Text

Description automatically generatedISO/IEC JTC 1/SC 29/WG 03 N0360

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

**Document type:** Output Document

**Title:** Technologies under Consideration for ISOBMFile Format

**Status:** Approved

**Date of document:** 2021-10-15

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

**No. of pages:** 63 (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 N0360**

**October 2021, Virtual**

|  |  |
| --- | --- |
| **Title** | **Technologies under Consideration for ISOBMFile Format** |
| **Source** | **WG 03, MPEG Systems** |
| **Status** | **Approved** |
| **Serial Number** | **20837** |

**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 2](#_Toc87544151)

[2 Edit lists and movie fragments 5](#_Toc87544152)

[3 Multiplexed timed metadata tracks 8](#_Toc87544154)

[4 Embedded Metadata Signaling 14](#_Toc87544155)

[5 On MIME type parameters 14](#_Toc87544156)

[6 On relation of entity groups and sample groups 15](#_Toc87544157)

[7 Stronger defaulting in Track Runs 16](#_Toc87544158)

[8 Sample reordering in Track Runs 25](#_Toc87544159)

[9 Extending Segment Index 30](#_Toc87544160)

[10 Segment Index and Level Assignment 32](#_Toc87544161)

[11 Haptics support 38](#_Toc87544162)

[12 Generic sub-picture track grouping extensions 41](#_Toc87544163)

[13 Signaling of Multi-Layer Picture Compositing Information in the VVC File Format 45](#_Toc87544164)

[14 Picture-in-Picture (from m57985) 51](#_Toc87544165)

# Handling lost or corrupted samples

## 'loss', 'null' and 'crpt' sample entries

[Ed. (MH): This section results from the following decisions in MPEG#116 and MPEG#117: In the MPEG#117 file format minutes regarding M39936: "The ‘null’ and ‘crpt’ sample entries should be in the TuC for part 12." In the MPEG#116 file format minutes regarding M39333: "We add a section for now to the Partial file WD, but that would go into Part12, to document 'null' and 'crpt' samples (lost being a special case)."]

[Ed. (FM): In the MPEG#118 file format minutes regarding m40528: “In this section, we document the three sample entries above; but the receiver might like to know the original format for material that is lost or corrupted. Should these be seen as transformations of the sample entry (like encryption, or restricted) rather than independent? They then would vary an original format box etc.

This, of course, raises the question of whether you always know what the type of the lost samples was.”]

The sample entry type 'loss' indicates that the file writer does not know what media should be here, if anything. The sample entry type 'null' indicates the same as an empty edit. When the sample entry type is 'loss' or 'null', the sample size shall be equal to 0. [Ed. (MH): From M38920 with editorial changes.]

A corrupted media frame is media frame that cannot be correctly decoded, due to being partially received or due to missing data in the media frames that are in the inter frame prediction chain. If a decoder tries to decode a corrupted media frame, an unexpected behavior, which can be as bad as decoder crash, may occur.

For indication of corrupted media files, the sample entry type 'crpt' is defined. This way, samples that contain media frames that are known to be corrupted can be associated with a sample entry of type 'crpt'. [Ed. (MH): 'crpt' originates from M39333 with editorial changes.]

'loss' sample entries should carry an OriginalFormatBox, documenting the format of the lost samples. 'crpt' sample entries must carry both an original format box, and also any sample entry boxes required by the original format.

## Using sample groups (from [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))

### Discussion

As a reminder, the use case for these sample entries is recording of an elementary stream, not recording of an ISOBMF file transmitted over a lossy network. In the later case, the PartialFileFormat provide the tools to document lost parts of the source file.

As stated in m55062, there is a problem using sample entries to signal loss or corrupted samples:

* in regular files, this will increase the size of the stsc and stco as a new chunk will be needed for each set of contiguous loss/null/crpt
* in fragmented files, this will increase the size of moof by introducing new truns for each set of contiguous loss/null/crpt

Additionally, in NALU based formats or other derived specifications using subsamples, the granularity of the loss/corrupted info is likely at the NAL/subsample level, rather than at the sample. Using a sample entry will mark the entire sample as corrupted, but will not describe which part of the sample is corrupted or clean.

We are not sure if ‘null’ is needed here, absence of samples could be done using edits or empty track fragments.

It is also very hard to understand how a file writer would take the decision to flag as ‘null’, it will likely flag as loss or corrupted

If this TuC proposal is to move forward, we would suggest usage of sample groups, and potentially usage of NALU map entry for NAL-based videos.

### Proposal 1: a sample group for corrupted samples

*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:

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:

* ParameterSetCorruptedFlag (value 0x00000001): indicates that one or more parameter sets (DCI, VPS, SPS, PPS, APS, OPI?) in the associated data is corrupted
* SEICorruptedFlag (value 0x00000002): indicates that one or more SEI messages in the associated data is corrupted
* SliceHeaderCorruptedFlag (value 0x00000004): indicates that one or more slice headers or picture headers in the associated data is corrupted
* VCLCorruptedFlag (value 0x00000008): indicates that VCL data of one or more slices in the associated data is corrupted
* OtherNonVCLNALCorruptedFlag (value 0x00000010): indicates that one or more NAL units in the associated data with types different from the above types is 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).

[[Ed Note on ParameterSetCorruptedFlag : do we want to split per NAL unit type ?]]

### Proposal 2: Essential sample group descriptions

We notice that restricted/transformation sample entries are mostly a way to inform the file reader of properties that must be understood to process a sample entry, but because there is no proper tool to signal this in the file format, we “hide” the original sample entry format so that file readers to not process what they do not recognize. This has several drawbacks:

* many, potentially nested transformations.
* sample granularity requires multiple sample entries (with impact on file size and fragmentation explained above)
* no forward-compatible mechanism, resulting in new transformations whenever a new property is defined
* sample entry type transformation is very fragile, as it depends on the base class being transformed (video sample entry, audio sample entry vX, text sample entry etc …)

We believe that ISOBMFF could use a sample grouping mechanism indicating that the sample group shall be supported (parsed and understood) to process a sample, much like what is defined in item property association.

Once defined, this mechanism will allow definition of new essential properties to process a sample without having to define new brands or transformation of sample entry types.

We therefore propose to define a version 3 for SampleGroupDescriptionBox

In 8.9.3.1 add

“If version equals 3, the sample group description describes essential information for the associated samples, and file processors shall not attempt to decode any track for which unrecognized sample group descriptions marked as essential are present.”

We should then enforce support for version 3 of SampleGroupDescriptionBox through a new brand.

Note on the name:

The version 3 of a sample group description makes a sample property essential rather than descriptive. We originally thought about changing the name to “SampleGroupEssentialPropertyBox”, however introducing a new box results in major edits in the specification, definition of a “SampleToPropertyBox”, rules for sample properties in movie fragments, etc… We therefore believe using a version 3 on the existing mechanism is much simpler to handle.

Note on mime:

The essential properties defined in a file should be advertised through a MIME sub-parameter, for example:

codecs=hvc1.1.6.L186.80, essential=4CC1[,4CC2]

# Edit lists and movie fragments

## 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.]

## Original Duration

### Introduction

Considering the discussion in [m58087](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/77) and its issue, this proposes a way to signal:

* The original intended duration of a sample for player to know where to stop on the sample
* the "past" or elapsed duration of a cloned sample for player to know where to start rendering the content of the cloned sample
* without impacting tfdt or sample duration

[[ed Note: we propose here a new box for the sake of backward compatibility, but this feature could be directly added to ‘tfdt’ for compacity reasons, saving one full box header, as outlined in the contribution.]]  
[[ed: VTT has a more general technique that allows arbitrary sample splitting, but that is specific to VTT. We could maybe make a general technique here for splitting samples other than only at Fragment boundaries?]]

### Proposal

X Fragmented Sample Time Adjustment

X.1 Definition

Box Type: 'fsta'  
Container: TrackFragmentBox   
Mandatory: No  
Quantity: Zero or one

The FragmentedSampleTimeAdjustmentBox provides:

* the elapsed time, measured in media timescale, of the first sample in decoding order in the track fragment.
* the original duration, measured in media timescale, of the last sample in decoding order in the track fragment.

This can be useful to document that the first sample of the track fragment is a copy of the previous sample, if any, and that this sample original start time was intended to be before its actual sample decode time. This allows rewinding the sample playback time at tune-in but ignoring it in regular playback mode.

This can also be useful to document that the last sample duration was truncated to respect fragmentation constraints, and that the intended duration of the sample is longer than its actual duration in the fragment; this allows exact processing of samples with internal timing logic dependent on the sample duration (such as text animations).

The following flags are defined:

* FSTA\_ORIGINAL\_DURATION: flag value is 0x000001. If set, the box describes the original duration of the last sample in the track fragment
* FSTA\_ELAPSED\_DURATION: flag value is 0x000002. If set, the box describes the elapsed duration of the first sample in the track fragment

When this box is present and flag FSTA\_ORIGINAL\_DURATION is set, it indicates that the last sample of the track fragment has a shorter duration than originally authored, and the original duration is signaled. The originalDuration shall be equal to or greater than the duration of the last sample in this track fragment.

When this box is present and flag FSTA\_ELAPSED\_DURATION is set, the first sample of the track fragment is treated as if its associated sample\_flags value has sample\_depends\_on=2 and sample\_has\_redundancy=1.

If a previous sample was already received for this track, that previous sample duration is extended by the duration of this first sample and the elapsedDuration is ignored. If the previous sample had an originalDuration signaled, the extended duration shall be:

* less than originalDuration if this is the only sample of the track fragment and it has an originalDuration associated
* equal to originalDuration otherwise

If no previous sample were received for this track (tune in), the first sample is processed at its sample decode time as if it was being presented for the indicated elapsedDuration.

Note: FSTA\_ORIGINAL\_DURATION and FSTA\_ELAPSED\_DURATION may be set together in a track fragment with multiple samples (describing that first sample is a continuation and last sample is truncated) or with a single sample (describing the only sample is both a continuation and truncated, for example when splitting a sample of 20 second long into three or more fragments).

X.2 Syntax

aligned(8) class FragmentedSampleTimeAdjustmentBox extends FullBox('fsta', 0, flags)  
{  
 if (flags & FSTA\_ORIGINAL\_DURATION) unsigned int(32) originalDuration;  
 if (flags & FSTA\_ELAPSED\_DURATION) unsigned int(32) elapsedDuration;  
}

X.3 Semantics

originalDuration gives the original duration of the last sample of the track fragment, in media timescale of this sample

elapsedDuration gives the elapsed duration of the first sample of the track fragment, in media timescale of this sample

# 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 MetedataKeyBox (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

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 characteric 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.

# 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 1: 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):

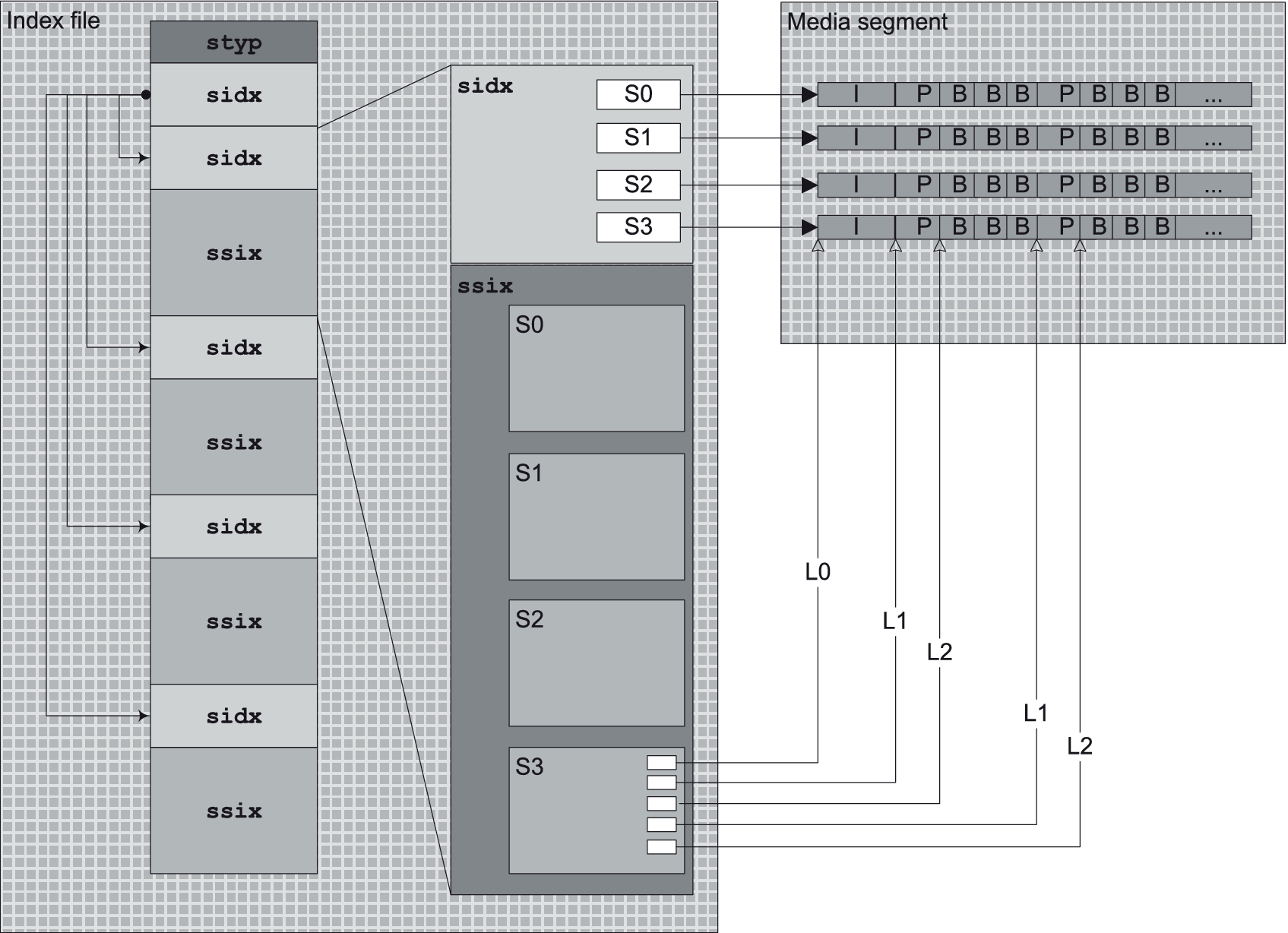


Figure 3: Mapping temporal sub-layers to Sub-Segment Index Box for trick mode in DASH

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;

# Extending Segment Index

## Introduction

Following discussions on 'ssix', this proposal suggests extending the 'sidx' to carry additional information. It basically proposes to add the sap size to the sidx, to enable efficient download of SAP when using MPEG DASH.

## Motivation

There are various ways to provide fast forward and fast rewind (a.k.a. trick modes, video scrubbing…) using http streaming. Some approaches may use a dedicated I-frame/low frame rate stream, while some other may use jumps between keyframes in a regular stream.

Using a dedicated stream increases the complexity of content generation and delivery, and it may be beneficial to provide trick modes based on „regular“ streams.

Assuming each segment begins by an I-frame, some implementations use „download and cancel“ approach: The download of the segment is started, then cancelled once the I-frame has been successfully downloaded and displayed. In this case, the ‚sidx‘ may be sufficient. However, using http 1.1, it requires closing the http session after each I-frame, which is not efficient.

‚ssix‘ may provide more precise indexing information (byte-ranges to portions of sub-segments) so that only the required amount of data is downloaded. It can be done by describing stream access points (sap) or temporal levels associated to stream samples, using sample groups. This avoids closing the http session, as precise byte-ranges are provided.

However, it comes with a significant increase in complexity, regarding content generation and player processing, because ‚ssix‘ requires many additional boxes to be present:

* ‚ssix‘, providing a mapping between levels and byte-ranges for a list of sub-segments
* ‚leva‘, providing a mapping between levels and the associated meaning. In our example, this meaning would typically be provided by a sample group, providing the sap type or the temporal level for each sample
* sample group description ‚sgpd‘ and sample to group box ‚sbgp‘ describing the criterion used for sample grouping and the group that each sample belongs to.

Further, when using MPEG DASH, the sample group information may be missing once the segment index has been downloaded, as it is partially or totally contained in the movie fragment header. This could thus require an additional download per subsegment to fill the missing information.

## Proposal

We would like a simple way to get the size of the SAP starting each subsegment.

We propose to extend the segment index information, to provide a range of bytes sufficient to download the SAP. This would allow downloading just the right amount of bytes to get the entire SAP, while keeping the http session open, and would enable simple seek/trick modes support.

aligned(8) class SegmentIndexBox extends FullBox('sidx', version, 0) {  
 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 (version>=2) {   
 if (starts\_with\_SAP) {  
 unsigned int(32) SAP\_range;  
 }  
 }  
 }  
}

SAP\_range provides a range of bytes, starting from the beginning of the subsegment sufficient to download the SAP (if any) associated to this subsegment.

*Note: could be made more constrained/explicit using SAP\_size instead of SAP\_range*

## Conclusion

We propose to consider the modification of the ‚sidx‘ syntax, to carry additional useful information, which may go beyond the modification proposed here.

While being flexible, the ‚ssix‘ design comes with a significant added complexty and is quite confusing. We believe the ‚sidx‘ could and should be sufficient to accomodate simple use cases.

## Notes from the minutes (Brussels)

Putting this inside the loop breaks backward compatibility; perhaps it would be safer in a separate loop? Would a flag be better than a version (unclear)? And should it be conditional on the reference\_type as well as starts\_with\_SAP and the feature being present? Do we need specific provision for "I don't know" (0?)?

Putting it all together, something like this

if (flags & mask) {  
 for(i=1; i <= reference\_count; i++)  
 {  
 if ((reference\_type[i] == 0) && (starts\_with\_SAP[i] == 1))   
 {   
 unsigned int(32) SAP\_end\_offset;  
 }  
 }  
 }

But can we achieve a similar effect by somehow telling the reader that level 0 in the ssix is, in fact, the SAP? Then no reading of the level assignment box, or sample groups, is needed. Also one could include two byte ranges for level 0 (the initial SAP, and enough bytes to include the second SAP).

We could do this with a version change to the ssix ("for version 1, the level assignment box is optional, and whether present or not, level 0 means the SAPs").

# Segment Index and Level Assignment

See also *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.

cyan text is additional in the definition.

yellow text and ~~red strikethrough~~ text are moved from previous definition and semantics.

**Definition**

Box Type: 'ssix'   
Container: File   
Mandatory: No  
Quantity: Zero or more

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.

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.

**Syntax**

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;   
 }  
 }  
}  
}

**Semantics**

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 subsegment, in which case assignment of remaining bytes to level is unknown but the remaining bytes shall 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. movie fragment). Media data boxes shall not be included in level 0 ranges.
  + Level 1 means the data is independently decodable (SAP 1, 2 or 3). A byte range assigned to level 1 may contain the initial part of the subsegment (e.g. movie fragment box). The start of a byte range assigned to level 1 shall coincide with the start of a top-level box in the subsegment.
  + Level 2 means the data is independently decodable (SAP 1, 2 or 3). The start of a byte range assigned to level 2 shall not coincide with the start of a top-level box in the subsegment.
  + Level N requires data from the preceding byte ranges with lower levels (level N-1 and below) to be processed, stopping at the last specified level 0 if specified, otherwise at the last specified level 1 or 2 if specified, otherwise at the first byte range . Ranges assigned to levels other than 2 may contain movie fragment box.
* 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 2. The remaining less significant bits of the level field give a trackID, which shall identify a track of the movie present in the indexed subsegment for level values other than 0, and shall be 0 if level value is 0. In this mode, each range consists only of data from the identified track, possibly with some meta-data boxes (movie fragments, etc...). The level value only gives dependency information within the track. This allows cross-track indexation within a same level.
* 3, 4, 5, 6, 7: 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. 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.

*Editor's notes :*

* the range\_count could be restricted to 1 or more.
* The incomplete flag is more an indication that size must be cross-checked with size indicated in sidx, we could remove the flag.
* the design here explicitly lists the flags required for parsing and the level assignment. These flags (2 bytes) could be removed from the box, using the full\_box flags instead (3 bytes).

## Modifications to leva

This relates to ISOBMFF TuC, section 16.2.2

*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."

Editors Note: there is a pending Defect Report (2.5) on the use of padding flag to be inserted here (See *https://github.com/MPEGGroup/FileFormat/issues/11)*

*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)

## Examples of use

### Single track indexing

A picture containing shape

Description automatically generated

Figure 4 - level\_assignment\_type=1 with 2 IDRs and no explicit range for moof

A picture containing shape

Description automatically generated

Figure 5 - level\_assignment\_type=1 with 2 IDRs and explicit range for moof

A picture containing text

Description automatically generated

Figure 6 - 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 7 - 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 8 - 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 9 - level\_assignment\_type=2, 1 moof with 2 trafs, 1 IDR/traf and explicit explicit range for moof

# Haptics support

## Context

This section proposes adding some common boxes that could be used by Haptics media, in the new subsection of Chapter 12.

### Sample Entry

#### Definition

The arrays of VibeGeneralConfigBox and KinestheticGeneralConfigBox boxes within the sample entry indicate whether the content in this track requires vibrator actuator(s), kinesthetic actuator(s), or both.

This provides information in the sample entry on the types of actuators the content in this track is designed to be played on. For example, if the content is intended for single vibrator actuator devices (e.g., basic phones, laptops, tablets), then the sample entry will have a single VibeGeneralConfigBox (array of one). If the track is intended for multiple vibrator actuators, the sample entry will have an array of VibeGeneralConfigBoxes, each of different frequency and motor type. A similar structure can be used for kinesthetic actuators.

NOTE it is possible to create a haptic track that can be played on both vibrator and kinesthetic actuators. The sample entry for such a track will have non-zero arrays of both VibeGeneralConfigBox and KinestheticGeneralConfigBox boxes.

#### Syntax

aligned(8) class HapticSampleEntry(codingname)extends SampleEntry(codingname) {  
 unsigned int(8) vibe\_entry\_count; // count of vibrator actuators   
 unsigned int(8) kinh\_entry\_count; // count of kinesthetic actuators   
 unsigned int(8) other\_entry\_count; // count of otherboxes  
 unsigned int(8) reserved = 0; // to pad to 8 bits  
  
 VibeGeneralConfigBox[vibe\_entry\_count] vibes;   
 // array of vibrator actuator configurations  
 KinestheticGeneralConfigBox[kinh\_entry\_count] kines;   
 // array of kinesthetic actuator configurations  
 Box()[other\_entry\_count] otherboxes;  
}

### Vibrator General Configuration

#### Definition

VibeGeneralConfigBox is a general-purpose configuration box in HapticSampleEntry for specific vibrator-based haptics systems. If the content is designed to be (ideally) played on a device with one or more vibrator actuators, there would be an array of one or more VibeGeneralConfigBoxes in the HapticSampleEntry.

#### Syntax

aligned(8) class VibeGeneralConfigBox extends FullBox('vibh') {   
 template unsigned int actuatorID(4);   
 // arbitrary value uniquely identifying this actuator configuration  
 template unsigned int actuatorFrequency(16);   
 template unsigned int actuatorConfiguration(4);  
 template unsigned int endpointConfiguration(4);  
 unsigned int(4) reserved = 0; // to pad to 8 bits  
}

#### Semantics

actuatorID is an arbitrary value that identifies this actuator and must be unique among the VibeGeneralConfigBoxes in a given sample entry; there are two reserved values, 0 and 0xF, which shall not be used.

actuatorFrequency is a 16-bit value representing the actual frequency of the actuator used. Currently known actuators typically use 200 Hz, 1000 Hz, or 8000 Hz. Newer actuators could use other frequencies.

endpointConfiguration is a 4-bit value representing the waveform data format to be used by the codec to interpret the haptic media.

The actuatorConfiguration and endpointConfiguration are described below:

Table 2: Actuator Configuration

|  |  |
| --- | --- |
| **Actuator Configuration Index** | **Description** |
| 0x0 | Eccentric Rotating Mass (ERM) |
| 0x1 | Narrowband Linear Resonant Actuator (LRA) |
| 0x2 | High-definition (HD) Motor |
| 0x3 – 0xF | Reserved |

Table 3: Endpoint Configuration

|  |  |
| --- | --- |
| **Endpoint Configuration** | **Description** |
| 0x0 | Amplitude Modulation |
| 0x1 | Unipolar Waveform |
| 0x2 | Bipolar Waveform |
| 0x3-0xF | Reserved |

### Kinesthetic General Configuration

#### Definition

KinestheticGeneralConfigBox is a general-purpose configuration box in HapticSampleEntry for specific kinesthetic haptics systems. If the content is designed to be (ideally) played on a device with one or more kinesthetic actuators, there would be an array of one or more KinestheticGeneralConfigBoxes in the HapticSampleEntry.

#### Syntax

aligned(8) class KinestheticGeneralConfigBox extends FullBox('kinh') {  
 template unsigned int actuatorID(4);   
 // arbitrary value uniquely identifying this actuator configuration  
 template unsigned int controlFrequency(4);  
 template unsigned int endpointConfiguration(4);  
 unsigned int(4) reserved = 0; // to pad to 8 bits  
}

#### Semantics

actuatorID is an arbitrary value that identifies this actuator and must be unique among the KinestheticGeneralConfigBoxes in a given sample entry; there are two reserved values, 0 and 0xF, which shall not be used.

endpointConfiguration is a 4-bit value representing the waveform data format to be used by the codec to interpret the haptic media, as described in Table 3 above.

controlFrequency is a 4-bit field specifying the control frequency of the kinesthetic feedback.

### Codec-specific Sample Entry Extensions

Each haptic codec may extend HapticSampleEntry and contain codec-specific information.

Each codec-specific sample entry extension may also specify the maximum number of actuators it needs for rendering the haptics track, using VibeGeneralConfigBoxes or KinestheticGeneralConfigBoxes.

# Generic sub-picture track grouping extensions

## 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 7.1.11.2 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.

## Two dimensional spatial relationship

### 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.

The SubPictureRegionBox is optional and either:

1. is present in the SpatialRelationship2DDescriptionBox and there shall be no associated SpatialRelationship2DGroupEntry in the associated track (this track has a constant, static, size and position); or
2. is not present in the SpatialRelationship2DDescriptionBox and there shall be one or more associated SpatialRelationship2DGroupEntry(s) in the associated track (this track possibly has a dynamic size and/or position).

When constructing the composition picture as specified below, gaps between sub-pictures carried in different sub-picture tracks are allowed, and overlaps between sub-pictures carried in different sub-picture tracks with different values of the layer field in TrackHeaderBox are allowed. However, overlaps between sub-pictures carried in different sub-picture tracks with the same value of the layer field in TrackHeaderBox are not allowed.

The spatial relationship is restricted according to the chroma sub-sampling format of the associated track; total\_width and total\_height, and object\_x, object\_y, object\_width and object\_height, shall all select an integer number of samples for all planes. In effect this means that:

* when the format is 4:4:4, there is no restriction;
* when the format is 4:2:2 the total\_width, object\_x and object\_width shall be even numbers;
* when the format is 4:2:0 all of these fields shall be even numbers.

The composition picture is reconstructed as follows, with values of object\_x, object\_y, object\_width, and object\_height obtained from SubPictureRegionBox if present or otherwise from the SpatialRelationship2DGroupEntry applying to the sample:

1. Out of all tracks belonging to the same '2dsr' track group, form them into subgroups such that each subgroup contains tracks in the same alternate group; then select exactly one track from each of those subgroups.
2. For each composition-time aligned sample of each of the selected tracks, the following applies, in the front-to-back ordering (layer) indicated in the TrackHeaderBox of the picked tracks:

For each value of i in the range of 0 to object\_width − 1, inclusive, and for each value of j in the range of 0 to object\_height − 1, inclusive, the pixel value of the composition picture at pixel position ( ( i + object\_x ) % total\_width, ( j + object\_y ) % total\_height ) is set equal to the pixel value of the sample of this track at pixel position (i, j).

### 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 SubPictureRegionBox extends FullBox('sprg',0,0) {  
 unsigned int(16) object\_x;  
 unsigned int(16) object\_y;  
 unsigned int(16) object\_width;  
 unsigned int(16) object\_height;  
 bit(14) reserved = 0;  
 unsigned int(1) track\_not\_alone\_flag;  
 unsigned int(1) track\_not\_mergeable\_flag;  
}

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

### Semantics

total\_width specifies, in pixel units, 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, in pixel units, 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.

object\_x specifies the horizontal position of the top-left corner of the samples in this track within the coordinate system specified by this 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 this SpatialRelationship2DDescriptionBox.

object\_y specifies the vertical position of the top-left corner of the samples in this track within the coordinate system specified by this 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 this SpatialRelationship2DDescriptionBox.

object\_width specifies the width of the samples in this track within the coordinate system specified by this 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 track within the coordinate system specified by this 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.

track\_not\_alone\_flag equal to 1 indicates that the current sub-picture track is not intended to be presented alone without at least one other sub-picture track belonging to the same track group of grouping type '2dsr'. The value 0 indicates that the current sub-picture track may or may not be intended to be presented alone without at least one other sub-picture track belonging to the same track group of grouping type '2dsr'.

track\_not\_mergeable\_flag equal to 1 indicates that the video bitstream carried in the current sub-picture track cannot be merged with the video bitstream carried in any other sub-picture tracks belonging to the same track group of grouping type '2dsr', to generate a single video bitstream without decoding mismatch by rewriting only header data of the bitstreams, where a decoding mismatch refers to the value of any pixel when decoding the video bitstream in the current track is not identical to the value of the same pixel when decoding the merged single video bitstream. An example of such bitstream merging is the reconstruction of an HEVC bitstream as specified in clause 10.1.3.4 when the untransformed sample entry type of the track with the given track\_ID is equal to 'hvc2'. track\_not\_mergeable\_flag equal to 0 indicates that the video bitstream carried in the current sub-picture track can be merged with the video bitstream carried in at least one other sub-picture track belonging to the same track group of grouping type '2dsr' to generate such a single video bitstream in such a manner as described above.

NOTE: When HEVC (i.e., Rec. ITU-T H.265 | ISO/IEC 23008-2) is the video codec used for encoding of the bitstreams carried in the sub-picture tracks, track\_not\_mergeable\_flag equal to 0 means that the HEVC bitstream carried in the current sub-picture track contains and only contains one or more MCTSs that can be indicated by a temporal MCTSs SEI message as specified in HEVC version 5 published by the ITU-T in Feburary 2018, or a later version of HEVC.

## 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.

There are restrictions both on the presence of this sample grouping, and on the values of the fields; see clause 7.1.11.2.1.

### 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;  
}

### 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.

# 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

# Picture-in-Picture (from [m57985](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/70))

## Summary of the proposal

The proposal is summarized as follows:

1. A new type of entity grouping, named picture-in-picture entity grouping, with the grouping\_type equal to 'pinp' is proposed.
2. Each entity in the entity group must be a video track.
3. A PicInPicEntityGroupBox is defined, by extending the EntityToGroupBox, to carry the following pieces of information:
   1. The number of main bitstream tracks N. The entities (i.e., tracks in this context) identified by the first N entity\_id values in the EntityToGroupBox are main bitstream tracks, while the entities identified by other entity\_id values in the EntityToGroupBox are supplementary bitstream tracks. For playing back of the picture-in-picture experience, one of the main bitstream tracks is chosen, and one of the supplementary bitstream tracks is chosen.
   2. An indication that indicates whether it is enabled to replace the coded video data units representing the target picture-in-picture region in the main video with the corresponding video data units of the supplementary video.
   3. A list of region IDs, for indicating which coded video data units in each picture of the main video represent the target picture-in-picture region.
   4. The position and size in the main video for embedding/overlaying the supplementary video, which is smaller in size than the main video, is indicated using the matrix of the overlay track, and the layer header field to layer it in front of the main video.

## Picture-in-picture entity grouping

### Definition

Picture-in-picture services offer the ability to include a video with a smaller spatial resolution within a video with a bigger spatial resolution, referred to as the supplementary video and the main video, respectively. Tracks in the same entity group with grouping\_type equal to 'pinp' can be used for support of picture-in-picture services, by choosing one of the tracks that are indicated to contain the main video and one of the other tracks (which contain the supplementary video).

All entities in a picture-in-picture entity group shall be video tracks.

### Syntax

aligned(8) class PicInPicEntityGroupBox extends EntityToGroupBox('pinp',0,0) {  
 unsigned int(8) num\_main\_video\_tracks;  
 unsigned int(1) data\_units\_replacable;  
 bit (7) reserved = 0;  
 if(data\_units\_replacable) {  
 unsigned int(8) num\_region\_ids;  
 for(i=0; i<num\_region\_ids; i++)  
 unsigned int(16) region\_id[i];  
 }  
}

### Semantics

num\_main\_video\_tracks specifies the number of tracks in this entity group that carry the picture-in-picture main video.

data\_units\_replacable indicates whether the coded video data units representing the target picture-in-picture region in the main video can be replaced by the corresponding video data units of the supplementary video. The value 1 indicates that such video data units replacement is enabled, and the value 0 indicates that such video data units replacement is not enabled.

When data\_units\_replacable is equal to 1, the player may choose to replace the coded video data units representing the target picture-in-picture region in the main video with the corresponding coded video data units of the supplementary video before sending to the video decoder for decoding. In this case, for a particular picture in the main video, the corresponding video data units of the supplementary video are all the coded video data units in the decoding-time-synchronized sample in the supplemental video track. In the case of VVC, when the client chooses to replace the coded video data units (which are VCL NAL units) representing the target picture-in-picture region in the main video with the corresponding VCL NAL units of the supplementary video before sending to the video decoder, for each subpicture ID, the VCL NAL units in the main video are replaced with the corresponding VCL NAL units having that subpicture ID in the supplementary video, without changing the order of the corresponding VCL NAL units.

num\_region\_ids specifies the number of the following region\_id[i] fields.

region\_id[i] specifies the i-th ID for the coded video data units representing the target picture-in-picture region.

The concrete semantics of the region IDs need to be explicitly specified for specific video codecs. In the case of VVC, the region IDs are subpicture IDs, and coded video data units are VCL NAL units. The VCL NAL units representing the target picture-in-picture region in the main video are those having these subpicture IDs, which are the same as the subpicture IDs in the corresponding VCL NAL units of the supplementary video.