INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2016/M38642
May 2016, Geneva (CH)

Source **Canon Research Centre France, Telecom ParisTech**
Status **For consideration at the 115th MPEG meeting**
Title **Storage of Tiled HEVC Video**
Author Franck Denoual, Frédéric Mazé, Jean Le Feuvre, Cyril Concolato

1 Introduction

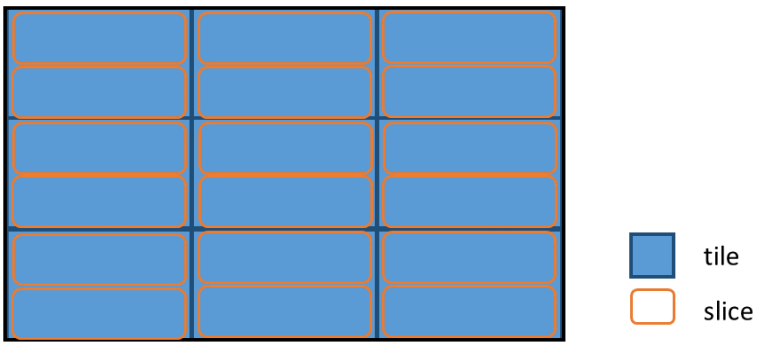
Following discussions and recent proposals on tile descriptors and tile tracks, this contribution explains how to use the current description tools related to HEVC tiles, as defined in MPEG-4 Part-15 (DIS w15640 and draft FDIS w15928) section 10.
In particular, this contribution provides a list of use cases that can be supported by these tools without any modifications recently discussed or suggested, for example in m38225. Each example section contains a discussion and proposes recommendations for the tile-related tools.

2 Discussion on tile access granularity

From ISO/IEC 23008-2 specification, relationships between tiles and slices can be summarized as follows (section 6.3.1):
A tile always contains an integer number of coding tree units, and may consist of coding tree units contained in more than one slice. Similarly, a slice may consist of coding tree units contained in more than one tile.

2.1 Tile as one or more slices

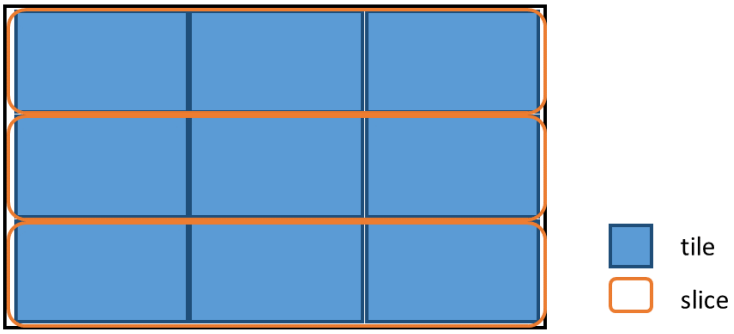
To have access to tile data in an ISOBMFF track, we have to distinguish whether the track samples contain one tile or multiple tiles.



When the track samples contain one tile, sample-based description is sufficient as shown in section 4.1. Sample grouping mechanism from Part-12 allows that.
When the track sample contains multiple tiles, then a mapping of the NAL units to the tiles is required as shown in section 3.1. The `NALUMapEntry` from Part-15 allows that. In case the access is limited to the region formed by this set of tiles sample grouping is sufficient.

2.2 Slice as one or more tiles

In this case, NAL unit granularity is not enough (since one NAL unit corresponds to one slice hence multiple tiles) and then description at sub-sample level is required. This requirement appeared during the votes on w15640 with proposed changes to include sub-sample level mapping in trif or in `NALUMapEntry`. We discuss how to provide sub-sample access granularity without modification of any of the existing tools in section 5.



2.3 Summary on the different levels of description for tile access

The Table below summarizes the appropriate tools to retrieve a tile depending the tile to track mapping, the tile to slice and slice to tile mapping. In red, appear the non-supported cases by the current definition of NALUMapEntry.

Tile per track	Tiles per slice	Trif sample group	Tsif sample group	NALUMapEntry sample group	Sub-Sample
1	< 1*	Recommended	Not Needed if independent tiles Recommended otherwise	Not needed	Not needed
	1				
	> 1**				
> 1	< 1*	Used for tile description		Recommended	
	1				
	> 1***			Not supported	Recommended

* one tile contains more than one slice

** one slice contains more than one tile, however the slice is "split" through extractors to have only one tile per track. This is a corner case where the track type would be hev2/hvc2 yet containing only one tile from the complete bitstream.

*** one slice contains more than one tile

Table 1: Appropriate tools for tile access for various tile-slice and tile-tracks modes

3 On the use of NALU map entry

The `NALUMapEntry` structure, as the name indicates, allows mapping NAL units to a sample group description. `NALUMapEntry` is defined as a specific kind of `VisualSampleGroupEntry`. As such, it is used in a `SampleGroupDescriptionBox`.

Note: As **currently specified (w15928), the NALUMapEntry does not offer mapping at a granularity finer than the NAL Unit.**
How to enable a finer granularity in the mapping is discussed in section 5.

3.1 Example of use for a tiled single-layer video

In this example, we assume one single layer video track embedding two tiles as illustrated on Figure 1. Each tile contains 2 VCL NAL units.

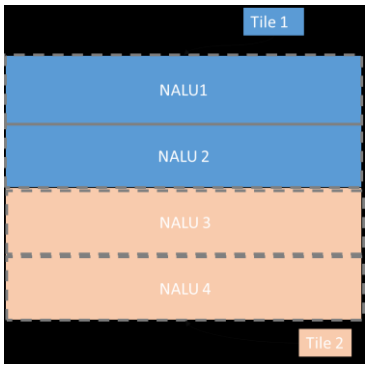


Figure 1: Example of tiled video

The NAL unit mapping to the tiles is described in a visual sample group description entry of type '`trif`' as shown on Figure 2. Since we're working at the NAL unit level, we use the `groupID` of '`trif`' to reference the different tiles.

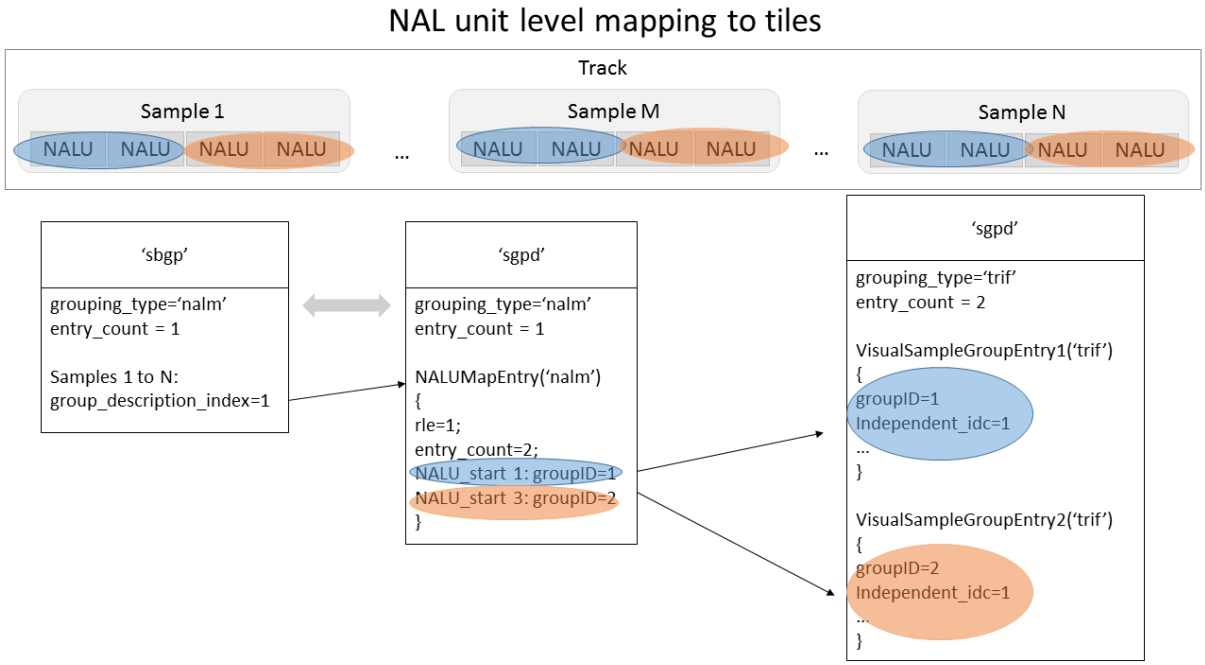


Figure 2: Example of tiled video encapsulation in a single track

The same description could be used for:

- a 2x2 tiled video where one tile exactly matches one NAL unit: the '`sgpd`' box in the middle would then have an entry_count = 4 and each NAL unit would be mapped to one '`trif`' in the '`sgpd`' table on the right (with an entry_count set to 4 and four `VisualSampleGroupEntry`).

Discussion:

There should be a note stating that when nalm is used, nal units of the sample are not directly associated to trif/tsif sample descriptions; consequently, there could be no sample to group of type trif or tsif in these case (although one could always map a sample to a bigger tile, for example the full picture).

3.2 Example of use for a tiled multi-layer video

3.2.1 Tile to full layer dependency, discussing independent_idc

The configuration for the video samples is given on Figure 3 below: the base layer has no tiles while the spatial enhancement layer contains two tiles, each tile containing two slices/NAL units. Each tile depends upon the full base layer, but each tile is independent in the enhancement layer. Two tracks are created, each to encapsulate one layer. Only the NAL units in the enhancement track require a mapping to tiles. For the base layer, since there are no tiles, there is no need for `NALUMapEntry`. The only requirement for the base layer is to define a sample group containing all the samples of the base layer track and associate them to a description box of type '`trif`' so that the whole layer is described as one tile region (`groupID`=1, `full_picture`=1 in '`trif`') that can be referenced from the enhancement layer.

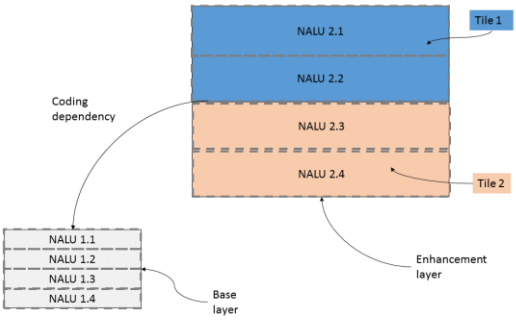
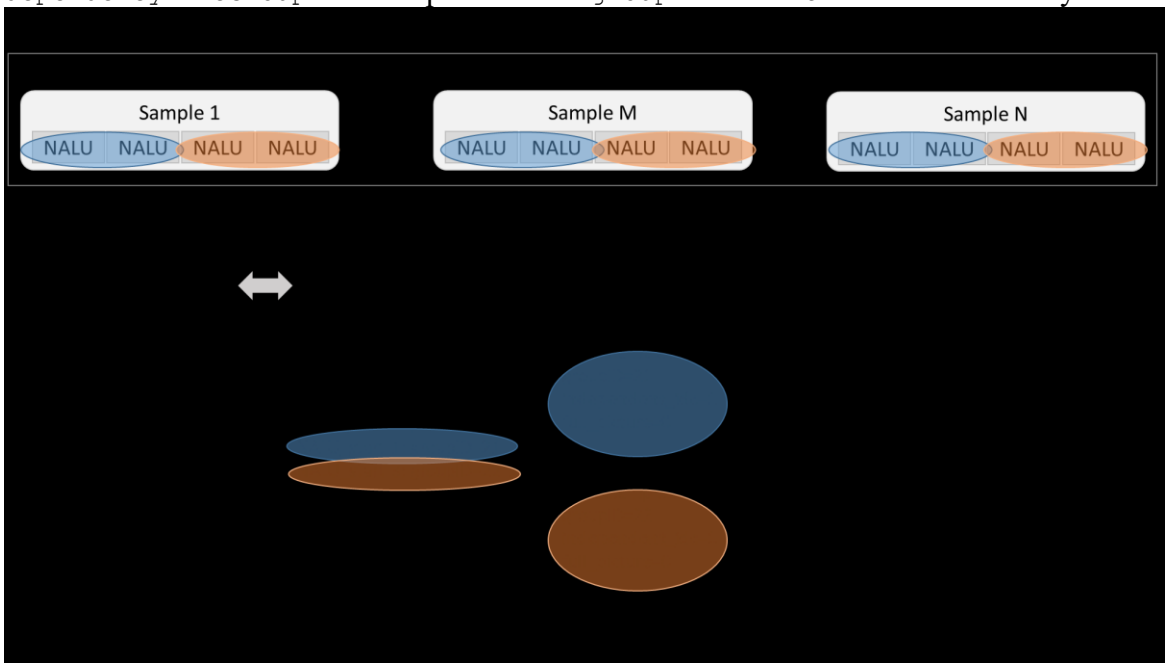


Figure 3: Example of multi-layer video with tiles

The Figure 4 illustrates the encapsulation of the enhancement layer track, using the `NALUMapEntry` and the '`tsif`' descriptor as in w15928. The value 1 for `dependencyTileGroupID` corresponds to the `groupID` of the '`trif`' in the base layer track.



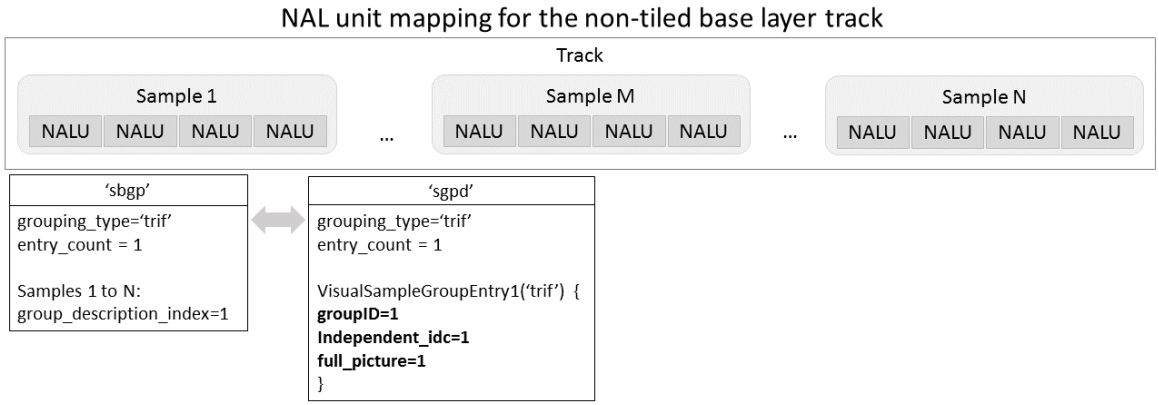


Figure 4: Example of encapsulation of a tiled video layer in a single track

It has to be noted that from w15928, the dependencies to other tiles are put in a single list. Until w15640, we distinguished between intra-layer tile dependencies and inter-layer tile dependencies for some samples (this was the object of the second dependency list: `irap_dependencyTileGroupID`).

Discussion:

independent_flag as defined in w15640 specifies coding dependencies between pictures "of the same layer". We have no indication from the 'trif' that inter-layer dependencies may exist. In the enhancement layer, since the tiles are independent from each other, they should be declared with `independent_idc`=1. Since they have dependency to the base layer (which is described through the tile in the base layer with `full_picture`=1 and `groupID`=1), this means that a '`tsif`' should also be declared to provide the inter-layer tile dependency list.

This is due to:

- the current definition of `independent_idc`: when set to 1, tile region depends only on tile regions with the same `groupID` on the given layer.
- the scope of `groupID`: even if tiles are aligned across layers, they have a different `groupID` value.
- the declaration of dependencies in tsif (then at least one tsif is present).

We then recommend adding a note saying:

"in order to find dependencies to lower layers, the 'tsif' have to be inspected."

3.2.2 Tile to tile inter-layer dependency

This example is close to the previous one, except that this time the base layer also contains tiles, as illustrated on Figure 5.

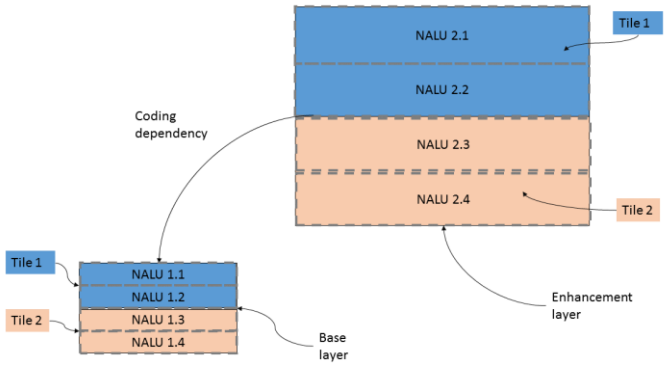


Figure 5: Example of encapsulation of a multi-layered tiled video

The main difference compared to the previous example will be the need for:

- tile description in the base layer track (independently decodable tiles)
- tile to tile dependency signaling in the enhancement layer track.

The Figure 6 illustrates these differences.

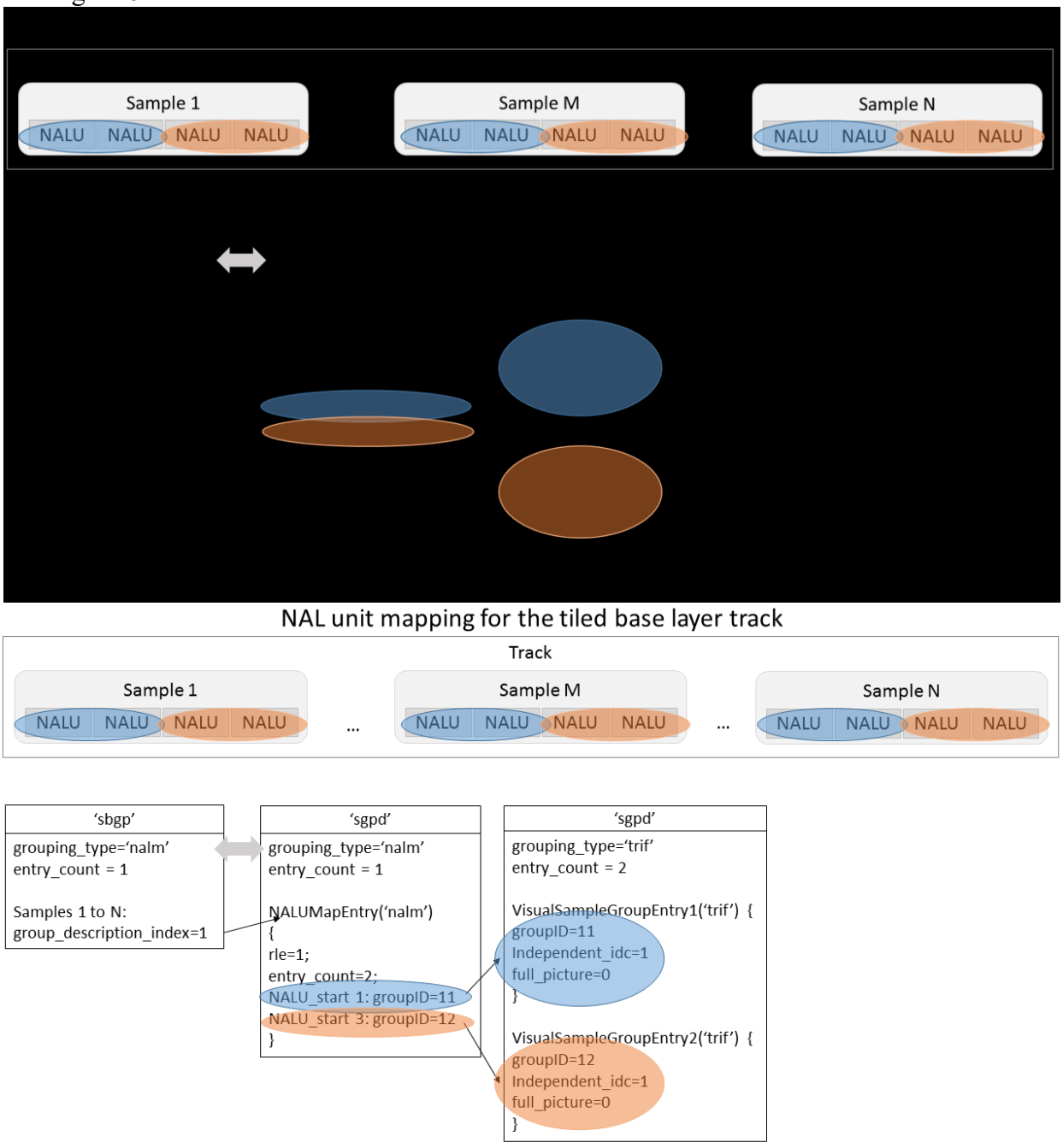


Figure 6: Example of encapsulation of a multi-layered tiled video

Discussion:

Concerning the declaration of dependencies, we have the same issue as discussed in section 3.2.1.

We see through these two examples that the `full_picture` flag enables to describe in a unified way the inter-layer tile dependencies, whatever one layer is tiled or not. Especially, it enables finer description (tile to tile) of the inter-layer dependencies than the simple use of '`linf`'.

Again, a 'tsif' is required in the enhancement layer track to provide the dependency to the tile in the base layer. Contribution m38645 discusses this point.

We then recommend to add the note proposed in section 3.2.1 and to keep the full_picture flag in 'trif'.

3.3 Use case: one ROI and background tiles

Another use case for tiles and NAL unit mapping was suggested after the votes on w15640 in the contribution m38225. It consists in one single track encapsulating four kinds of NAL units:

- Non VCL NAL units applying to whole track like SPS, VPS, PPS ,
- VCL NAL units for a tile region (Region of interest)
- other VCL NAL Units outside the tile region of interest (background).
- Non VCL NAL units applying only to the tile region.

The requirement is to provide easy sub-bitstream extraction of the region of interest.
Let's see how this could be described with current tools from w15928.

Extracting the region of interest from the video track corresponds to a sub-bitstream extraction process. This process, in HEVC, is defined as a "*process by which NAL units in a bitstream that do not belong to a target set, determined by a target highest TemporalId and a target layer identifier list, are removed from the bitstream, with the output sub-bitstream consisting of the NAL units in the bitstream that belong to the target set".* This means that to be able to extract a decodable bitstream for the region of interest, the NAL units have to be mapped as relevant for the region of interest or not. For those applying to the whole track, no specific mapping is required: by default they apply to whole track and then would be kept in the extracted sub-bitstream. For those applying to the region of interest, they could be mapped to a 'trif' providing description of the region of interest. Finally, the remaining NAL units applying to the background could be mapped to one or more other tile region descriptors as shown on Figure 7. These latter tile region descriptors would then contain the list of NAL units to remove during the sub-bitstream extraction process.

We can use the rle mode of the NALUMapEntry to encode the 8 runs of NAL units as shown on the bottom left of Figure 7. Only runs 3 and 6 correspond the region of interest: the corresponding NAL units are then mapped to the tile region descriptor providing the position and sizes of this ROI (the groupID=3 on the right of Figure 7). The remaining VCL NAL units (gray ones) plus eventual non-VLC ones are mapped to different rectangular regions:

- groupID=1 corresponding to the first line of tiles;
- groupID=2 corresponding to the two vertical tiles on the middle left of the picture,
- groupID=4 corresponding to the two vertical tiles on the middle right of the picture and
- groupID=5 corresponding to the last line of tiles.

The non-VCL NAL units are left unmapped so that they keep on applying to the track resulting from any tile selection: ROI only or all tiles.

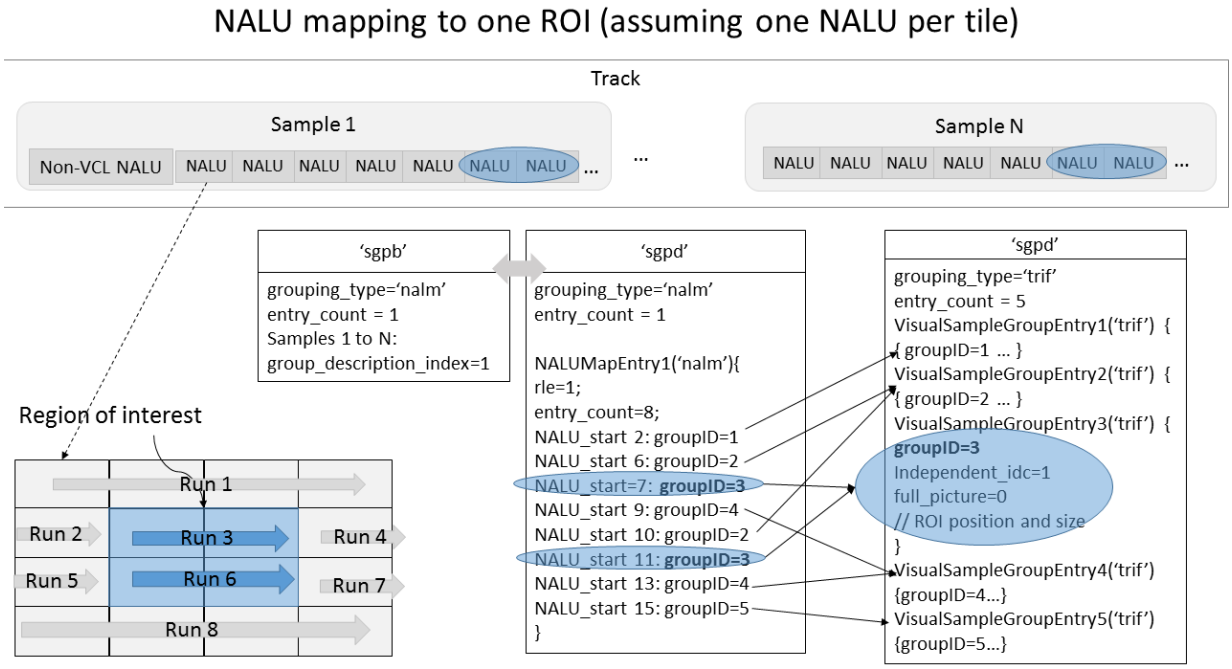


Figure 7: Example of encapsulation of a ROI in a single track

Discussion:

As can be seen, current tools enable the description of a use case where only a subset of tiles is of interest for sub-bitstream extraction. As such, there is no need for another reserved `groupID` value in the `NALUMapEntry` as proposed in m38225.

3.4 Summary on NALUMapEntry use

3.4.1 List of supported features

This considers the NALUMapEntry as currently defined (in both w15640 and w15928).

What can be done

- Mapping a set of NAL units to a 'trif' or 'tsif' `VisualSampleGroupEntry` in one track when each tile is a set of one or more slices (see example 3.1)
- Mapping NAL units in multi-layer tiled video (see examples in 3.2)
- Mapping NAL units from a region of interest for extraction, decoding and display of a subset of tiles in a track (see example 3.3)
-

What cannot be done

- Mapping NAL units to tiles in one track containing multiple tiles when the NAL units contain more than one tile (discussed in section 5).

3.4.2 List of recommendations

Following the above discussions, contribution m38645 should be studied for inter-layer tile dependencies.

From the above, we see that the `full_picture` flag should be kept in '`trif`' since it provides a unified way to describe tile to tile and tile to full-picture dependencies.

The last point is that there is no need to reserve another `groupID` value in `NALUMapEntry`.

4 On the use of tile tracks

4.1 Use case: tile track with a single tile in non-layered HEVC

When each single tile is stored in a dedicated track, the following structures will be found:
A base track containing common information to all the tile tracks, with hev2/hvc2 or lhv1/lhe1 sample entry. Which sample entry to use remains unclear: see contribution m38646 that discusses the appropriate code point to use in tile base track).
One or more tile tracks with '`hvt1`' sample entry and having a track reference type of type '`tbas`' to the base track.
One `trif` descriptor per tile track with a default sample grouping of all the samples of the track.

The tile tracks inherit parameter sets from the tile base track, but there are use cases where all tiles share the same set of properties with the base tile track: sync sample, dependency, sap types, 'rap ' and 'roll', likely most of the defined sample groups (except tiling). Some tables cannot be omitted in a track as their absence already has meaning (namely sync sample table). In order to avoid duplicating this info in NxM tile tracks (N being the number of tiles in horizontal dimension and M the number of tiles in the vertical dimension).

We then propose to add in section 10.6.1 on tile tracks the following text:
"When a sample group description (resp. sample to group) of a given grouping_type value is not present in a tile track but is present in the base tile track, the sample group description (resp sample to group) of the given grouping_type of the base track applies to the samples of this tile track.
For example, if the base tile track has a 'roll' sample group description and the tile track does not, the roll distance for samples in the tile track is the same as the roll distance for samples in the base track."

4.2 Use case: tile track in multi-layer HEVC

The use case described in section 3.2 can also be realized with tile tracks.
In case of a non-tiled base layer, a standard 'hvc1' track would be used containing a single tile region descriptor with the full_picture parameter set to 1. In the enhancement layer, there would be one base track with hev2/hvc2 or lhv1/lhe1 (remains unclear, see m38646) sample entry and as many tile tracks (lht1 sample entry) as tiles in the enhancement layer, each tile track containing a trif plus a tsif for the tile dependencies, in particular to the tile formed by the full picture in the base layer. The two base tile track in the different layers would be related to each other with track reference type to indicate coding dependencies (for example 'scal' when using extractors or through the 'oinf' sample group when using implicit reconstruction).
In case of a tiled base layer, this layer would then be encapsulated into multiple tracks: one base track and multiple tile tracks, each with a trif descriptor with a full_picture parameter set to 0. In the enhancement layer, the dependencies could then be described at tile level rather than at layer level, allowing sub-bitstream extraction for tile-based decoding in enhancement layer (only the referred tiles in the base layer would be needed). As above the two base track would be related to express coding dependencies.

4.3 Summary on tile tracks

Tile tracks offer a convenient way to organize and describe data so that they can be addressed spatially. This is useful for tile-based streaming, or tile-based storage as suggested in m37873.

There are still needs for clarification, for example on the sample entry to use in the tile base track and on the inheritance of properties declared in the tile base track.

5 Multiple tiles per slice (mapping at sub-sample level)

5.1 Combining NALUMapEntry and sub-sample information

Contribution m37873 describes a use case using tiles to provide access to a region of interest in a video. It seems from the description that one slice contains more than one tile, since the authors mention the tile addresses that are given in the slice header. Then, the configuration is like depicted on Figure 8 below, and indeed current tile descriptors in w15928 do not provide this level of description. However, the 'subs' box for HEVC file format natively contains tile information when flags=2: "*Tile-based sub-samples. A sub-sample either contains one tile and the associated non-VCL NAL units, if any, of the VCL NAL unit(s) containing the tile, or contains one or more non-VCL NAL units*".

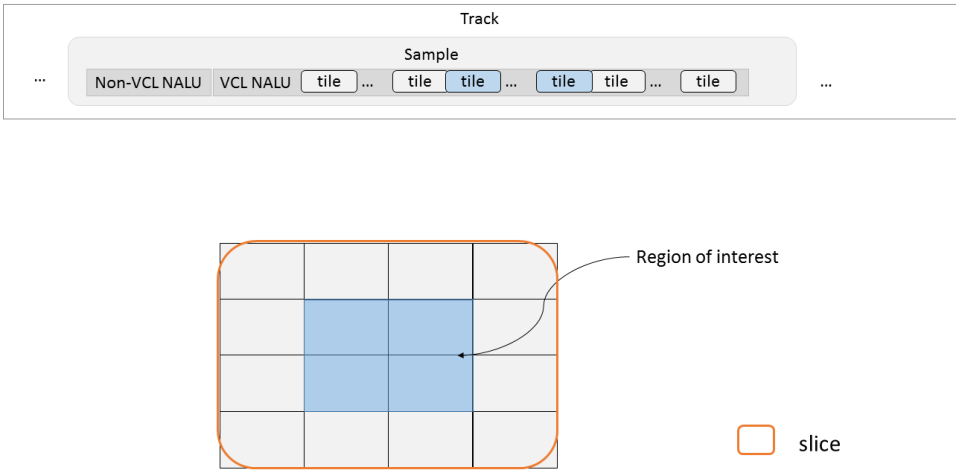


Figure 8: Example of encapsulation of a ROI in a single track

Considering such an encapsulated bitstream (one track with one slice containing multiple tiles), a parser could gain access to one specific tile region through the following steps: first by parsing the sample group descriptions, the parser would gather all the 'trif' descriptions for the track, then obtaining the number of tile regions (like for example the blue region on Figure 8) with their groupID, dependency information and their position and sizes.
In this particular encapsulation, no SampleToGroupBox and no NALUMapEntry would be available to respectively map samples and NAL units to a tile region, since mapping is at sub-sample level. The parser should look at sub-sample level, through 'subs' box with flags=2 and retrieve sub-samples with ctb_x and ctb_y parameters corresponding to the horizontal and vertical offsets of the tile. The parser would then obtain the byte ranges to extract if the application wants to record the sub-bitstream corresponding to one specific tile region.

While feasible with current tools, the access to one specific tile region when tiles are contained in one slice is not direct (requiring loop on subs box to decode and match tile positions). The decoding and matching of the coordinates can be avoided by considering the previous proposal in m31439 where a new value of flags is reserved to associate sub-sample to a groupID (for example for a tile region identifier):

```
} else if (flags == 6) {
```

```
        unsigned int(16) groupID;
        unsigned int(16) reserved;
    }
```

With the following semantics:

"6: groupID based sub-samples. A sub-sample is mapped to a HEVC tile region identified by its groupID.

A track may associate full sample to a given tile region using trif/tsif or individual NALUs to a given tile region (using NALUMapEntry, trif/tsif), while at the same time associating subsamples to other tiles. If this is the case, the tile description shall be correct for all mapping, i.e. the tile indicated by the sub-sample description shall match or be included by the tile indicated through sample grouping tools
"

Discussion:

This use case could be supported with existing tools, but it requires some parsing and check of tile parameters. If we want a more direct mapping between sub-sample and tiles, the proper way to do this is by using sub-sample tools.

Moreover, we don't think it is a good design to have different levels of mapping in NALUMapEntry as suggested in m38225:

- NALUMap Entry should exclusively remain mapping of NAL units.
- All these approaches rely on the presence of subsample box (to get the subsample index), therefore having yet another indirection through NALU doesn't solve anything

Finally if mapping at sub-sample level is really needed, a better design is to consider extension of 'subs' box as suggested above.

6 Conclusion

This contribution explained current tools for file description in ISOBMFF files. For additional use cases that would not be possible with current tools, we suggest to address these through new proposals and eventually consider them in a future amendment. We think last minute modifications to try to cover more use cases may break the initial design that took care of flexibility for description depending on different use cases.

7 Additional comments on w15928

Shouldn't we define HEVC NAL units in Section 3 (was done for AVC, 3D-AVC, MVC, MVD and SVC)

8.2: A non-layered video stream is represented by one video track in a file**, except in the case of tiles (see Section 10)**