input sample entries and if they differ produce 2 output sample entries". Alternatively, having specific merge rules such that “if the fields A, B, C differ between sample entries 1 and 2, create a sample entry with max (A1,A2), max(B1,B2) max(C1,C2)"

A good way to reduce the amount of information in the sample entry is to use **sample groups**. All samples of a track can be easily made part of a common sample group, for example using the default sample group in the ‘sgpd’ box without even using an ’sbgp’ box. This should be considered in particular instead of defining a new sample entry child box type (e.g. colr, pasp, dmix, …).

## Adjusting a codec under development for better integration

In order to get integrated simply and to produce less error-prone files, codecs should have some good properties as discussed here. Designers of high level syntaxes of new codecs should try to follow these recommendations:

* **Enable shallow parsing.** The basic operation of an ISOBMFF packager is to read an elementary stream and produce an ISOBMFF file. Therefore, it should be simple for the packager to identify what will constitute a sample. It should be able to parse the stream (without decoding it fully) and to identify elements that are needed by the packager.
  + Examples of stream types following this good practice are: NALU-based streams, OBU-based streams, TLV-based streams.
  + Example of stream types not following such approach: AAC (non-ADTS) streams.
* **Separate information with different update rates**. Streams typically have information that vary over time, but at different rates. For example, in a video stream, slice-level information will vary frequently and be different for each sample. Picture Parameter Set may vary at each sample but typically varies at GoP granularity or more. Sequence Parameter Sets vary even less, e.g. only a few times in the scope of a sample entry. Profile and level values typically don’t vary within a track. Designers of new codecs should structure data in a way that the packager can easily identify data that it needs and that varies at different rates to store them at the appropriate place (sample entry, sample group, sample). This can mean creating new units (NALU, OBU, Packet) specifically for each type of data.
* **Reduce elementary stream/container file redundancy.** In today’s designs, there is often redundancy between elementary stream information and file level information. For example, width and height are stored in the elementary stream but also in the visual sample entry. The same is true for a lot of information: color (VUI and colr box), HDR static metadata (SEI and ‘mdcv’/’clli’ boxes), sample rate, frame rate, etc. Often the information is replicated at the ISOBMFF level in a codec-agnostic way and to simplify the processing at the client side. But this is a source of error when only one level is modified and not the other one. Approaches to overcome this problem include:
  + Design data units that are codec-agnostic, leveraging CICP or equivalent whenever possible.
  + Consider the possibility of the packager replacing entire data units with ISOBMFF structures when storing in ISOBMFF and the demuxer restoring these data units from the ISOBMFF structures. For example, if a stream had a color (NAL/OB)unit, the packager could avoid storing that unit, replacing it with a ‘colr’ box, and upon reading the demuxer would recreate a (NAL/OB)unit from the ‘colr’, if needed or directly communicate the ‘colr’ box to the decoder/renderer.
* **Ease identification of encryptable data**. Often when encrypting streams, payload headers are kept in the clear. However, it is not always easy to determine where those headers end without having to parse the entire header. Codec designs should consider facilitating this (adding header length, using fixed headers, …).

# MovieFragmentHeaderBox update

Issue : [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/97*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/97)

MPEG is considering defining a new version of the ‘mfhd’ box as follows:

aligned(8) class MovieFragmentHeaderBox

extends FullBox('mfhd', version, flags){

if (version == 0) {

unsigned int(32) sequence\_number;

else if (version == 1) {

unsigned int(64) sequence\_number;

}

}

With the additional semantics:

When version 1 is used, the following ‘flags’ values have the following meaning:

0 : the creator makes no statements, promises, warranties about how sequence\_number is updated

1 : the sequence\_number in this moviefragment is larger than the sequence\_number in the preceding one

3 : the sequence\_number in this moviefragment is one greater than the sequence\_number in the preceding one

# Sample Run Sample Group

Issue : [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/98*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/98)

## Introduction

The overhead of file format metadata, such as the TrackRunBox(es), could be significant especially when it comes to applications using multi-track approach. The overhead of TrackRunBox(es) has been clearly established and thoroughly studied in the context of OMAF tile based streaming [1-3]. For convenience, an example from [1] is copied below.

*“Imagine a 4K video being HEVC encoded with 50 tiles, each tile measuring 384x384 pixels. In such a tile, it is not uncommon for a NAL unit containing a P or B-slice to be as small as 40 bytes, and in areas with little motion sometimes even significantly less. In such a case, the 32-bit sample\_size already results in at least a 10% TrackRunBox overhead, and that’s assuming all of the other optional TrackRunBox fields have been disabled.”*

Conventionally, file writers operate by parsing the high-level syntax of a given input video bitstream and generate the file format metadata from the information of the bitstream. Consequently, under certain constraints, the information present in TrackRunBox(es) could instead be regenerated at the client side by parsing the high-level syntax of the received bitstreams.

More details are provided in Section 14.2, below.

## Discussion

For video bitstreams encapsulated as a track, the information of the TrackRunBox could be tightly packed at a coarser granularity and additional finer-level information can be concluded in the file reader based on the received MediaDataBox for a movie fragment. Figure 19.2.1 shows the conventional encapsulation of TrackRunBox where a encoded segment has m access units, with each access unit having either VCL NAL units or non-VCL units or both. Information about each access unit is encapsulated in a sample of the track as part of the TrackRunBox.

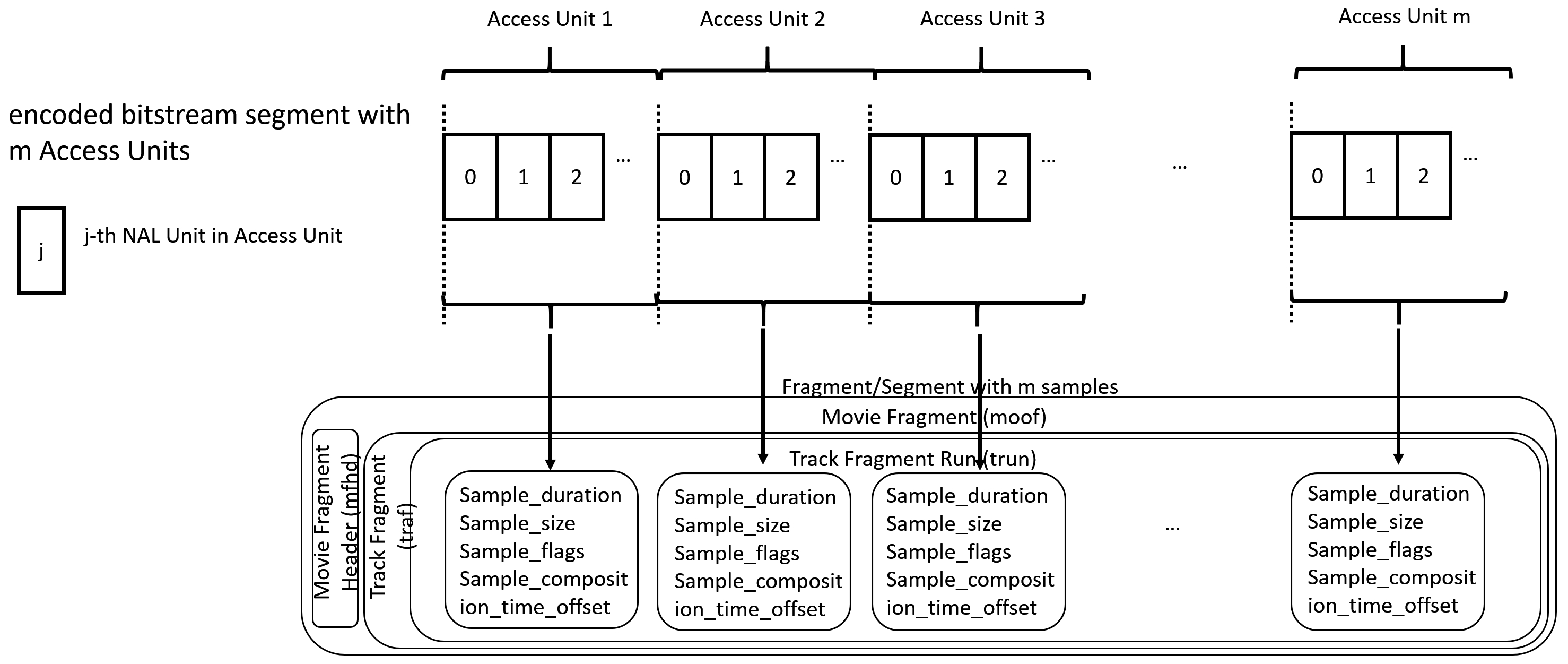


Figure 19.2.1: Conventional sample information in TrackRunBox

As shown in Figure 19.2.2, the sample-level information of the TrackRunBox (m samples) are tightly packed into a single sample information in the SampleRun sample group. Kindly note that the packing of m samples into a single sample in Figure 19.2.2 is only for demonstrating the usage of SampleRun sample group. The SampleRun sample group may contain information about n < m samples where more than one sample of the original TrackRunBox are tightly packed into the SampleRun sample group samples.

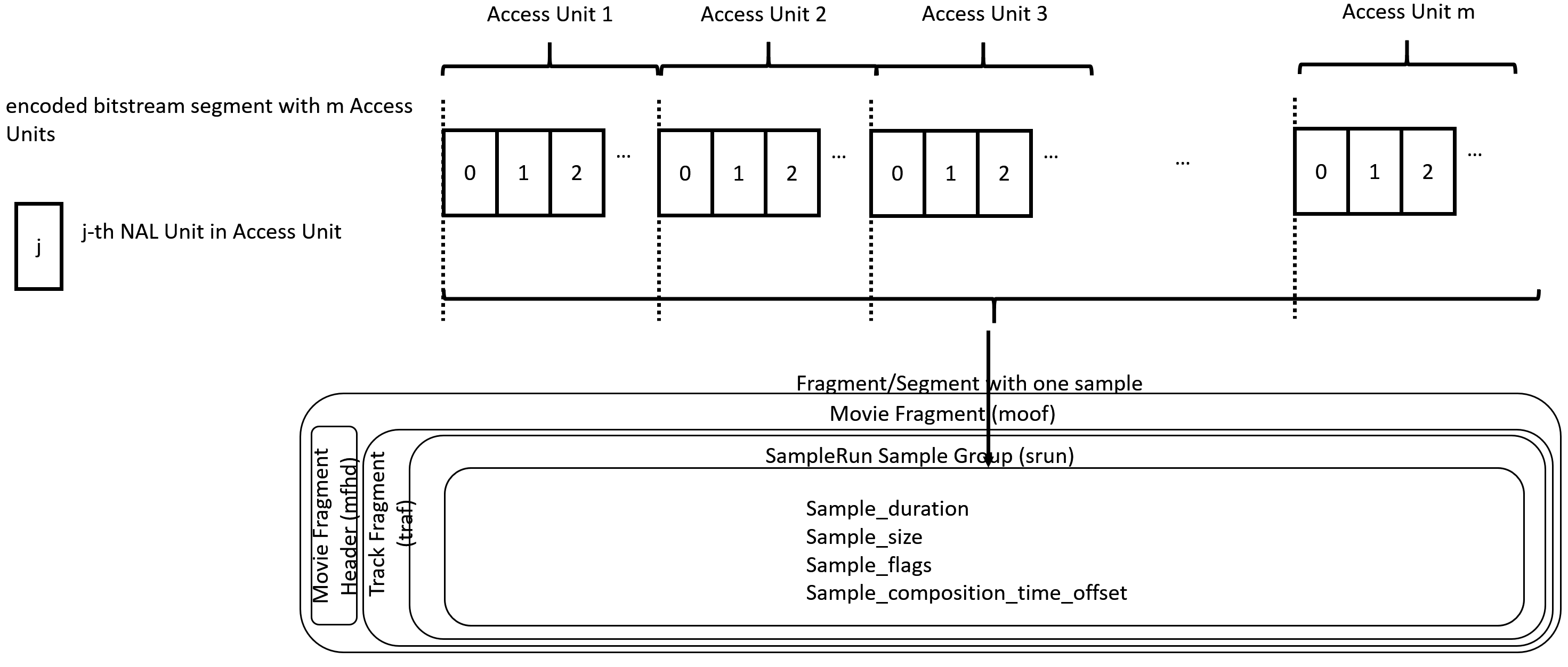


Figure 19.2.2: An example usage of SampleRun Sample group.

An approach for optimizing the delivery of MovieFragmentBox for video could be designed as follows.

* A file writer performs the following to pack a MovieFragmentBox:
  + Identifying GOP structures where pictures have the same pattern of duration, composition time offset and sample flags.
  + Generating a SampleRun sample group description entry for each identified GOP structure.
  + Merging the samples of each GOP into one sample in an 'tspt' transformed video track.
  + Mapping the samples in the 'tspt' transformed video track to the respective SampleRun sample group description entries in SampleToGroupBox(es).
* A player:
  + Performs the access unit boundary determination as specified in AVC, HEVC, VVC, or EVC for each sample in an 'tspt' transformed video track.
  + Generates the original TrackRunBox information in MovieFragmentBoxes based on the signalled SampleRun sample group and the sample sizes determined from the access unit boundaries.

## Proposal

It is proposed to support the new SampleRun sample group as follows.

### Definition

A sample run sample group may be present in an 'tspt' transformed video track. A sample run documents contiguous samples of the untransformed track where each access unit in the mapped sample of the 'tspt' transformed video track is its own sample.

Note: The sample run sample group is limited to be used in an 'tspt' transformed video track containing bitstreams which support picture boundary detection. For example, VVC, which supports signalling of picture header information and the client need not maintain a parsing context to detect picture boundaries.

### Syntax

class SampleRunEntry() extends SampleGroupDescriptionEntry('srun') {  
 unsigned int(24) flags;

if (flags & 0x000008)

unsigned int(32) default\_sample\_duration;

if (flags & 0x000010)

unsigned int(32) default\_sample\_size;

if (flags & 0x000020)

unsigned int(32) default\_sample\_flags;

unsigned int(32) sample\_count;

if (flags & 0x000004)

unsigned int(32) first\_sample\_flags;

{

if (flags & 0x000100)

unsigned int(32) sample\_duration;

if (flags & 0x000200)

unsigned int(32) sample\_size;

if (flags & 0x000400)

unsigned int(32) sample\_flags;

if (flags & 0x000800)

signed int(32) sample\_composition\_time\_offset;

}[ sample\_count ]  
}

### Semantics

flags is a map of flags

The following flags are allowed to be set in the flags:

0x000008 default-sample-duration-present

0x000010 default-sample-size-present

0x000020 default-sample-flags-present

0x000004 first-sample-flags-present; this overrides the default flags for the first sample only. This makes it possible to record a group of frames where the first is a key and the rest are difference frames, without supplying explicit flags for every sample. If this flag and field are used, sample-flags-present shall not be set.

0x000100 sample-duration-present: indicates that each sample has its own duration, otherwise the default is used.

0x000200 sample-size-present: each sample has its own size, otherwise the default is used.

0x000400 sample-flags-present; each sample has its own flags, otherwise the default is used.

0x000800 sample-composition-time-offsets-present; each sample has a composition time offset.

default\_sample\_duration: indicates the default duration of the samples in the sample run

default\_sample\_size: indicates the default size of the samples in the sample run

default\_sample\_flags: indicate the default flags values for the samples in the sample run.

sample\_count the number of samples in this sample run

first\_sample\_flags provides a set of flags for the first sample only of this sample run.

# Improvements to movie fragments and dynamic tracks

The File Format group contained in its mandate the following topic:

1. Study improvements to movie fragments especially when recording e.g. at gateways, and dynamic tracks.

The inputs below are considerations and proposals addressing this topic.

## Dependent Movie Fragments

Reference to discussions: <http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/148>

### Discussion

With low-latency delivery techniques for ISOBMFF, it is quite common that demultiplexers synchronize on the first fragment in a series (e.g., segment) but do not try to demultiplex the following fragments first, typically because the SAP frame is in the first fragment of the series. This means that a lot of information is duplicated in the following fragments due to ISOBMFF rules, but are identical with the information sent in the first segment used to synchronize. This applies to:

* sample group descriptions inserted in movie fragments
* MetaBox, UserDataBox
* Any possible extension of ISOBMFF allowing box injection in movie fragments that do not carry per-sample information

This is because ISOBMFF makes no difference between movie fragments in terms of random access for the parser, while higher delivery protocols impose different constraints on fragments (first fragment of a segment is a SAP1, must have a TFDT, etc…).

For example, a typical ‘seig’ sample group description inserted in each fragment to allow for key rolling will cost, assuming a single entry (one key) is used:

* 61 bytes if constant IV 128bits is used (cbcs)
* 46 for 128 bits IV (cenc)

At 100ms fragment duration, this leads to costs of 3.7 / 4.9 kbps.

In very low latency cases (one sample per fragment):

* for 25fps video : 9.2 / 12.2 kbps
* for AAC at 44100Hz: 16 / 21 kbps.

The signaling rate will obviously increase:

* when adding other sample group descriptions to be updated in the fragments or other boxes such as MetaBoxes
* when more complex sample descriptions are used, e.g. multiple keys per sample.

### Proposal

It is proposes to define a new version in the movie fragment header box:

aligned(8) class MovieFragmentHeaderBox extends FullBox('mfhd', version, 0){   
 unsigned int(32)sequence\_number;   
}

If version is not 0, any SampleGroupDescriptionBox, UserDataBox or MetaBox defined in the last movie fragment, or in the last TrackFragmentBox in the last movie fragment, whose MovieFragmentHeaderBox version is 0 also apply for this movie fragment, and there shall not be any SampleGroupDescriptionBox or MetaBox defined for this movie fragment.

Note that this new version of a movie fragment can still be parsed, in terms of box structure and sample information, independently of the previous fragment, however its interpretation usually re-use information from a previous movie fragment.

### Comments received at MPEG#140

* We already have a new version in the TuC (see Section 13), should we use flags instead?
* Should we pre-declare in the ’mvex’? or ’moov’?
* The use of 'dmof' instead of 'moof' as an option should be studied, so that old readers don't get puzzled when they see something that is erroneous in their opinion.
* we may not always have sample group or encryption for true low latency
* meta and udta should not be in the fragments, they are not mandatory at all therefore rotating keys based on the segments, maybe better to use the manifest for key rotation instead? (but would assume that a manifest is always present).

## On dynamic tracks in fragments

This topic comes from the discussions related to the input contribution m61140:

<http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/147>

The following comments were received at MPEG#140:

- Relying only on brands for parsers to correctly understand the features can be dangerous, and a possible rename of movie fragments boxes should be considered, for example ‘dmof’ instead of reusing ‘moof’.

- Some of the proposed functionalities can be achieved using external signaling such as DASH MPD.

### Discussion

#### Signaling cost

As noted in m58085, the proposed signaling of TrackBox in movie fragments is 4kbps on average for a 1 sec duration segment. Striping the TrackBox from non-essential box (empty sample table, data references) gives around 3.1 kbps for 1s segments.

It is proposed hereafter a further reduction of the track box for inclusion in movie fragments, with an achieved overhead of roughly 1.3 kbps for 1s segments.

#### Change tracking

In use cases such as MPEG-2 TS to fMP4 gateways, the proposed dynamic tracks could be further improved by signaling (un)changed configurations in the fragments, usually called “carousel” in broadcast terminology.

It is noted that sample can be signaled as repeated samples using dependency flags (MPEG-4 systems), but we lack support for such signaling for non-sample data:

* MetaBox
* UserDataBox in track fragment
* Sample Group Description

Obviously relying on the box version is a bad choice:

* Some boxes ('udta') do not have versions
* Other boxes have a version field but already use it (sample group description)
* It is not the intended purpose of box version, which is to indicate variation in the binary syntax and not in the payload.

One approach to deal with this problem is to perform a comparison of past and current boxes, typically through a hashing function. While this allows detecting identical configurations, it has several drawbacks:

* It is costly in client resources
* It does not allow for signaling a repeated configuration with slight variations (in the meta-data for example) that do not require reparsing of the data

If we want dynamic tracks or sample description changes in movie fragments (regardless of the method), it could be useful to introduce some change detection mechanism so we can properly identify repeated information across fragments.

#### Track Removal

In MPEG-2 TS, media streams can be removed due to a PMT update. We noticed that track fragments with duration-is-empty flag set could be used to signal no samples apply to this track. It has however some drawbacks:

* The flag cannot be used if an edit list is present for this track in the movie box,
* It cannot reliably be used as a hint that the track is no longer present; this implies that
  + resource optimization (closing decoders/buffers/etc.) cannot be fully performed.
  + Remixing to e.g. MPEG-2 TS might trigger unused PIDs in the PMT (i.e. PID declared but no packet for this PID), with unpredictable result at the demuxer side
* Moreover, support for this flag is not very good, typically triggering rebuffering in some MSE implementations (because ignored)

If we want a proper/unambiguous signaling of track removal, we need a dedicated box/field to indicated tracks no longer present.

#### File Concatenation, splicing or time-aligned track addition

MPEG-2 TS and some other delivery formats can embed splicing information, allowing a media pipeline to identify temporary service reconfigurations.

While investigating support for such signaling in ISOBMFF along with dynamic tracks, we faced an old but never resolved topic in ISOBMFF: how can we simply build files out of a collection of files, whether as a sequence (‘append a presentation to another one’), or as a set of additional tracks (‘add these tracks, time-aligned’).

In fragmented mode, addition of time-aligned tracks requires editing of the 'moov' to inject the track, and then direct concatenation can work, assuming track IDs do not conflict and timestamps origin is 0 for each “track file”. This obviously will result in one media track being stored after all other ones, but the result is a valid ISOBMFF.

Extending a fragmented presentation with another one is more problematic:

* trackIDs may need to be rewritten: if not the same in each presentation, some tracks may have to be added to the initial 'moov'
* decoder configuration may need to be updated
* timing needs to be rebuilt:
  + alignment of timestamps at boundaries
  + introduction of “gaps” in the timeline to deal with AV sync
* Signaling codec priming (e.g. AAC) is even more complex: a dedicated edit list needs to be created to remove the priming period at the beginning of the second stream.

When concatenating files to splice new content (A1->B->A2), we faced the following issues:

* If the splice is a content replacement,
  + The timing of the injected content B must be rewritten
  + There is no guarantee that the injected content B ends up exactly at the right position, in which case the timing of A2 may need to be rewritten
* If the splice is a content injection (timeline is extended),
  + The timing of the injected content B must be rewritten
  + The timing of A2 must be rewritten

We therefore thought that it would be interesting to design dynamic tracks such that file concatenation is possible without any rewrite. We designed the proposal to allow:

* “early-splicing”, where the source timeline is spliced but contiguous and the only need is to signal the splice points (e.g. MPEG-2 TS to fMP4 gateway)

“late-splicing”, where splicing is done after the fMP4 muxing stage (non-contiguous timelines)

### Proposal

#### Design

We reworked the proposal from m58085 to provide a single method for declaring new tracks and declaring new sample description(s) in an existing track, based on the previous observations.

The proposal defines a DynamicMovieBox in a movie fragment, containing

* Indication on source (timeline/splicing) and configuration changes compared to previous DynamicMovieBox
* Zero or one UserData box, Zero or one Meta box
* Zero or more DynamicTrackBox, each containing
  + a specific DynamicTrackHeaderBox containing
    - track setup (width/height/delay…)
    - indications on modification changes compared to previous DynamicTrackBox
  + some common boxes found in TrackBox.

##### Dynamic movie box

###### Definition

Box Type: 'dymv'  
Container: MovieFragmentBox  
Mandatory: No  
Quantity: Zero or one

A DynamicMovieBox completely or partially overrides the MovieBox setup (track list, user data and meta) for the current fragment.

Each DynamicMovieBox has an associated source\_id, which indicates how the movie fragment extends the initial MovieBox:

* A source\_id equal to 0 indicates that the MovieBox is modified by the movie fragment,
* A source\_id different from 0 indicates that the MovieBox is ignored (i.e. considered not present)

A change of source\_id between two consecutive movie fragments N and N-1 in a single bytes sequence (file, remote resource, etc.) indicates that tracks in N shall be considered as new tracks and tracks in N-1 shall no longer be considered present in the file. In this case, there is no guarantee that the timeline is contiguous between fragments N and N-1. How file readers handle such discontinuities is out of scope of this specification and usually driven by the processing pipeline capabilities. Implementations should however avoid introducing long playback gaps at source\_id change points.

When dynamic tracks are used, the first track fragment of each track in the parent movie fragment shall have a TrackFragmentBaseMediaDecodeTimeBox.

If two consecutive movie fragments N and N-1 have the same value for source\_id, the timeline of all tracks active in both fragments is contiguous, i.e. the constraints on TrackFragmentBaseMediaDecodeTimeBox of each track shall be respected: for a track fragment with the same trackID, the first TrackFragmentBaseMediaDecodeTimeBox in movie fragment N is equal to or greater than the first TrackFragmentBaseMediaDecodeTimeBox in movie fragment N-1 plus the sum of the sample durations in movie fragment N-1.

If several tracks need to be inserted or replaced, these tracks may be declared

* all in a single DynamicMovieBox (i.e. a single movie fragment),
* each one in its own DynamicMovieBox (i.e. one movie fragment per new track) or
* a mix of both approaches.

If a track from the MovieBox is not listed either for update or removal in a DynamicMovieBox with source\_id value of 0, it is valid, but there could be no track fragment for this track in the movie fragment, as is the case with regular movie fragments.

A DynamicMovieBox may contain zero or more DynamicTrackBox.

NOTE Usage of DynamicMovieBox with source\_id different from 0 and zero DynamicTrackBox can be used to force a discontinuity between two movie fragments. Usage of DynamicMovieBox with source\_id equal to 0 and zero DynamicTrackBox can be used to update MetaBox or UserDataBox.

Tracks declared or modified in a DynamicMovieBox may have the same configuration for several consecutive movie fragments. source\_flags allow a file parser to detect that a DynamicMovieBox is a repetition of the previous DynamicMovieBox with the same values for source\_id and bundle\_id. For a same value of source\_id, if multiple dynamic tracks or movie-related metadata (user data, meta) are modified or declared in more than one DynamicMovieBox, then each of these DynamicMovieBoxes shall use a different bundle\_id.

Presence of dynamic tracks in movie fragments shall be indicated using the brand ‘dytk’ in the ExtendedTypeBox or by using the brand ‘isod’ or higher in the FileTypeBox. The MovieBox is not mandatory when using dynamic tracks, but in that case the first movie fragment loaded shall have a FileTypeBox or ExtendedTypeBox indicating support for dynamic track.

EDITOR’S NOTE: we could make FileTypeBox optional in this case but have SegmentTypeBox mandatory.

For DynamicMovieBox the following flags are defined:

* 0x000001 source-info-present if set, indicates that source information is present; if not set, source\_id, bundle\_id and source\_flags take the value 0.
* 0x000002 in-splice if set, indicates that the tracks described in the DynamicMovieBox correspond to a content splice period and will soon move back to previous configuration. By monitoring this flag and the source\_id field, the processing media pipeline can be optimized if desired (e.g. avoid unloading/reloading decoder resources for instance). This flag shall not be set if source\_id is 0.

EDITOR’S NOTE: we could also use a flag in 'styp' to signal this, to simplify edition of files during concatenation

###### Syntax

aligned(8) class DynamicMovieBox extends FullBox('dymv', version=0, flags){   
 if (flags & 1) {  
 unsigned int(32) source\_id;  
 unsigned int(32) bundle\_id;   
 unsigned int(24) source\_flags;  
 } else {  
 source\_id=0;  
 bundle\_id=0;  
 source\_flags=0;  
 }   
 DynamicTrackBox track; // optional: zero or more  
 UserDataBox user\_data; // optional: zero or one   
 MetaBox meta; // optional: zero or one   
}

###### Semantics

source\_id identifies the origin of the fragment. The value 0 indicates that the source is the MovieBox andthe DynamicMovieBox modifies the MovieBox. Other values identify another source than the MovieBox and all tracks, UserDataBox, MetaBox and any other properties defined in the MovieBox shall be ignored.

bundle\_id provides an identifier for tracking partial configuration changes for a given source\_id.

NOTE: The bundle\_id is typically needed when two or more dynamic tracks are inserted in initial movie or during a splice period, each in their own movie fragment. This allows a file reader to detect that the changes advertised for a given bundle have already been processed in a preceding movie fragment.

source\_flags identify the modifications declared in this DynamicMovieBox compared to the previous DynamicMovieBox with the same value of source\_id and bundle\_id. The following flags are defined:

* 0x000001 if set, indicates that one or more tracks configurations have changed
* 0x000002 if set, indicates that the global (MovieBox-level) user data has changed
* 0x000004 if set, indicates that the global (MovieBox-level) meta box has changed
* 0x800000: if set, indicates that the modifications are functionally equivalent to the previous DynamicMovieBox with the same source\_id and bundle\_id. When this flag is set, a file reader may safely skip processing the DynamicMovieBox if a previously parsed DynamicMovieBox has the same source\_id and bundle\_id. Otherwise (this is the first DynamicMovieBox parsed with this source\_id and bundle\_id values), the flag may be set but shall be ignored (i.e. considered as not set) by file readers.

When source\_flags is not set (either explicitly or per the above rule) or has the value 0, the box shall not be skipped. In this case, there is no information regarding modifications of child boxes compared to previous DynamicMovieBox; and the entire content of the box must be re-evaluated.

When source\_flags is not set to 0x000001, any DynamicTrackBox present in this DynamicMovieBox shall have the 0x800000 modification\_flags set.

##### Dynamic track box

###### Definition

Box Type: 'dytk'  
Container: DynamicMovieBox  
Mandatory: No  
Quantity: Zero or more

A DynamicTrackBox declares a new track or modifies an existing track for the duration of the parent movie fragment.

Tracks declared by a DynamicTrackBox, and for which

* the associated source\_id is not 0,
* or there are no tracks with a matching track\_ID in the MovieBox

implicitly declare a TrackExtendsBox with the value default\_sample\_description\_index set to 1 and the values default\_sample\_duration, default\_sample\_size, default\_sample\_flags set to 0.

NOTE This implies that default values will likely need to be set in the TrackFragmentHeaderBox; if multiple track fragments are used for a dynamic track within one movie fragment, the default values may need to be re-coded for each track fragment.

###### Syntax

aligned(8) class DynamicTrackBox extends Box('dytk'){  
 DynamicTrackHeaderBox dyn\_tkhd;//mandatory, must be first  
 SampleDescriptionBox stsd;//conditionally mandatory  
 Box minf\_header\_info; //optional, one of vmhd, smhd, sthd, hmhd…  
 DataInformationBox data\_info; //optional  
 UserDataBox udta; //optional  
 MetaBox meta; //optional  
 TrackReferenceBox tref; //optional  
 TrackGroupBox trgr; //optional  
}

###### Semantics

data\_info if present, gives the source(s) of the samples’ data for this dynamic track. If not present, the samples’ data is present in the container.

minf\_header\_info if present, gives the media-specific header box usually found in the MediaInformationBox of a track with the same handler\_type as this dynamic track. Derived specification may mandate its presence.

Other boxes contained in the DynamicTrackBox (except DynamicTrackHeaderBox) have unchanged semantics. When present, they replace their counterpart boxes in the TrackBox (and children) of the MovieBox.

##### Dynamic track header box

###### Definition

Box Type: 'dtkh'  
Container: DynamicTrackBox  
Mandatory: Yes  
Quantity: One

A DynamicTrackBox can be used to disable an existing track from the MovieBox, override the definition of an existing track from the MovieBox or define a completely new track.

NOTE The DynamicTrackBox is a compaction of TrackHeaderBox, EditListBox and MediaHeaderBox in a single container, in order to keep the track signalling overhead low.

The following flags are defined for a dynamic track header box:

0x000001 dyn\_tk\_ignore\_track indicates, when set, that the track declared in the MovieBox or in a previous DynamicMovieBox with the same source\_id as the parent DynamicMovieBox should be ignored (treated as if not present) until a next movie fragment is received for this track.

Flag values 0x000002, 0x000004, 0x000008, 0x000010, 0x000020 and 0x000040 are used for box parsing. Other values are reserved.

If flag dyn\_tk\_ignore\_track is not set in DynamicTrackHeaderBox, the parent DynamicTrackBox overrides an existing track or declares a new track. In this case:

* If stsd is present in the parent DynamicTrackBox: if there is a track with same ID in the MovieBox, overwrite it with current track, otherwise add the new track to the presentation

NOTE1: derived specification can mandate that the handler type / timescale / width/height remain the same in this case

NOTE2: This is typically used to update a sample description for a track

* If stsd is not present in the parent DynamicTrackBox : there shall be a track with same ID in the MovieBox and all fields in DynamicTrackHeaderBox shall match their counterpart fields in the track/handler/media header boxes declared in the MovieBox. This is used to update UserDataBox, MetaBox, TrackReferenceBox, TrackGroupBox of the track. The sample descriptions of the track remain unchanged in this case.

EDITOR’S NOTE: we could also use a dedicated flag for the case where no 'stsd' is present, in order to avoid re-listing things that shall match what is in the 'moof'’s track

###### Syntax

aligned(8) class DynamicTrackHeaderBox extends FullBox('dtkh', version=0, flags){  
 unsigned int(32) track\_ID;  
 if (!(flags & 1)) {  
 if (flags & 2)   
 unsigned int(24) modification\_flags;  
 unsigned int(32) handler\_type;  
 unsigned int(32) media\_timescale;  
 if (flags & 4)   
 signed int(32) delay;  
 if (flags & 8)   
 unsigned int(24) track\_flags; //as in TrackHeaderBox  
 bit(1) lang\_3cc;  
 if (lang\_3cc) {  
 unsigned int(5)[3] language; // ISO-639-2/T language code  
 } else {  
 bit(7) reserved;  
 utf8string extended\_language

}  
 int(16) alternate\_group;  
 if (flags&16) {  
 unsigned int(32) width;  
 unsigned int(32) height;  
 int(16) layer;  
 if (flags&32) {  
 int(32)[9]matrix ;  
 }  
 }  
 if (flags&64) {  
 int(16)volume;  
 }  
 }  
}

###### Semantics

track\_ID indicates the ID for the track declaration

modification\_flags identify the modifications in the parent DynamicTrackBox compared to the previous DynamicTrackBox with the same value of track\_ID and the same value of source\_id in the parent DynamicMovieBox. These flags may be used by the reader to optimize processing of consecutive track fragments with the same track\_id and source\_id. The following flags are defined:

* 0x000001: the track configuration has changed (changes in one or more fields other than modification\_flags in DynamicTrackHeaderBox, or changes in the associated media header box (‘vmhd’, ‘smhd’, etc.) or changes in the associated data information box)
* 0x000002: the media configuration has changed (new sample description, ‘stsd’)
* 0x000004: the track UserDataBox has changed
* 0x000008: the track MetaBox has changed
* 0x000010: the track TrackReferenceBox has changed
* 0x000020: the track TrackGroupBox has changed
* 0x800000: the track modifications are functionally equivalent to the previous DynamicTrackBox with the same track\_id and source\_id. This flag may be set but shall be ignored (i.e. considered as not set) by file readers when the previous DynamicMovieBox has a different source\_id or when this is the first DynamicTrackBox parsed for this track\_id and source\_id.

When modification\_flags is not set (either explicitly or per the above rule) or has the value 0, there is no information available on possible changes of the track compared to previous DynamicTrackBox with the same track\_id and source\_id and the entire content of the DynamicTrackBox must be re-evaluated.

handler\_type same as handler\_type in HandlerBox

media\_timescale same as timescale in MediaHeaderBox

delay indicates media delay of the track, in media\_timescale. The presentation time of any sample in the track is the sum of the composition time of the sample and this value. A negative presentation time indicates that (part of) the sample data shall not be presented (media skip). If not coded, the value 0 is used.

track\_flags same as flags in TrackHeaderBox. If this field is not coded:

* If source\_id of the dynamic track is 0 and there is a matching track with the same track\_ID in the MovieBox, the flag values from the TrackHeaderBox are used,
* Otherwise, this field value is inferred to be 0x000003

language same as language in MediaHeaderBox

extended\_language same as in ExtendedLanguageBox

alternate\_group same as in TrackHeaderBox

width same as in TrackHeaderBox. If not coded, the media width, after pixel aspect ratio and clean aperture applied, is used

height same as in TrackHeaderBox If not coded, the media height, after pixel aspect ratio and clean aperture applied, is used

layer same as in TrackHeaderBox. If not coded, the layer is 0

matrix same as in TrackHeaderBox. If not coded, the identity matrix is used

volume same as in TrackHeaderBox. If not coded, full volume (1.0) is used

### Examples

### Sample description update

When a sample description changes for fragments N to K, this can be expressed as:

* Fragment N->K :
  + 1 DynamicMovieBox with source\_id=0, (bundle\_id= any), source\_flags=0x800000 (or 0x800001 as a hint).
  + 1 DynamicTrackBox with one stsd
  + DynamicTrackHeaderBox with track\_ID of the updated track and no modification\_flags or
    - modification\_flags=0x800002 if no track layout change
    - modification\_flags=0x800003 if track layout change
* Fragment K+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Track addition

When **a new track appears** for fragments N to K, this can be expressed as:

* Fragment N->K :
  + 1 DynamicMovieBox with source\_id=0 (bundle\_id= any), source\_flags=0x800000 (or 0x800001 as a hint).1 DynamicTrackBox with one stsd
  + DynamicTrackHeaderBox with new track\_ID (different from all previous tracks due to source\_id=0) with
    - modification\_flags=0x800000
* Fragment K+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Splicing AV contained in a single movie fragment

When splicing AV content with T tracks contained in a single movie fragment for fragments N to K, this can be expressed as:

* Fragment N->K :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id>0, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (any track\_ID) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800000
* Fragment K+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Splicing AV contained in separate movie fragments

When splicing AV content contained in separate movie fragments for fragments Ni to Ki, i representing each media, this can be expressed as:

* Fragment Ni->Ki :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id>0, bundle\_id=***i***, source\_flags=0x800000 (or 0x800001 as a hint)
  + 1 DynamicTrackBox (any track\_ID) with one stsd
    - DynamicTrackHeaderBox with modification\_flags=0x800000
* Fragment Ki+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Splicing with configuration changes in a single movie fragment

When splicing AV content contained in a single movie fragment for fragments N to K with codec configuration change happening from C1 to C2 (N < C1 < C2 < K), this can be expressed as:

* Fragment N->C1 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (any track\_IDs called *splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800000
* Fragment C1 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x000001
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x000002
* Fragment C1+1->C2 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800002
* Fragment C2+1 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x000001
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags= 0x000002
* Fragment C2+2 -> K :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800002
* Fragment K+1 :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Efficiency

With this proposal, signaling a new track in simple cases costs STSD plus:

* 12 bytes for DynamicTrackHeaderBox header (cost of the FullBox)
* 17 bytes for audio (no volume), 27 bytes for video with size/layer and without track matrix (63 with matrix)
* 8 bytes for DynamicTrackBox header (cost of the Box)
* 12 bytes for DynamicMovieBox header

Overhead for 1s fragments is 0.392 kbps for audio, 0.47 kbps for video (0.76 kbps if full visual matrix is specified) hence roughly 1.3 kbps for full signaling (i.e. track and sample descriptions (stsd)) versus 4+ kbps for TrackBox injection as proposed in m58085.

Signaling removal of a track costs 12+4+8+12 bytes, 0.288 kbps

### Usage with dependent movie fragments (defined in section16.1)

The proposed definition for “dependent movie fragments” can be updated when combined with dynamic movie fragment as follows:

If version is not 0, any SampleGroupDescriptionBox, DynamicMovieBox or MetaBox defined in the last movie fragment, or in the last TrackFragment in the last movie fragment, whose MovieFragmentHeaderBox version is 0 also apply for this movie fragment, and there shall not be any SampleGroupDescriptionBox, DynamicMovieBox or MetaBox defined for this movie fragment.

# Support for Haptics

This topic comes from the discussions related to the input contribution m61136:

<http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/157>

This section contains the parts that were not merged in the Merged Carriage of Haptic Streams in ISOBMFF (m61418). The related issue is:

<http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/160>

## Grouping Haptics Tracks for the Same Perception

When multiple Haptics Tracks are used to carry the band data of the various channels of the haptics experience perceptions, track grouping shall be used to identify which Haptic Tracks are associated with a certain haptics perception. This is done by defining a new track group type by extending TrackGroupTypeBox defined in ISO/IEC 14496-12 which contains a track\_group\_id that represents an identifier for the track group and a track\_group\_type field which stores a four-character code identifying the group type. The pair of track\_group\_id and track\_group\_type identifies a track group within the container file.

A new HapticsTrackGroupBox is defined as follows.

Box Type: 'hptg'  
Container: TrackGroupBox  
Mandatory: No  
Quantity: Zero or more

aligned(8) class HapticsTrackGroupBox extends TrackGroupTypeBox('hptg') {  
 unsigned int(32) perception\_id;

}

# Event-based media signaling

This topic comes from the discussions related to the input contribution m61071 at MPEG#140 :

[m61071 [Haptics] [7.2] [ISOBMFF] Event Based Media Signalling for Haptics Media in ISOBMFF (#145) · Issues · MPEG / Systems / File Format / ISOBMFF · GitLab (int-evry.fr)](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/145)

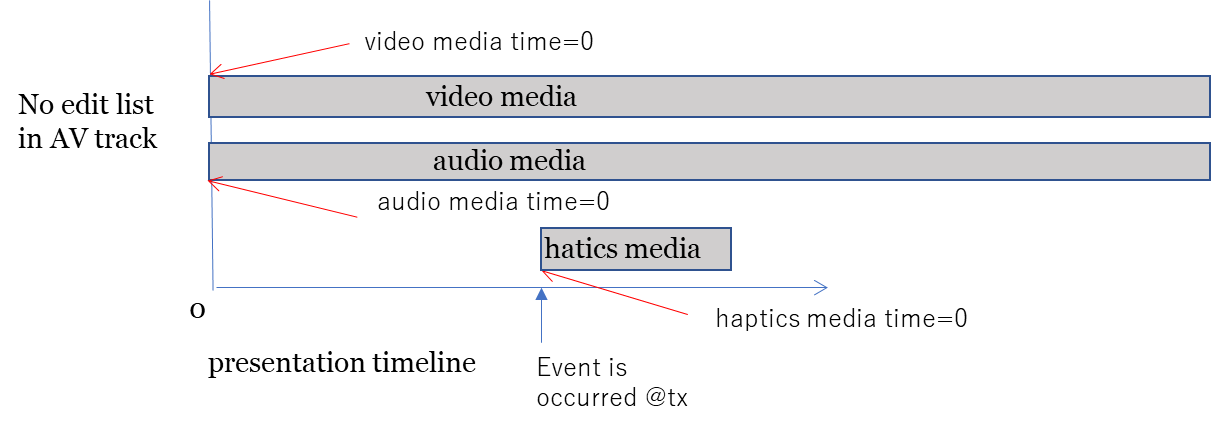
Updated at MPEG#141 from the discussions to the input contribution m61794:

[m61794 [Haptics] [7.2] [ISOBMFF]On Event-Based Media Signalling for Haptics (#167) · Issues · MPEG / Systems / File Format / ISOBMFF · GitLab (int-evry.fr)](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/167)

## Problem statement about Event-based Haptics media play back

In the 6DoF scene, 3D video and audio are played back in presentation timeline. When an event such as a collision between user and 3D object occurs at any time in presentation timeline, event-based haptics media would be started playing back in the presentation timeline at the above event time.

In ISO Base Media File Format, this type of event-based media signalling is not defined yet. If no event-based media signalling, event-based media might be interpreted as time-based media and be played back synchronously with the other medias. To avoid this kind of miss interpretation in File Format layer, this event-based media should explicitly be signalled in ISO Base Media File Format.



## Two alternatives were proposed for study at MPEG#140 :

1. Reuse the ‘sync’ track reference from MPEG-4 Systems, allowing a track to have a ‘sync’ reference to itself, 0 or with no listed track\_IDs to indicate it runs on its own timeline. The question of allowing two-way ‘sync’ track references (as in QuickTime) was raised.

2. Extend edit-list (as suggested in m61071 and described hereafter)

## The combination of two solutions was proposed for study at MPEG#141

### Considerations for two alternative proposal at MPEG#141

Followings are the considerations on two alternative proposals.

(1)Edit List extension

The Edit list could be used to map a media in a track into the Presentation timeline for event-based media

**Pros:**

This is an explicit information what playback is required for event-based media track.

**Cons:**

No disadvantage. If the functionality of edit list extension is as same as sync track reference with the value 0, no new definition is needed.

(2) sync track reference

‘sync” track reference with the value 0 indicates that this track is not synchronized with other tracks.

**Pros:**

When indicating a relationship among multiple tracks, using a Track reference fits the file format design principle. And this has already defined in MP4 File Format.[4]

**Cons:**

This indicates that it is not synchronized with other tracks, but it is not explicit what playback is required.

In MP4 file format[4], the specification defines what playback is required as follows

* If a track (stream) contains a track reference of type ‘sync’ whose value is 0, then the hinter or server shall set the OCRStreamFlag field in the MPEG-4 ESDescriptor to FALSE and shall not insert any OCR\_ES\_ID field. This means that this stream is not synchronized to another, but other streams may be synchronized to it.

### The combination proposal of two solutions

Based on the consideration in section 3, reuse of sync track reference with the value 0 seems to be better as file format design principle. However, it is not enough information for event-based media signalling. An edit list extension compensates the disadvantage of sync track reference with the value 0 for event-based media signalling. Hence, the combination of two solution would be the best solution for event-based media signalling as follows.

(1) Reuse the sync track reference with the value 0

To indicate that this track plays on its timeline

Then, client could determine that this track shall not be played back with other tracks.

(2) Extend edit list flags with the value 2

To indicate how this track plays on its timeline

Then, client could determine that this track shall be played back by the interactive condition.

### List of questions for the combination proposal of two solutions

Following list is extracted from the discussion in the issue#167

(1) What is the expected behavior in legacy players?

(2) Could Haptics File Format define a specific major brand to ignore a track for legacy player?

(3) Does edit list extension really give explicit information? It seems to be no additional information to sync track reference with the value 0.

(4) This proposal does not seem to be specific to Haptics. Event-based signalling for audio/video/text tracks might be carefully considered.

(5) How many tracks for a event? Multiple tracks for a event? Single track is for a event typically. So, the number of events might equal to the number of tracks.

## Proposed Text for event-based media signalling at MPEG#141

Yellow highlight: same as the proposal in m61071 and TuC[2]

Blue highlight: addition to m61071

**New terminology definitions**

**Asynchronous media**

is a media in a track, which intends to be played back on its timeline.

**Event-based media**

Is a media in a track, which intends to be played back when an event is occurred at any time within presentation timeline. E.g., An event may be an interaction between user avatar and 3D object in a 6DoF scene.

**Reuse sync’ track reference defined in MP4 File Format**

Add new track reference type into semantics of track reference box in ISOBMFF[3]

Following text is already defined in MP4 file format[4]. And some updates for event-based media signalling.

'sync' this track uses the referenced track as its synchronization source. Track reference of type 'sync' with the value 0 indicates that a media in this track is an asynchronous media.

**Edit list box extension**

The definitions of flags in edit list box in 8.6.6.3 in ISOBMFF are as follows

1st bit was defined for the function to repeat media in HEIF specification.

**flags**

the following values are defined. The values of flags greater than 1 are reserved.

RepeatEdits 1

Additional function to signal event-driven media in reserved flags is proposed as follows:

**flags**

the following values are defined. The values of flags greater than 3 are reserved.

RepeatEdits 1

EventBasedMedia 2 /\*This indicate 2nd bit in flags\*/

flags&2 equal to 0 (EventBasedMedia =0) indicates that media in this track is time-based media and presents in presentation timeline as described in edit list entry

flags&2 equal to 1 (EventBasedMedia =1) indicates that media in this track is event-based media and presents in presentation timeline as described in edit list entry when an event is occurred at any time within presentation timeline. This event condition may be defined in upper layer specification and/or handled by the application.

Note that when flags&2 equal to 1 (EventBasedMedia =1), track reference of type 'sync' with the value 0 shall be present in this track.

Note: EventBasedMedia bit in flags shall equal to 0 for Haptics phase 1 codec to be synchronized with Audio, Video, and other medias.

This note should be in the text “Carriage of haptics codec in ISO base media file format”.

(comment to be considered)

New version=2 in this edit list box might be considered to distinguish this extension.

# Use of Preselection for signaling Picture in Picture in ISOBMFF

The “DASH Manifest writing Guidelines based on ISOBMFF” (WG03 N00569) uses the track groups to signal the preselection sets in ISOBMFF. This section of the ISOBMFF TuC indicates that **the File Format group is now considering 2 options for PiP and welcome contributions**:

* Option 1: as discussed in m61182, we have a base PreselectionProcessingBox and derived processing boxes for audio, video ... and the region id goes into the video one. The existing one becomes audio. details to be worked out
* Option 2: use track references to indicate PiP relationship (move only the generic part from part 15 to part 12)

Reference to discussions:

<http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/140>

## Picture in Picture use cases (from m61182)

The solution should indicate how two or more tracks can be used for the PiP experience, i.e. the decoded pictures of the tracks can be overlayed over a region of the decoded pictures of the other track(s). The following information needs to be signaled:

1. The combination of tracks creates a PiP experience.
2. The main and susbsitute tracks are identified.
3. Allow different grouping of a track with other tracks, e.g.,
   * PiP1: Track 2 can be overlayed on Track 1
   * PiP2: Track 3 can be overlayed on Track 1
4. Tracks can be
   * 1. Independently decodable tracks, i.e. to be decoded by independent decoders
     2. The main picture may consist of multiple independent tracks, each of which is independently decodable or then can be merged to decode together.

The main picture may be a single track consisting of substitutable subpictures

## Implementation of the PiP using preselection track groups

Figure below demonstrates the use of the preselection track group for signaling PiP:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | **tkhd** |  |  |  |  | **tkhd** |  |  |  | **tkhd** |  |  |  |
|  |  |  |  |  | trak\_id =1 |  |  |  |  | trak\_id =2 |  |  |  | trak\_id =3 |  |  |
| **mvhd** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **tkgd** |  |  |  | **trgr** |  |  |  |  | **trgr** |  |  |  | **trgr** |  |  |  |
|  | **prse** |  |  |  | **pres** |  |  |  |  | **pres** |  |  |  | **pres** |  |  |
|  |  | track\_group\_id =1 |  |  |  | track\_group\_id =1 |  |  |  |  | track\_group\_id =1 |  |  |  | track\_group\_id =2 |  |
|  |  | num\_tracks =2 |  |  | **pres** |  |  |  |  |  |  |  |  |  | **prsp** |  |
|  |  | Preselection\_tag = “pip” |  |  |  | track\_group\_id =2 |  |  |  |  |  |  |  |  |  | track\_order=1 |
|  |  | **kind** |  |  |  | **prsp** |  |  |  |  |  |  |  |  |  |  |
|  | **prse** |  |  |  |  |  | track\_order=0 |  |  |  |  |  |  |  |  |  |
|  |  | track\_group\_id =2 |  |  |  |  | sample\_merge\_flag =1 |  |  |  |  |  |  |  |  |  |
|  |  | num\_tracks =2 |  |  |  |  | region\_ids =’1’ |  |  |  |  |  |  |  |  |  |
|  |  | **kind** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| trak |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| trak |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| trak |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

* The “prse” box in the track group indicates the existence of a PIP experience:
  + track\_group\_id is used as the identifier.
  + num\_tracks indicates the total number of components in this PiP preselection.
  + The kind box includes the DASH Role scheme with “pip” value.
* The “pres” box in each track indicates that the track is a part of one PiP experience.
  + trackgroup\_id identifies the corresponding ‘prse’ box.
  + The ‘prsp’ box defines the processing of PiP:
    - track\_order =0 defines the main PIP picture, while track\_order>0 defines the substitute PiP picture. If the main PiP consists of multiple tracks, all those tracks must set their track\_order to 0.
    - sample\_merge\_flag =1 defines that the merge with this track is possible for single decode.
    - region\_ids provides the list of region ids in the track that can be replaced with the substitute PiP picture. This field is ignored when track\_order>0.

Note: Use of sample\_merge\_flag is not needed for PIP. A region\_ids with a value of NULL can also indicate that the track doesn’t have replaceable regions.

In the above example, two PiP experiences are defined as the following:

* Preselection with track\_group\_id =1 defines a PiP experience with two independent decodes.
* Preselection with track\_group\_id =2 defines a PiP experience with the video possibility of substituting the subpicture streams of the main PiP picture with the substitute picture stream.

## Proposed changes in ISOBMFF text (from m61182, yellow-highlighted)

Preselection processing box

Definition

Box Type: 'prsp'  
Container: PreselectionGroupBox  
Mandatory: No  
Quantity: Zero or one

This box contains information about how the tracks contributing to the preselection can be processed. Media type specific boxes may be used to describe further processing.

Syntax

aligned(8) class PreselectionProcessingBox   
 extends FullBox('prsp', version=0, flags ){   
 unsigned int(8) track\_order;  
 unsigned int(1) sample\_merge\_flag;  
 unsigned int(7) reserved;

utf8string region\_ids;  
 // further attributes and Boxes defining additional processing of  
 // the track contributing to the preselection  
}

Semantics

track\_order defines the order of this track relative to other tracks in the preselection as described below.

sample\_merge\_flag equal to 1 indicates that this track is enabled to be merged with another track as described below.

region\_ids specifies a list of white-spaced IDs for the coded data units that their corresponding substreams can be replaced with the other tracks in this preselection. A NULL string means that no region is replaceable. This field can have a non-NULL value only if track\_order =0. The use of this field is described below.

[Editor’s Note(MPEG#140): *The contribution proposes to add the region\_ids to the PreselectionProcessingBox, but these will never be used in the audio context. We propose to have a generic version of the PreselectionProcessingBox from which specific audio and pip versions can be inherited instead*.]

For picture-in-picture applications, every track that is a part of the main picture shall have its track\_order=0. Any track in the picture-in-picture applications that is intended to be used as an overlay or substitute of one or more regions of the main picture (known as the substitute picture) shall have its track\_order to be set with a value higher than 0. A lower track\_order value indicates the higher priority to be used as a substitute picture in the main picture.

Sample entry specific specifications might require the tracks for a preselection to be provided to the respective decoder instances in a specific order. Since other means, such as the track\_id, are not reliable for this purpose, the track\_order is used to order tracks in a preselection relative to each other. A lower number indicates that at a given time the sample of the containing track is provided to the decoder before the sample with the same given of other tracks with higher number. If two tracks in a preselection have their track\_order set to the same value or if the preselection processing box is absent for at least one of the tracks, the order of these tracks is not relevant for the preselection, and samples can be provided to the decoder in any order.

A merge group is defined as a group of tracks, sorted according to track\_order, where one track with the sample\_merge\_flag set to 0 is followed by a group of consecutive tracks with the sample\_merge\_flag set to 1. All tracks of a merge group shall be of the same media type and shall have all samples time-aligned.

If the sample entry type is associated with a codec-specific process to merge samples of a preselection, this process shall be used.

The combination of sample\_merge\_flag = 1 and a non-NULL region\_ids indicates that any of the coded data units represented by the ids in the region\_ids can be replaced by the other tracks with a track\_order value larger than 0. The concrete semantics of the region IDs need to be explicitly specified for specific codecs.

NOTE 1 If the tracks in the merge group are all of sample entry type of “mhm2” (MPEG-H 3D Audio), the merging process is defined in ISO/IEC 23008-3:2019, subclause 14.6.

NOTE 2 Tracks in a merge group may have different sample entry types.

If the sample entry type is not associated with a codec-specific process to merge samples of a preselection and when region\_ids is NULL, then the following process shall be used:

Merging within the merge group shall proceed by forming tuples of track samples with the same time stamp across contributing tracks. The ordering of samples within the tuple shall be determined by track\_order. These tuples shall be formed by byte-wise concatenation of the samples resulting in a single sample with having the respective time stamp assigned. If generation of new tracks is targeted, each merge group shall result in a separate output track conformant to a media type derived from the media types of the merged tracks.

For tracks not part of a merge group, a merging process is not specified by the present document.

##### Track group entry definitions

* + - * 1. Preselection track group entry box

Definition

Box Type: 'prse'  
Container: TrackGroupDescriptionBox  
Mandatory: No  
Quantity: Zero or More

Preselections can be qualified, for example, by language, kind or media specific attributes like audio rendering indications, audio interactivity or channel layouts. Attributes signalled in a PreselectionTrackGroupEntryBox take precedence over attributes signalled in contributing tracks.

PreselectionTrackGroupEntryBox shall describe only track groups identified by track\_group\_type equal to 'prse'.

All preselections with at least one contributing track having the track\_in\_movie flag set to 1 shall be qualified by PreselectionTrackGroupEntryBoxes. Otherwise, the presence of the PreselectionTrackGroupEntryBoxes is optional.

All attributes uniquely qualifying a preselection shall be present in PreselectionTrackGroupEntryBox of the preselection.

Syntax

aligned(8) class PreselectionTrackGroupEntryBox  
 extends TrackGroupEntryBox('prse', version=0, flags)  
{  
 unsigned int(8) num\_tracks;  
 utf8string preselection\_tag;  
 if (flags & 1) {  
 unsigned int(8) selection\_priority;  
 }  
 if (flags & 2) {  
 unsigned int(8) segment\_order;  
 }  
 // Boxes describing the preselection  
}

Semantics

This box contains information on what experience is available when this preselection is selected.

Boxes suitable to describe a preselection include but are not limited to the following list of boxes defined in this document:

* AudioElementBox (subclause 12.2.9**Error! Reference source not found.**)
* AudioElementSelectionBox (subclause 12.2.13**Error! Reference source not found.**)
* ExtendedLanguageBox (subclause 8.4.6**Error! Reference source not found.**)
* UserDataBox (subclause 8.10.1**Error! Reference source not found.**)
* KindBox (subclause 8.10.4)
* LabelBox (subclause 8.10.5)
* AudioRenderingIndicationBox (subclause 12.2.8)
* ChannelLayout (subclause 12.2.4)

If a UserDataBox is contained in a PreselectionTrackGroupEntryBox, then it shall not carry any of the above boxes.

num\_tracks specifies the number of non-alternative tracks grouped by this preselection track group.

A track grouped by this preselection track group is a track that has the 'pres' track group with track\_group\_id equal to the ID of this preselection.  
The number of non-alternative tracks grouped by this preselection track group is the sum of the following:

* the number of tracks that have alternate\_group equal to 0 and are grouped by this preselection track group,
* the number of unique non-zero alternate\_group values in all tracks that are grouped by this preselection track group.

The value of num\_tracks shall be greater than or equal to the number of non-alternative tracks grouped by this preselection track group in this file.  
A value equal to 0 indicates that the number of tracks grouped by this track group is unknown or not essential for processing the track group.  
NOTE 1 The value of num\_tracks can be greater than the number of non-alternative tracks containing a PreselectionGroupBox with the same track\_group\_id in this file when the preselection is split into multiple files.  
NOTE 2 When a player has access to fewer non-alternative tracks grouped by this preselection track group than indicated by num\_tracks, the player might need to omit the tracks grouped by this preselection track group.

preselection\_tag is a codec specific value that a playback system can provide to a decoder to uniquely identify one out of several preselections in the media.

selection\_priority is an integer that declares the priority of the preselection in cases where no other differentiation such as through the media language is possible. A lower number indicates a higher priority.

segment\_order specifies, if present, an order rule of segments that is suggested to be followed for ordering received segments of the Preselection. The following values are specified with semantics according to ISO/IEC 23009-1:2022, subclause 5.3.11.5:

0: undefined

1: time-ordered

2: fully-ordered

Other values are reserved. If segment\_order is not present, its value shall be inferred to be equal to 0.

NOTE 3 Not all tracks contributing to the playout of a preselection may be delivered in the same file.

NOTE 4 The kind box might utilize the Role scheme defined in ISO/IEC 23009-1:2022, subclause 5.8.5.5 as it provides a commonly used scheme to describe characteristics of preselections.

NOTE 5 This box carries information about the initial experience of the preselection in the referenced tracks. The preselection experience can change during the playback of these tracks, e.g., audio language can change during playback. These changes are not subject to the information presented in this box.

Further media type specific boxes may be used to describe properties of the preselection. Readers may ignore and skip boxes that are not recognized.

A KindBox may be used to signal the picture in picture experience, by using the DASH Role schemeIdURI (ISO/IEC 23009-1) and the value “pip”, meaning that one or more regions of the main picture can be overlayed with one or more substitute pictures.

# Attaching collection timestamps and large metadata to every samples

This topic comes from the discussions related to the input contribution m61022:

http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff#143

## Scope:

Certain motion imagery applications require knowledge of the absolute time-of-day (referenced to start of exposure) for each image in a sequence. GEOINT and time-space-position-information (TSPI) measurement applications, security video, sporting events, and scientific testing are common examples. The specific time related to a media sample collection time (the start of exposure of an image, capture of a timed metadata item in a timed metadata track, and capture of an audio sample), is referred to in this section as the “collection time”.

This section proposes to:

* + - * + Provide basic use cases where this capability is a requirement.
        + Define the requirements for information to be included in the box.
        + Describe a Collection Timestamp box (‘colt’) inside the Sample Table box, which provides for close coupling of the timestamps with their image frames.
        + Include quality and status information in the form of static and dynamic metadata providing information on the clock source, lock status of the source, and accuracy of the timestamps.

Timing information in a video file provides utility for a variety of user purposes. At a fundamental level, timing is required to understand:

* + - * + the ordering of samples,
        + the time difference between samples (frame rate or sample rate),
        + the consistency of time between samples (jitter),
        + the presence of dropped/lost samples,
        + for video, the timing of pixel information within a frame (global vs rolling shutter, push-broom sensors, etc.),
        + the temporal correlation between individual samples in separate tracks or files

## Use cases

* Correlation of image frames taken from different locations but looking at the same scene content (3D analysis)
* Correlation of sample information, in real-world time, for different media types
* Correlation of scene content to real world events and information derived from other sensors
* Confirmation of event times and sequences in one or more security cameras
* Confirmation of an event time or sequence at a sporting event, race, etc.
* Temporal correlation of still media with timed media

Ex: Correlation of a phone camera still image with a security camera video frame

Ex: Correlation of an image frame with a motion sensor trigger

## Requirements

* Provide the ability to annotate each sample (motion imagery, metadata, audio, etc.) with a high precision and high accuracy collection timestamp.
* Provide sufficient metadata to determine the quality and accuracy of the timestamps.
  + - * + Timestamps in the Timestamp box are collection timestamps, representing the “real-world” time a sample is generated. The values are not intended to be modified once captured during initial recording into a file unless the modifications are part of an intended process to refine and improve the accuracy of the timestamp information.
* Timestamps are based on “atomic” time standards, such as TAI, GPS, etc. Conversion of timestamp information to UTC is enabled in user applications through the ability to carry leap seconds information in the metadata.

## Analysis of Alternatives

To implement a generalized collection timestamp capability, a few alternative approaches are examined:

1. Items associated to samples
   * One metadata item is created to carry the static collection timestamp information for the entire track
   * Additional metadata items are created, one for each track sample, to carry the time stamps and associated dynamic time quality metadata.
   * The items are referenced to each individual sample in a track.

Pros:

* + The method is straightforward conceptually.
  + There are defined locations for both the static information and the dynamic information.
  + The method leverages existing format capabilities, although an encoding method for the timestamps and quality metadata will need to be selected.

Cons:

* + There is a significant amount of overhead associated with creating the large number of items necessary to annotate each individual sample in a track. Security cameras, for instance, may record continuously for many hours or days.
  + The timestamps are loosely coupled with the track through referencing and may get handled improperly, or possibly get dropped during editing and file export operations.

1. Creation of a ‘collection timestamp’ metadata track, time-synchronized with another video/media track
   * In this approach, individual metadata items, that carry the collection timestamp information, are generated and referenced to each individual sample in a track.
   * Static metadata for the collection timestamps is carried in the static metabox of the track.
   * Collection timestamps and associated dynamic time quality metadata are carried in the track samples
   * The track is referenced to the track carrying the associated video/media samples
   * The composition timeline is used to correlate the collection timestamps with the individual image frames

Pros:

* + The method is straightforward conceptually.
  + There are defined locations for both the static information and the dynamic information.
  + The method leverages existing format capabilities, although an encoding method for the timestamps and quality metadata will need to be selected.

Cons:

* + There is a fair amount of overhead associated with creating a separate track to carry just the timestamps.
  + The timestamps are loosely coupled with the track through referencing and may get handled improperly, or possibly get dropped during editing and file export operations.

1. Utilization of the Sample Auxiliary Information Sizes/Offsets boxes (‘saiz’/’saio’) to carry the ‘collection timestamp’ metadata information for a specific media track. See Annex A - Proposed Expanded ‘saiz’/’saio’ box solution for a detailed description.
   * In this approach, the timestamp information is carried using expanded versions of the existing ‘saiz’ and ‘saio’ boxes inside the Sample Table box
   * Static collection timestamp metadata is carried

Pros:

* + This leverages an existing capability within 14496-12
  + The timestamps are included within the track structure of the media they apply to, resulting in a tighter coupling

Cons:

* + The ‘saiz’/’saio’ boxes have some constraints on placing the data, based on matching the chunking of the track media, etc.
  + If the ‘saio’ and ‘saio’ boxes are not closely accounted for, or if the application doesn’t understand the data carried in the ‘saiz’/’saio’ boxes, editing and export operations might potentially corrupt the alignment of the timestamps with their respective image frames
  + The boxes need to be modified to support larger sample auxiliary information (although this then provides a benefit)

1. Utilization of T.35 messages to carry the collection timestamp information
   * In this approach, the timestamp information is carried using Supplemental Enhancement Information (SEI) message areas
   * Static collection timestamp metadata can be carried in the track metabox, which will require a definition and defined coding mechanism.

Pros:

* + Information is carried as part of the track
  + SEI messages can be generated for each frame

Cons:

* + Not all video and other media formats support T.35 messages

1. Implementation of a new Collection Timestamp (‘colt’) box. See Annex B - Proposed Collection Timestamp (‘colt’) box solution for a detailed description.
   * The box will be located in the Sample Table box
   * The top portion of the box contains static collection timestamp metadata, followed by the addressing for the dynamic sample-based timestamp information.
   * The dynamic information is allowed to be carried within the box or located externally to the box via offsets.

Pros:

* + The collection timestamp information is tightly coupled within the media track
  + The information is clearly defined and should be well understood

Cons:

* + This approach requires a new box

## Annex A - Proposed Expanded ‘saiz’/’saio’ box solution

Note: This solution is general in nature and provides additional capability beyond the carriage of collection timestamp information. It is proposed here in a form that supports the general solution.

### Scope:

* In a generic sense, expand the ability to attach metadata to each individual sample of a track using the ‘saiz’/’saio’ boxes.
* In a specific sense, use this capability to attach collection timestamp information to each individual sample of a track.
* The current capability is limited to 255 bytes of information.
* This contribution proposes to:
  + - * + Expand the size of a metadata block that can be attached to each frame.
        + Allow an application to signal the encoding format for frame metadata identified in the ‘saiz’ box
        + Allow for multiple instances of ‘saiz’/’saio’ boxes within a single track and provide IDs for each instance so the corresponding ‘saiz’ and ‘saio’ boxes can be determined.
        + Include an ID within the ‘saiz’ box to uniquely label the stream of synchronous metadata, which can then be available for referencing.
* As an additional observation, this approach might provide a mechanism for labeling user defined regions of interest within each frame:

The ROI’s may change from frame to frame

There may be more than one ROI for each frame

The number of ROIs may not be the same for all frames.

### Use Case Descriptions:

* Annotate each frame with collection timestamp and associated metadata (primary purpose here)
* Annotate image frames with information created through analysis
* Attached metadata to individual tracked objects across multiple frames. Maintain information on each region of interest location and size.
* Attach metadata to each pixel in an image, such as inclusion of latitude, longitude, and height above ellipsoid for each pixel.

### Requirements:

* Support frame metadata sizes using an addressing size of up to 64-bits.
* Support pixel level metadata, where there are metadata elements for each pixel in an image.
* Support metadata that is specific to each band or to all bands.
* Support metadata that is specific to a region of interest within a frame and/or within a band of the frame. The region of interest may change location and size from frame to frame. Provide for ROI IDs to provide traceability of regions over the course of a track.
* Provide a method for linking/referencing ‘saiz’/’saio’ metadata to other metadata in a file
* Provide a mechanism for applying an ID to an individual frame in a track

### Syntax:

aligned(8) class SampleAuxiliaryInformationSizesBox extends FullBox('saiz', version, flags){   
 if (flags & 1){   
 unsigned int(32) aux\_info\_type; // a registered FourCC code set to ‘misb’  
 unsigned int(32) aux\_info\_type\_parameter; // a coded 32-bit word   
 }  
 if (version == 0){   
 unsigned int(8) default\_sample\_info\_size; //size of synchronous metadata block (255 bytes //max), 0 if sizes vary  
 unsigned int(32) sample\_count; //number of samples in track  
 if (default\_sample\_info\_size == 0) {  
 unsigned int(8) sample\_info\_size[sample\_count]; // (255 bytes max)   
 }  
 }  
 else if (version==1){   
 unsigned int(32) saiz\_ID; //ID that uniquely identifies frame metadata stream  
 //need to determine if a matching ID is needed in the ‘saio’ box  
 ItemInfoEntry[1]; //Implement a generic method for signaling the metadata //encoding method  
  
 if(flags & 2){   
 unsigned int(64) default\_sample\_info\_size; //size of synchronous metadata block //0 when sizes vary  
 unsigned int(32) sample\_count; //number of samples in track  
 if (default\_sample\_info\_size == 0) { //signals sizes are different  
 unsigned int(32) sample\_info\_size[sample\_count]; //size of each synchronous //metadata block  
 }  
 }  
 else if(flags & 4){   
 unsigned int(32) default\_sample\_info\_size; //size of synchronous metadata block //(4GB max), 0 if sizes vary  
 unsigned int(32) sample\_count; //number of samples in track  
 if (default\_sample\_info\_size == 0) {  
 unsigned int(32) sample\_info\_size[sample\_count]; //size of each synchronous metadata //block (4GB max)   
 }  
 }  
 else if(flags & 8){   
 unsigned int(16) default\_sample\_info\_size; //size of synch metadata block //(65,535 bytes max), 0 if sizes vary  
 unsigned int(32) sample\_count; //number of samples in track  
 if (default\_sample\_info\_size == 0) {  
 unsigned int(16) sample\_info\_size[sample\_count]; //size of each synch metadata block //(65,535 bytes max)   
 }  
 }  
 else{  
 unsigned int(8) default\_sample\_info\_size; //size of synch metadata block (255 bytes //max), 0 if sizes vary  
 unsigned int(32) sample\_count; //number of samples in track  
 if (default\_sample\_info\_size == 0) {   
 unsigned int(8) sample\_info\_size[sample\_count]; //size of each synch metadata block //(255 bytes max)   
 }  
 }  
 }  
}

### Semantics

aux\_info\_type is an integer that identifies the type of the sample auxiliary information. At most one occurrence of this box with the same values for aux\_info\_type and aux\_info\_type\_parameter shall exist in the containing box.

‍aux\_info\_type\_parameter identifies the “stream” of auxiliary information having the same value of aux\_info\_type and associated to the same track. The semantics of aux\_info\_type\_parameter are determined by the value of aux\_info\_type.

saiz\_ID is an ID used to uniquely identify a specific ‘saiz’ metadata track, in the event more than one ‘saiz’/’saio’ box pair is included within a single track.

default\_sample\_info\_size is an integer specifying the sample auxiliary information size for the case where all the indicated samples have the same sample auxiliary information size. If the size varies then this field shall be zero.

sample\_count is an integer that gives the number of samples for which a size is defined. For a SampleAuxiliaryInformationSizesBox appearing in the SampleTableBox this shall be the same as, or less than, the sample\_count within the SampleSizeBox or CompactSampleSizeBox. For a SampleAuxiliaryInformationSizesBox appearing in a TrackFragmentBox this shall be the same as, or less than, the sum of the sample\_count entries within the TrackRunBoxes of the track fragment. If this is less than the number of samples, then auxiliary information is supplied for the initial samples, and the remaining samples have no associated auxiliary information.

sample\_info\_size gives the size of the sample auxiliary information in bytes. This may be zero to indicate samples with no associated auxiliary information.

### Modification to ‘saio’ box to include an ID to match with the ‘saiz’ box

*Note: If “paired” ‘saiz’/’saio’ boxes are mandated to follow each other in order, a change to the ‘saio’ box is not necessary. Otherwise, a method is needed to align matching ‘saiz’ and ‘saio’ boxes.*

**Syntax**

aligned(8) class SampleAuxiliaryInformationOffsetsBox extends FullBox('saio', version, flags){   
 if (flags & 1) {   
 unsigned int(32) aux\_info\_type;   
 unsigned int(32) aux\_info\_type\_parameter;   
 }   
 unsigned int(32) entry\_count;   
 if ( version == 0 ) {  
 unsigned int(32) offset[ entry\_count ];   
 }   
 else if (version == 1){ //required to meet current guidance on version signaling  
 unsigned int(64) offset[ entry\_count ];   
 }

else if (version == 2) {  
 unsigned int(32) saio\_ID; //required to match the corresponding saiz\_ID  
 }  
}

**Semantics**

aux\_info\_type and aux\_info\_type\_parameter are defined as in the SampleAuxiliaryInformationSizesBox

entry\_count gives the number of entries in the following table. For a SampleAuxiliaryInformationOffsetsBox appearing in a Sample Table Box this shall be equal to one or to the value of the entry\_count field in the ChunkOffsetBox or ChunkLargeOffsetBox. For a SampleAuxiliaryInformationOffsetsBox appearing in a TrackFragmentBox, this shall be equal to one or to the number of TrackRunBoxes in the TrackFragmentBox.

offset gives the position in the file of the Sample Auxiliary Information for each Chunk or Track Fragment Run. If entry\_count is one, then the Sample Auxiliary Information for all Chunks or Runs is contiguous in the file in chunk or run order. When in the SampleTableBox, the offsets are relative to the same base offset as derived for the respective samples through the data\_reference\_index of the sample entry referenced by the samples. In a TrackFragmentBox, this value is relative to the base offset established by the TrackFragmentHeaderBox in the same track fragment (see 8.8.14).

saio\_ID is an ID used to uniquely match the ‘saio’ box to its corresponding ‘saiz’ box within a track, in the event more than one ‘saiz’/’saio’ box pair is included within a single track.

## Annex B - Proposed Collection Timestamp (‘colt’) box solution

Description of content to be included in the Timestamp box:

### Static information (stored at the front of the box):

* + - * + **Timing System Level**
  + Units: Encoded integer enum
  + Type: UINT8
  + Encoding of value:

|  |  |
| --- | --- |
| 0 | Level 0 – inconsistent time system |
| 1 | Level 1 – proper ordering time system |
| 2 | Level 2 – relative difference time system |
| 3 | Level 3 – absolute time system |
| 4-255 | Reserved |

* The ‘colt’ box shall carry labeling metadata to identify the level of timing capability, as defined here:

**Level 0** – A system labels a collection of samples without sufficient timing information to determine the original order of collection. The timing in this situation is defined to be ***inconsistent***.

[Ex] – Media (imagery, etc.) and associated metadata relating to pointing angles and location are collected without any annotated timing information.

**Level 1** – A collection of samples where the timing information allows for determining the ***proper ordering*** of all samples, across all tracks.

[Ex] – a running counter, driven by an oscillator of unknown frequency, is used to tag each measurement with a counter value for all samples and associated metadata measurements. The tagged measurements are sufficient to determine the order of samples, but not for calculating a difference of time between samples.

**Level 2** – A system labels all samples with timing information in a manner that enables computing the ***relative difference in time***, using SI Seconds, between two timestamps

[Ex] – a running counter, driven by an oscillator of known and stable frequency, is used to tag each measurement with a counter value for all frames and associated metadata measurements.

**Level 3** – A system labels all sample information with ***absolute time***, based on a known universal atomic time standard, with a known statistical measurement uncertainty. This enables temporal correlation of events outside the span of the imaging/media system with sample information produced by the imaging/media system.

[Ex] – camera systems operating in distant locations, tag their imagery and associated metadata with highly accurate GPS timing system derived timing information. This allows for precise 3D measurements of dynamic moving objects within overlapping scenes.

* + - * + **Time Format** (note, choice of epoch is defined by the time format):

Units: Encoded integer enum

Type: UINT8

Encoding of value:

|  |  |
| --- | --- |
| 0 | Unknown Time Format |
| 1 | International Atomic Time (TAI)   * Epoch: 1958-01-01T00:00:00.0Z |
| 2 | Global Positioning System (GPS)   * Epoch: 1980-01-06T00:00:00.0Z * For date calculations:   GPS Time = TAI – 19 seconds   * For seconds calculations:   GPS is equal to TAI with an offset of the GPS epoch seconds-count plus the 19 seconds |
| 3 | MISP Time (U.S. DoD)   * Epoch: 1970-01-01T00:00:00.0Z * For date calculations:   MISP Time = TAI – 8.000082 seconds   * For seconds calculations: * MISP Time is equal to TAI with an offset of the MISP epoch seconds-count plus the 8.000082 seconds |
| 4-255 | Reserved |

* + - * + **Clock Source** (Time Transfer Method):

Units: Encoded integer enum

Type: UINT8

Encoding of value:

|  |  |
| --- | --- |
| 0 | Unknown Time Transfer Method |
| 1 | Global Positioning System (GPS) 1 PPS |
| 2 | Precision Time Protocol (PTP) - Version 1 |
| 3 | Precision Time Protocol (PTP) – Version 2 |
| 4 | Network Time Protocol (NTP) - Version 3 .3 |
| 5 | Network Time Protocol (NTP) – Version 4 |
| 6 | Inter-range Instrumentation Group (IRIG-A) |
| 7 | Inter-range Instrumentation Group (IRIG-B) |
| 8-255 | Reserved |

* + - * + **Synchronization Pulse Frequency**

The Synchronization Pulse Frequency specifies the periodicity of the pulse transmitted by a reference time source. The unit of measure is Pulse per [SI] Second (PPS) or Hertz (Hz). The default value 1.0 PPS.

Units: Hertz (Hz)

Type: FLT32

* + - * + **Leap Seconds Offset** (supports conversion to UTC for display, etc.):

Description: conversion to UTC is a user software implementation requirement. This value is provided to help facilitate conversion. Users are reminded a small probability exists of a leap second change in the middle of a track and may take steps to address this type of rare event.

Units: seconds

Type: INT16

* + - * + **Receptor Clock Measurement Uncertainty**:

Description: the measurement uncertainty of the receptor clock

Units: nanoseconds

Type: UINT64

* + - * + **Drift Rate:**

Description: the maximum drift rate of the source clock when freewheeling

Units: microseconds/second

Type: FLT32

* + - * + **Signal Source Delay:**

Description: The Signal Source Delay is the estimated or measured latency of the time signal from the reference clock to the receptor clock.

Units: nanoseconds

Type: UINT32

* + - * + **Unlock Time:**

Description: Time when synchronization with reference clock was lost. This value can be used when a track is edited and a front-end portion that contains the frame where synchronization was lost is removed. Including this value allows a user to estimate error due to drift over time. A value of zero shall be used when the clock always remains locked.

Units: The units for this measurement are the same as the units of the parent time.

### Dynamic Information, one entry per media sample:

* + - * + **Timestamp:**

64-bit timestamp (counter since epoch of the Time Format)

Units: nanoseconds

Type: UINT64

* + - * + **Time Transfer Parameters:**

Description: Dynamic metadata about the clock source:

Units: Encoded integer

Type: UINT8

|  |  |
| --- | --- |
| Bits 7 (msb) - 4 | Reserved (set to zero) |
| Bits 3 and 2 | **Correction Method** |
| 00 | Unknown |
| 01 | Jam Correction |
| 10 | Slew Correction |
| 11 | Reserved for future use |
| Bits 1 and 0 (lsb) | **Clock Synchronization Status** |
| 00 | Clock source synchronization status is unknown |
| 01 | Clock source is not synchronized to an atomic source |
| 10 | Clock source is synchronized to an atomic source |
| 11 | Reserved for future use |

### (Draft) Syntax:

aligned(8) class CollectionTimestampBox extends FullBox('colt', version, flags){   
 unsigned int (32) entry\_count;   
 unsigned int(8) location;   
 unsigned int(8) size\_of\_timestamp\_block; //set equal to size of timestamp and //clock\_synch\_status  
 unsigned int (8) time\_system\_level;   
 unsigned int (8) time\_format;   
 unsigned int (8) clock\_source;   
 float (32) synchronization\_pulse\_frequency;   
 int (16) leap\_seconds\_offset; //defined as valid for first frame  
 unsigned int (64) clock\_measurement\_uncertainty;   
 float (32) drift\_rate;   
 unsigned int (32) signal\_source\_delay;   
 unsigned int (64) unlock\_time;   
 if (location == 0) {  
 //timestamps are stored in Timestamp box  
 for(i=0; I < entry\_count; i++) {  
 unsigned int (64) collection\_time\_stamp[i];   
 unsigned int (8) time\_transfer\_parameters [i];   
 }  
 }  
 else if(location == 1){   
 //timestamps are contiguous in a “well-ordered”, self-contained array  
 //and stored elsewhere (mdat, etc.)   
 unsigned int (64) offset; //Array size = entry\_count \* size\_of\_timestamp\_block  
 }  
 else if (location ==2){   
 //timestamps are not contiguous and stored elsewhere, in mdat, etc.   
 for(j=0; j<entry\_count; j++) {  
 unsigned int (64) offset[j]; //points to individual timestamp/status blocks  
 }  
 }  
}

# Carriage of video with depth (RGBD)

Reference to discussions:

* <http://mpeg.expert/software/MPEG/Systems/FileFormat/isobmff/-/issues/164>

The original proposal (below) initiated a discussion on whether there are enough tools to support packaging and signalling of video with depth (RGBD). The group decided that indeed there is a need for further study.

## Introduction

With the advent of XR, there has been an increasing need for carriage of depth information alongside the video (i.e. RGB+D). This becomes especially needed now for several applications, for example:

1) Split rendering (used for image corrections)

2) XR teleconferencing (move PoV, background removal etc.)

3) MR placement (model/texture adaptation)

## Existing Tools

Currently ISOBMFF supports accompanying video with depth via auxiliary tracks – with reference\_type as ’auxl’ or ‘vdep’ (or both), as following:





Text

Description automatically generated

The only other provision in ISOBMFF spec is for auxiliary video metadata (item of type ‘auvd’), that references ISO/IEC 23002-3 “Auxiliary Video Data Representation”:

Text

Description automatically generated

We examined how ISO/IEC 23002-3 specifies the usage of related auxiliary video data, and we found out that mostly focuses on parallax video, with little provisions for depth streams (outside the parallax domain). Additionally, this standardised information is designed for injection in the video bitstream and no carriage at the encapsulation level is enabled. Lastly, at the time of writing of this standard, the targeted applications where completely different from the current media ecosystem therefore even the existing few provisions are not useful today.

ISO/IEC 23001-17 “Carriage of Uncompressed Video in ISOBMFF” has provisions for depth carried as a “component”. A component is typically used to carry colour information (of one channel) and can be used for depth. As such, it does not have any depth-specific features, besides the DepthMappingInformationBox which describes how the values are mapped to distance with its usage refering to ISO/IEC 23002-3.

## Current State

What the ecosystem is lacking right now for RGBD transmission is an interoperable, content-centric, but application and codec agnostic way to encapsulate the depth accompanying the video, while at the same time exposing the properties for the video and associated depth stream to distribution mechanisms (i.e. to be streaming-ready).

To enable XR applications, there is depth information that must be available prior to accessing the data. Such information includes (among others):

- Depth range

- Depth projection type (point/plane)

- Depth projection properties

- Depth coding properties (RGB/Grayscale)

That currently is not carried in a standardized way.

The mentioned properties, combined with some inherent FF features (ease of segmentation, tool for random access etc.) can produce an RGBD encapsulation mechanism that would be both useful today and extensible for the future.