ETSI TS 126 118 V16.1.1 (2021-01)

5G;
Virtual Reality (VR) profiles for streaming applications
(3GPP TS 26.118 version 16.1.1 Release 16)

Reference
RTS/TSGS-0426118vg11

Keywords
5G

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Legal Notice

This Technical Specification (TS) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Specification has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

Introduction

The present document provides technologies for interoperable Virtual Reality services with focus on streaming and consumption.

Virtual Reality (VR) is the ability to be virtually present in a space created by the rendering of natural and/or synthetic image and sound correlated by the movements of the immersed user allowing interacting with that world.

Suitable media formats for providing immersive experiences are specified to enable Virtual Reality Services in the context of 3GPP bearer and user services.

1 Scope

The present document defines interoperable formats for Virtual Reality for streaming services. Specifically, the present document defines operation points, media profiles and presentation profiles for Virtual Reality. The present document builds on the findings and conclusions in TR 26.918 [2].

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document.*

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TR 26.918: "Virtual Reality (VR) media services over 3GPP".

[3] Recommendation ITU-R BT.709-6 (06/2015): "Parameter values for the HDTV standards for production and international programme exchange".

[4] Recommendation ITU-R BT.2020-2 (10/2015): "Parameter values for ultra-high definition television systems for production and international programme exchange".

[5] Recommendation ITU-T H.264 (04/2017): "Advanced video coding for generic audiovisual services" | ISO/IEC 14496-10:2014: "Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding".

[6] Recommendation ITU-T H.265 (02/2018): "High efficiency video coding" | ISO/IEC 23008-2:2018: "High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 2: High Efficiency Video Coding".

[7] void.

[8] 3GPP TS 26.247: "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[9] ISO/IEC 14496-15: "Information technology - Coding of audio-visual objects - Part 15: Carriage of network abstraction layer (NAL) unit structured video in ISO base media file format".

[10] ISO/IEC 23001-8: "Information technology -- MPEG systems technologies -- Part 8: Coding-independent code points".

[11] Recommendation ITU-R BT.2100-1: "Image parameter values for high dynamic range television for use in production and international programme exchange".

[12] 3GPP TS 26.116: "Television (TV) over 3GPP services; Video profiles".

[13] ISO/IEC 23090-2: "Coded representation of immersive media -- Part 2: Omnidirectional media format".

[14] ISO/IEC DIS 23091-2: "Information technology -- Coding-independent code points -- Part 2: Video".

[15] 3GPP TS 26.260: "Objective test methodologies for the evaluation of immersive audio systems".

[16] 3GPP TS 26.259: "Subjective test methodologies for the evaluation of immersive audio systems".

[17] ISO/IEC 14496-12: "Information technology -- Coding of audio-visual objects -- Part 12: ISO base media file format".

[18] ISO/IEC 23009-1: "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats".

[19] ISO/IEC 23008-3:2015: "Information technology -- High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio", ISO/IEC 23008-3:2015/Amd2:2016: "MPEG-H 3D Audio File Format Support ", ISO/IEC 23008-3:2015/Amd 3:2017: "MPEG-H 3D Audio Phase 2", ISO/IEC 23008-3:2015/Amd 5: "Audio metadata enhancements".

[20] IETF RFC 6381: "The 'Codecs' and 'Profiles' Parameters for "Bucket" Media Types", R. Gellens, D. Singer, P. Frojdh, August 2011.

[21] AES69-2015: "AES standard for file exchange - Spatial acoustic data file format", 2015.

3 Definitions, symbols and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

bitstream: a bitstream that conforms to a video encoding format and certain Operation Point.

field of view: the extent of visible area expressed with vertical and horizontal angles, in degrees in the 3GPP 3DOF reference system.

operation point: a collection of discrete combinations of different content formats including spatial and temporal resolutions, colour mapping, transfer functions, rendering metadata and the encoding format.

pose: position derived by the head tracking sensor expressed by (azimuth; elevation; tilt angle).

receiver: a receiver that can decode and render any bitstream that is conforming to a certain Operation Point.

viewport: the part of the 3DOF content to render based on the pose and the field of view.

3.2 Symbols

For the purposes of the present document, the following symbols apply:

α	yaw of the 3GPP 3DOF coordinate system
β	pitch of the 3GPP 3DOF coordinate system
γ	roll of the 3GPP 3DOF coordinate system
ϕ	azimuth of the 3GPP 3DOF coordinate system
θ	elevation of the 3GPP 3DOF coordinate system

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

3DOF	3 Degrees of freedom
ACN	Ambisonics Channel Number
API	Application Programming Interface
AVC	Advanced Video Coding
BMFF	Base Media File Format
BRIR	Binaural Room Impulse Response
CMP	Cube-Map Projection

CIBR	Common Informative Binaural Renderer
DASH	Dynamic Adaptive Streaming over HTTP
DRC	Dynamic Range Control
EOTF	Electro-Optical Transfer Function
ERP	EquiRectangular Projection
ESD	Equivalent Spatial Domain
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FOA	First Order Ambisonics
FOV	Field Of View
GPU	Graphics Processing Unit
HDR	High Dynamic Range
HDTV	High Definition TeleVision
HEVC	High Efficiency Video Coding
HMD	Head Mounted Display
HOA	High Order Ambisonics
HRD	Hypothetical Reference Decoder
HRIR	Head-Related Impulse Responses
HRTF	Head-Related Transfer Function
HTTP	HyperText Transfer Protocol
IFFT	Inverse FFT
IRFFT	Inverse RFFT
MAE	MPEG-H Audio Metadata information
MCC	Metrics Collection and Computation
MHAS	MPEG-H Audio Stream
MIME	Multipurpose Internet Mail Extensions
MPD	Media Presentation Description
MPEG	Moving Pictures Experts Group
NAL	Network Abstraction Layer
OMAF	Omnidirectional MediA Format
PCM	Pulse Code Modulation
RAP	Random Access Point
RFFT	Real FFT
RWP	Region-Wise Packing
SDR	Standard Dynamic Range
SEI	Supplemental Enhancement Information
SN3D	Schmidt semi-normalisation
SOFA	Spatially Oriented Format for Acoustics
SPS	Sequence Parameter Set
SRQR	Spherical Region-wise Quality Ranking
VCL	Video Coding Layer
VST	Virtual Studio Technology
VUI	Video Usability Information
VR	Virtual Reality

4 Architectures and Interfaces for Virtual Reality

4.1 Definitions and Reference Systems

4.1.1 Overview

Virtual reality is a rendered version of a delivered visual and audio scene. The rendering is designed to mimic the visual and audio sensory stimuli of the real world as naturally as possible to an observer or user as they move within the limits defined by the application.

Virtual reality usually, but not necessarily, assumes a user to wear a head mounted display (HMD), to completely replace the user's field of view with a simulated visual component, and to wear headphones, to provide the user with the accompanying audio as shown in Figure 4.1-1.

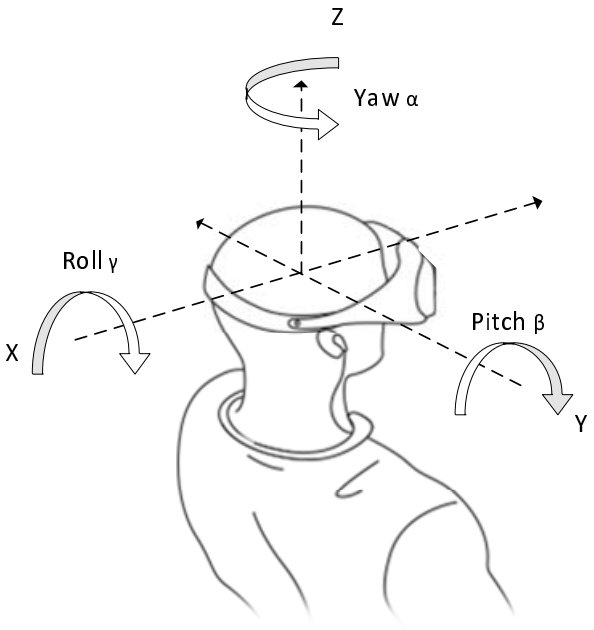

Figure 4.1-1: Reference System

Some form of head and motion tracking of the user in VR is usually also necessary to allow the simulated visual and audio components to be updated in order to ensure that, from the user's perspective, items and sound sources remain consistent with the user's movements. Sensors typically are able to track the user's pose in the reference system. Additional means to interact with the virtual reality simulation may be provided but are not strictly necessary.

VR users are expected to be able to look around from a single observation point in 3D space defined by either a producer or the position of one or multiple capturing devices. When VR media including video and audio is consumed with a head-mounted display or a smartphone, only the area of the spherical video that corresponds to the user's viewport is rendered, as if the user were in the spot where the video and audio were captured.

This ability to look around and listen from a *centre point* in 3D space is defined as 3 degrees of freedom (3DOF). According to the figure 4.1-1:

- tilting side to side on the X-axis is referred to as *Roll*ing, also expressed as γ
- tilting forward and backward on the Y-axis is referred to as *Pitch*ing, also expressed as β
- turning left and right on the Z-axis is referred to as *Yaw*ing, also expressed as α

It is worth noting that this *centre point* is not necessarily static - it may be moving. Users or producers may also select from a few different observational points, but each observation point in 3D space only permits the user 3 degrees of freedom. For a full 3DOF VR experience, such video content may be combined with simultaneously captured audio, binaurally rendered with an appropriate Binaural Room Impulse Response (BRIR). The third relevant aspect is the interactivity: Only if the content is presented to the user in such a way that the movements are instantaneously reflected in the rendering , then the user will perceive a full immersive experience. For details on immersive rendering latencies, refer to TR 26.918 [2].

4.1.2 3GPP 3DOF Coordinate System

The coordinate system is specified for defining the sphere coordinates azimuth (ϕ) and elevation (θ) for identifying a location of a point on the unit sphere, as well as the rotation angles yaw ($\alpha\alpha$), pitch ($\beta\beta$), and roll ($\gamma\gamma$). The origin of the coordinate system is usually the same as the centre point of a device or rig used for audio or video acquisition as well as the position of the user's head in the 3D space in which the audio or video are rendered. Figure 4.1-2 specifies principal axes for the coordinate system. The X axis is equal to back-to-front axis, Y axis is equal to side-to-side (or lateral) axis, and Z axis is equal to vertical (or up) axis. These axis map to the reference system in Figure 4.1-1.

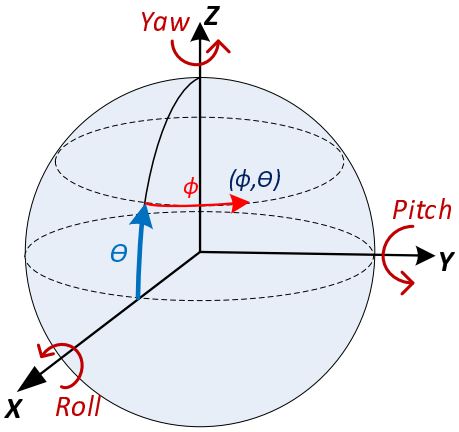

Figure 4.1-2: Coordinate system

Signals defined in the present document are represented in a spherical coordinate space in angular coordinates (ϕ, Θ) for use in omnidirectional video and 3D audio. The viewing and listing perspective are from the origin sensing/looking/hearing outward toward the inside of the sphere. Even though a spherical coordinate is generally represented by using radius, elevation, and azimuth, it assumes that a unit sphere is used for capturing and rendering of VR media. Thus, a location of a point on the unit sphere is identified by using the sphere coordinates azimuth (ϕ) and elevation (θ). The spherical coordinates are defined so that ϕ is the azimuth and Θ is the elevation. As depicted in Figure 4.1-2, the coordinate axes are also used for defining the rotation angles yaw ($\alpha\alpha$), pitch ($\beta\beta$), and roll ($\gamma\gamma$). The angles increase clockwise when looking from the origin towards the positive end of an axis. The value ranges of azimuth, yaw, and roll are all −180.0, inclusive, to 180.0, exclusive, degrees. The value range of elevation and pitch are both −90.0 to 90.0, inclusive, degrees.

Depending on the applications or implementations, not all angles may be necessary or available in the signal. The 360 video may have a restricted *coverage* as shown in Figure 4.1-3. When the video signal does not cover the full sphere, the coverage information is described by using following parameters:

- *centre azimuth*: specifies the azimuth value of the centre point of sphere region covered by the signal.
- *centre elevation*: specifies the elevation value of the centre of sphere region.
- *azimuth range*: specifies the azimuth range through the centre point of the sphere region.
- *elevation range*: specifies the elevation range through the centre point of the sphere region.
- *tilt angle*: indicates the amount of tilt of a sphere region, measured as the amount of rotation of the sphere region along the axis originating from the origin passing through the centre point of the sphere region, where the angle value increases clockwise when looking from the origin towards the positive end of the axis.

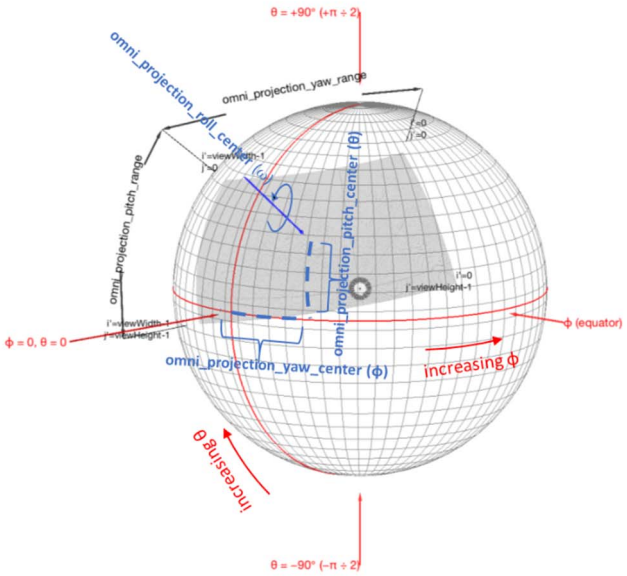

Figure 4.1-3: Restricted coverage of the sphere region covered by the cropped output picture with omni_projection_{yaw | pitch | roll}_center the center of the coverage region.

For video, such a centre point may exist for each eye, referred to as *stereo* signal, and the video consists of three color components, typically expressed by the luminance (Y) and two chrominance components (U and V).

The coordinate systems for all media types are assumed to be aligned in 3GPP 3DOF coordinate system. Within this coordinate system, the *pose* is expressed by a triple of azimuth, elevation, and tilt angle characterizing the head position of a user consuming the audio-visual content. The pose is generally dynamic, and the information may be provided through sensors in a frequently sampled version.

The *field of view (FoV)* of a rendering device is static and defined in two dimensions, the horizontal and vertical FoV, each in units of degrees in the angular coordinates (ϕ, θ). The pose together with the field of view of the device enables the system to generate the user viewport, i.e., the presented part of the content at a specific point in time.

4.1.3 Video Signal Representation

Commonly used video encoders cannot directly encode spherical videos, but only 2D textures. However, there is a significant benefit to reuse conventional 2D video encoders. Based on this, Figure 4.1-4 provides the basic video signal representation in the context of omnidirectional video in the context of the present document. By pre-processing, the spherical video is mapped to a 2D texture. The 2D texture is encoded with a regular 2D video encoder and the VR rendering metadata (i.e. the data describing the mapping from the spherical coordinate to the 2D texture) is encoded and provided along with the video bitstream, such that at the receiving end the inverse process can be applied to reconstruct the spherical video.

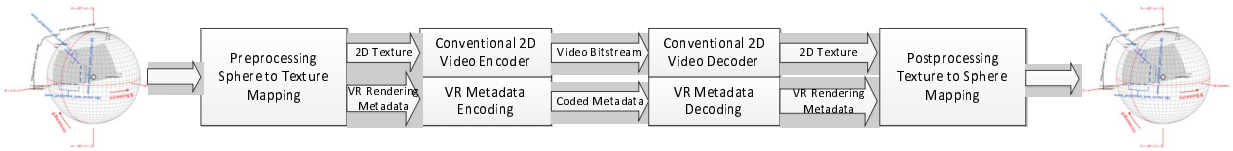

Figure 4.1-4: Video Signal Representation

Mapping of a spherical picture to a 2D texture signal is illustrated in Figure 4.1-5. The most commonly used mapping from spherical to 2D is the equirectangular projection (ERP) mapping. The mapping is bijective, i.e. it may be expressed in both directions.

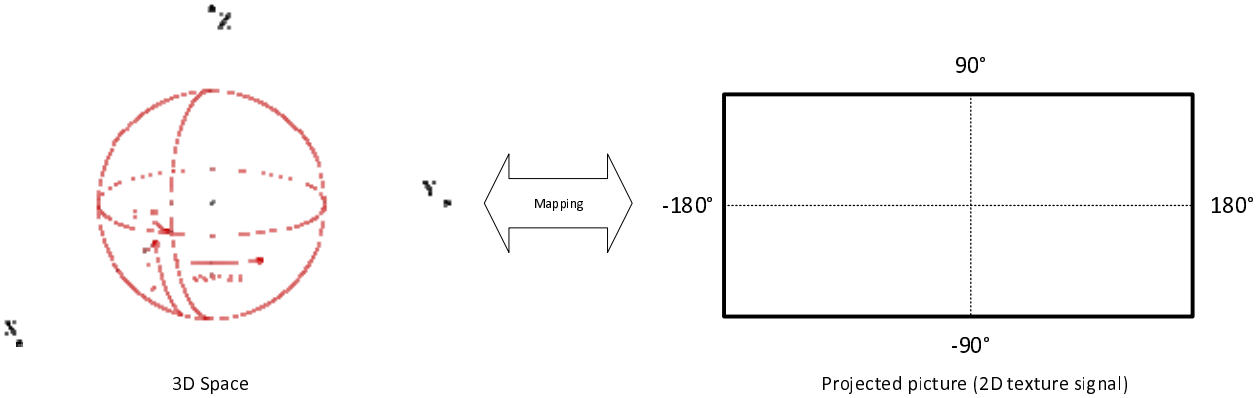

Figure 4.1-5: Examples of Spherical to 2D mappings

Following the definitions in clause 4.1.2, the mapping of the color samples of 2D texture images onto a spherical coordinate space in angular coordinates (ϕ, θ) for use in omnidirectional video applications for which the viewing perspective is from the origin looking outward toward the inside of the sphere. The spherical coordinates are defined so that ϕ is the azimuth and θ is the elevation.

Assume a 2D texture with `pictureWidth` and `pictureHeight`, being the width and height, respectively, of a monoscopic projected luma picture, in luma samples and the center point of a sample location (i, j) along the horizontal and vertical axes, respectively, then for the *equirectangular* projection the sphere coordinates (ϕ, θ) for the luma sample location, in degrees, are given by the following equations:

$$\phi = (0.5 - i \div pictureWidth) * 360$$
$$\theta = (0.5 - j \div pictureHeight) * 180$$

Whereas ERP is commonly used for production formats, other mappings may be applied, especially for distribution. The present document also introduces cubemap projection (CMP) for distribution in clause 5. In addition to regular projection, other pre-processing may be applied to the spherical video when mapped into 2D textures. Examples include region-wise packing, stereo frame packing or rotation. The present document defines different pre- and post-processing schemes in the context of video rendering schemes.

4.1.4 Audio Signal Representation

Audio for VR can be produced using three different formats. These are broadly known as channels-, objects- and scene-based audio formats. Audio for VR can use any one of these formats or a hybrid of these (where all three formats are used to represent the spherical soundfield). The audio signal representation model is shown in Figure 4.1-6.

The present document expects that an audio encoding system is capable to produce suitable audio bitstreams that represent a well-defined audio signal in the reference system as defined in clause 4.1.1. The coding and carriage of the VR Audio Rendering Metadata is expected to be defined by the VR Audio Encoding system. The VR Audio Receiving system is expected to be able to use the VR Audio Bitstream to recover audio signals and VR Audio Rendering metadata. Both signals, audio signals and metadata, are well-defined by the media profile, such that different audio rendering systems may be used to render the audio based on the decoder audio signals, VR audio rendering metadata and the user position.

In the present document, all media profiles are defined such that for each media profile at least one Audio Rendering System is defined as a reference renderer and additional Audio Rendering systems may be defined. The audio rendering system is described based on well-defined output of the VR Audio decoding system.

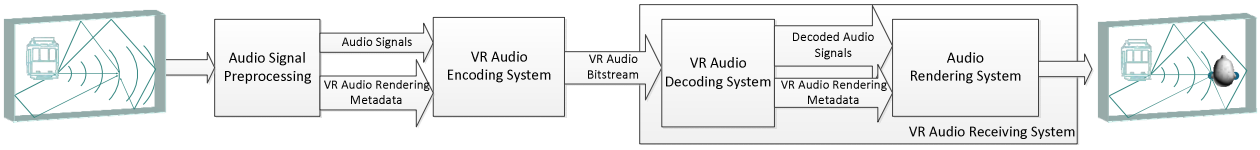

Figure 4.1-6: Audio Signal Representation

For more details on audio rendering, refer to clause 4.5.

4.2 End-to-end Architecture

The architecture introduced in this clause addresses service scenarios for the distribution of VR content in file-based download and DASH-based streaming services.

Figure 4.2-1 considers a functional architecture for such scenarios. VR Content is acquired and the content is pre-processed such that all media components are mapped to the 3GPP 3DOF coordinate system and are temporarily synchronized. Such pre-processing may include video stitching, rotation or other translations. The 3GPP VR Headend is responsible for generating content that can be consumed by receivers conforming to the present document. Typically, 3D Audio and spherical video signals are properly encoded. Especially for video, the processing follows the two step approach of mapping, projecting and pre-processing to 2D texture and then encoding with regular 2D video codecs. After media encoding, the content is made available to file format encapsulation engine as elementary streams. The encapsulated streams are referred to as 3GPP VR Tracks, i.e. they are spatially mapped to the same timing system for synchronized playback. For file based distribution a complete file for delivery is generated by multiplexing the 3GPP VR tracks into a single file. For DASH based delivery, the content is mapped to DASH segments and proper Adaptation Sets are generated, including the necessary MPD signaling. The Adaptation Sets are included in a VR Media Presentation, documented in a DASH MPD. Content may be made available such that it is optimized for a specific viewpoint, so the same content may be encoded in an ensemble of multiple viewport-optimized versions.

The content is delivered through file based delivery of DASH based delivery, potentially using 3GPP services such as DASH in PSS or DASH-over-MBMS.

At the receiving end, a VR application is assumed that communicates with the different functional blocks in the receivers' 3GPP VR service platform, namely, the DASH client or the download client, the file processing units for each media profile, the media decoding units, the rendering environment and the pose generator. The reverse operations of the VR Headend are performed. The operation is expected to be dynamic, especially taking into account updated pose information in the different stages of the receiver. The pose information is essential in the rendering units, but may also be used in the download or DASH client for delivery and decoding optimizations. For more details on the client reference architecture, refer to clause 4.3.

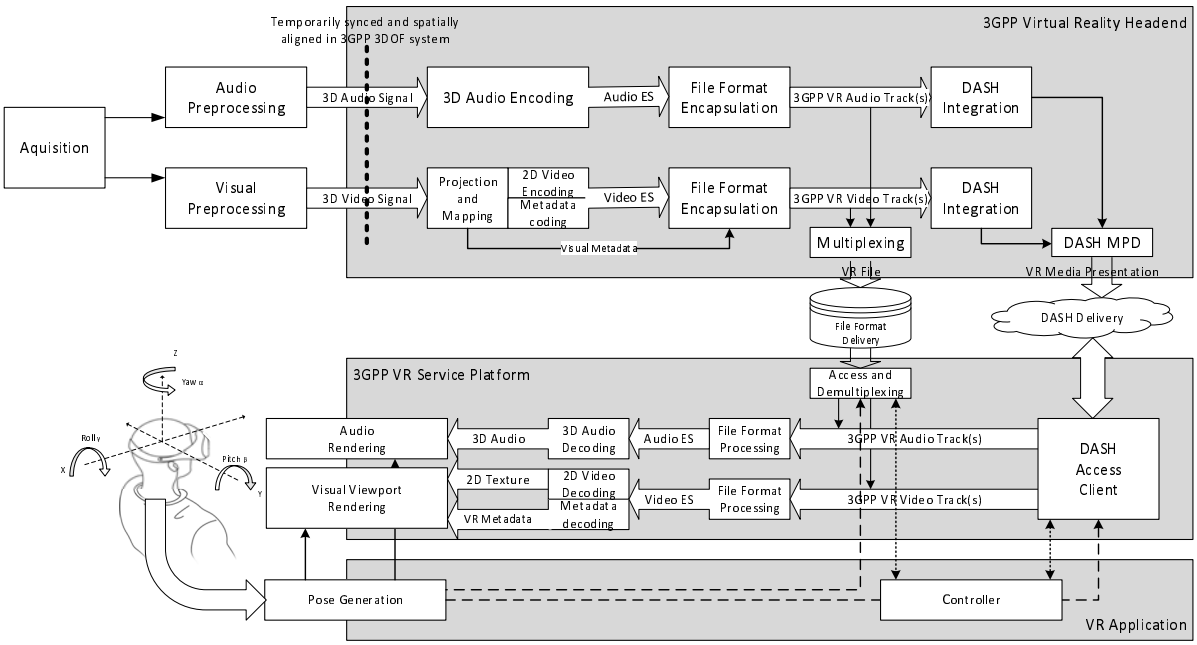

Figure 4.2-1: architecture for VR streaming services

Based on the architecture in Figure 4.2-1, the following components are relevant for 3GPP VR Streaming Services:

- Consistent source formats that can be distributed by a 3GPP VR Headend:
  - For audio that can be used by a 3D audio encoding profile according the present document.

  - For video that can be used by a spherical video encoding profile according to the present document.

- Mapping formats from a 3-dimensional representation to a 2D representation in order to use regular video encoding engines

- Encapsulation of the media format tracks to ISO file format together, adding sufficient information on to decode and render the VR content. The necessary metadata may be on codec level, file format level, or both.

- Delivery of the formats through file download, DASH delivery and DASH-over-MBMS delivery.

- Static and dynamic capabilities and environmental data, including decoding and rendering capabilities, as well as dynamic pose information.

- Media decoders that support the decoding of the formats delivered to the receiver.

- Information for audio and video rendering to present the VR Presentation on the VR device.

Based on the considerations above, to support the use case of VR Streaming, the following functions are defined in the present document:

- Consistent content contribution formats for audio and video for 360/3D AV applications including their metadata. This aspect should be considered informative, but example formats are provided to enable explaining the workflow.

- Efficient encoding of 360 video content. In the present document, this encoding is split in two steps, namely a pre-processing and projection mapping from 360 video to 2D texture and a regular video encoding.

- Efficient encoding of 3D audio including channels, objects and scene-based audio.

- Encapsulation of VR media into a file format for download delivery.

- The relevant enablers for DASH delivery of VR experiences.

- The necessary capabilities for static and dynamic consumption of the encoded and delivered experiences in the Internet media type and the DASH MPD.

- A reference client architecture that provides the signalling and processing steps for download delivery as well as DASH delivery as well as the interfaces between the 3GPP VR service platform, a VR application (e.g. pose information), and the VR rendering system (displays, GPU, loudspeakers).

- Decoding requirements for the defined 360 video formats.

- Decoding requirements for the defined 3D audio formats.

- Rendering requirements or recommendations for the above formats, for both separate and integrated decoding/rendering.

4.3 Client Reference Architecture

This clause provides more details of a client reference architecture for VR streaming applications and describes their components and interfaces.

Figure 4.3-1 and Figure 4.3-2 show a high-level structure of the client reference architecture for VR DASH streaming and VR local playback, respectively, which consist of five functional components:

- *VR Application*: The VR application controls the rendering depending on a user viewport or display capabilities. The application may communicate with other functional components, e.g., the access engine, the file decoder. The access engine or file decoder may parse some abstracted control information to the VR application and the application makes the decision on which adaptation sets or preselections to select or which tracks to choose taking into account platform or user information as well as the dynamic pose information.

- *Access Engine*: The access engine connects through a 3GPP bearer and provides a conforming VR presentation to the receiver. The access engine fetches the Media Presentation Description (MPD), constructs and issues requests and receives Segments or parts of Segments. In the case of local playback, the 3GPP VR Track is accessed from the local storage. The access engine may interface with the VR application function to

dynamically change the delivery session. The access engine provides a conforming 3GPP VR track to the file decoder.

- *File Decoder*: The file decoder processes the 3GPP VR Track to generate signals that can be processed by the renderer. The file decoder typically includes at least of two sub-modules; the file parser and the media decoder. The file parser processes the file or segments, extracts elementary streams, and parses the metadata, if present. The processing may be supported by dynamic information provided by the VR application, for example which tracks to choose based on static and dynamic configurations. The media decoder decodes media streams of the selected tracks into the decoded signals. The file decoder outputs the decoded signals and metadata which is used for rendering. The file decoder is the primary focus of the present document.

- *VR Renderer*: The VR Renderer uses the decoded signals and rendering metadata and provides a viewport presentation taking into account the viewport and possible other information. With the pose, a user viewport is determined by determining horizontal/vertical field of view of the screen of a head-mounted display or any other display device to render the appropriate part of decoded video or audio signals. The renderer is addressed in individual media profiles. For video, textures from decoded signals are projected to the sphere with rendering metadata received from the file decoder. During the texture-to-sphere mapping, a sample of the decoded signal is remapped to a position on the sphere. Likewise, the decoded audio signals are represented in the reference system domain. The appropriate part of video and audio signals for a current pose is generated by synchronizing and spatially aligning the rendered video and audio.

- *Sensor:* The sensor extracts the current pose according to the user's movement and provides it to the renderer for viewport generation. The current pose may for example be determined by the head tracking and possibly also eye tracking functionalities. The current pose may also be used by the VR application to control the access engine on which adaptation sets or preselections to select (for the streaming case), or to control the file decoder on which tracks to choose for decoding (for the local playback case).

The main objective of the present document is to enable the file decoder to generate decoded signals and the rendering metadata from a conforming 3GPP VR Track by generating a bitstream that conforms to a 3GPP Operation Point. Both, a 3GPP VR Track as well as a bitstream conforming to an Operation Point are a well-defined conformance points for a VR File decoder and a Media Decoder. Both enable to represent the contained media in the VR reference system (spatially and temporally).

NOTE 1: 3GPP VR Track represents media in container formats according to the ISO/IEC 14496-12 [17] ISO Base Media File Format and may consist of one or more ISO BMFF tracks following the requirements of this specification.

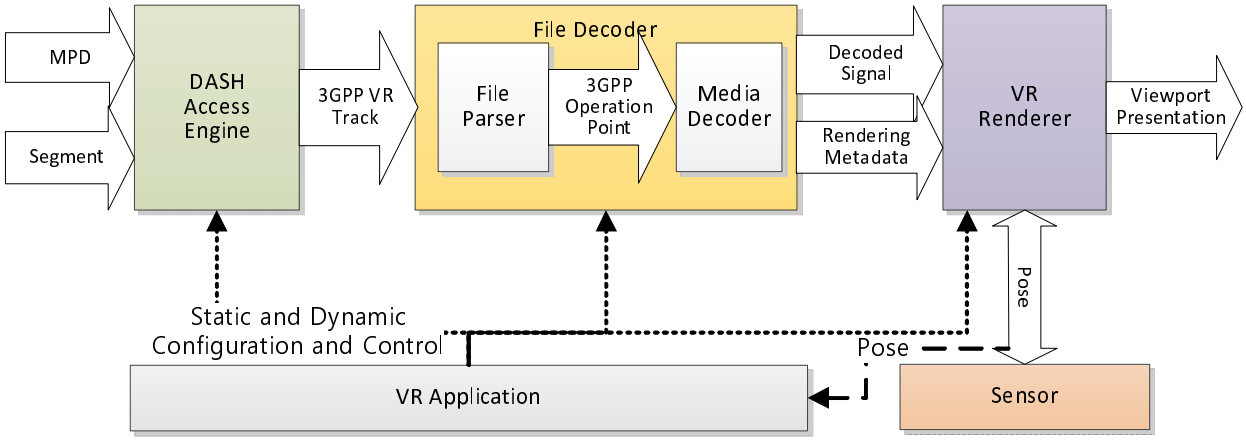

Figure 4.3-1: Client Reference Architecture for VR DASH Streaming Applications

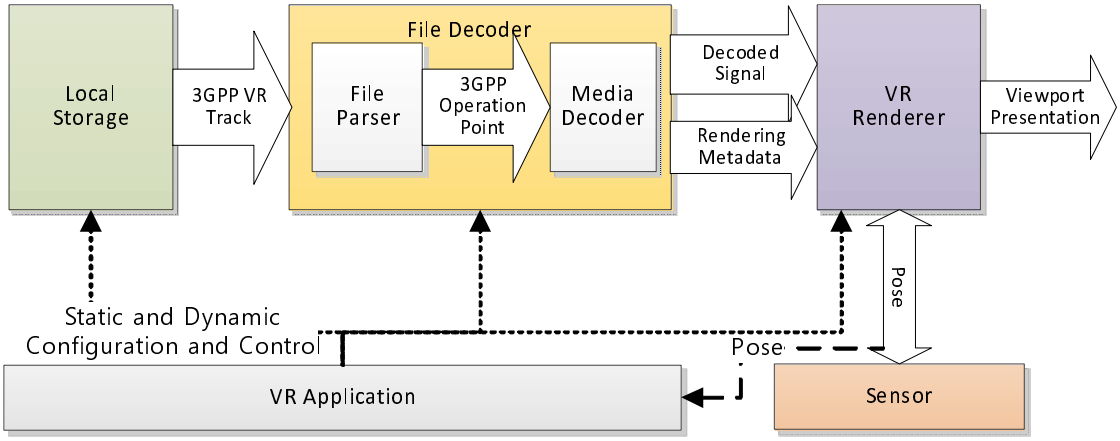

Figure 4.3-2: Client Reference Architecture for VR Local Playback

NOTE 2: The dashed arrows indicate optional interfaces between components in the Figure 4.3-2. Viewport information should optionally be input to the access engine and file decoder.

4.4 Rendering Schemes, Operation Points and Media Profiles

The present document provides several interoperability points that may be referred external specifications. These are:

- **Media profiles:** providing DASH, file format and elementary stream constraints for a single media type.
- **Operation Points**: a collection of discrete combinations of different content formats including spatial and temporal resolutions, colour mapping, transfer functions, rendering metadata and the encoding format.
- **Bitstream:** A video bitstream that conforms to a video encoding format and certain Operation Point including VR Rendering Metadata.
- **Rendering Scheme**: post-decoder processing of decoder output signals together with rendering metadata.

Note that this applies to both media types, audio and video. For audio, the 3GPP VR Rendering Scheme interoperability point serves as an informative output of the File Decoder. The 3GPP VR viewport interoperability point serves as the output of the entire file decoding process.

Both features provide clear requirements for interoperability for receiver. Figure 4.4-1 provides an overview on this.

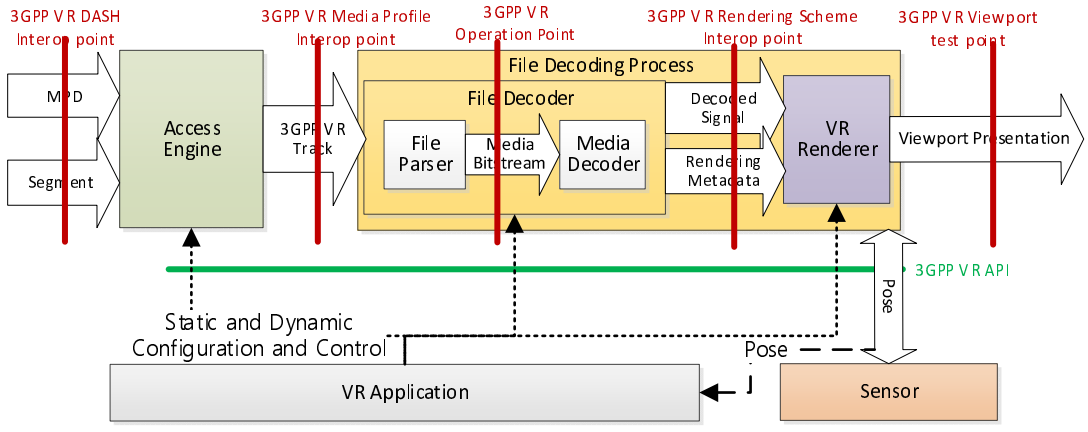

Figure 4.4-1: Interoperability aspects for 3GPP VR Profiles

Media profile for timed media is defined as requirements and constraints for a set of one or more 3GPP VR tracks of a single media type. The conformance of a set of one or more 3GPP VR tracks to a media profile is specified as a combination of:

- Specification of which sample entry type(s) are allowed, and which constraints and extensions are required in addition to those imposed by the sample entry type(s).

- Constraints on the samples of the tracks, typically expressed as constraints on the elementary stream contained within the samples of the tracks.

The elementary stream constraints of a media profile may be indicated by a requirement to comply with a certain profile and level of the media coding specification, possibly including additional constraints and extensions, such as a requirement of the presence of certain information for rendering and presentation.

Each media profile specified in the present document includes a file decoding process such that all file decoders that conform to the video media profile will produce:

- For video: numerically identical cropped decoded pictures when invoking the file decoding process associated with that video media profile for a set of 3GPP VR tracks conforming to the video media profile. A bitstream that conforms to the elementary stream constraints specified for the video media profile is reconstructed as an intermediate product of the file decoding process. Output of the file decoding process consists of all of the following:

  - a list of decoded pictures with associated presentation times;

  - for projected omnidirectional video VR rendering metadata.

- for audio: a set of audio signals when invoking the file decoding process associated with that audio media profile for a VR Track conforming to the audio media profile. A bitstream that conforms to the elementary stream constraints specified for the audio media profile is reconstructed as an intermediate product of the file decoding process. Output of the file decoding process consists of all of the following:

  - a sequence of audio samples with associated presentation times;

  - audio VR rendering metadata.

A file decoder conforms to the file decoding process requirements of the present document when it complies with both of the following:

- for video:

  - The file decoder includes a conforming decoder that produces numerically identical cropped decoded pictures to those produced by the file decoding process specified for the video media profile in clause 5 (with the correct output order or output timing, as specified in the video coding specification of the video media profile, respectively).

  - The file decoder outputs rendering metadata that is equivalent to that produced by the file decoding process specified for the video media profile in clause 5 (with the correct association of the rendering metadata to particular cropped decoded pictures, as specified in the present document).

- for audio:

  - The file decoder includes a conforming decoder that produces a sequence of audio samples with associated presentation times as defined in clause 6.

  - The file decoder outputs Audio rendering metadata that is equivalent to that produced by the file decoding process specified for the audio media profile in clause 6 (with the correct association of the rendering metadata to particular audio samples).

A player claiming conformance to a media profile shall include a file decoder complying with the file decoding process of that video media profile as specified above. While the player operation, with the exception of the file decoding process, is not specified normatively in the present document, specifications of a media profile may include an informative clause on expectations of a player operation, for example including recommendations for rendering.

In addition to the interoperability on track level, also a DASH level interoperability for each media profile is defined. This interoperability includes the signalling and content generation, such that by dynamic switching based on network constraints or sensor input a conforming 3GPP VR Track for this media profile may be obtained.

4.5 Audio Rendering

4.5.1 Audio Renderer Definitions

4.5.1.1 Reference Renderer

The purpose of the *Reference Renderer* is to provide a documented audio rendering solution in 3GPP for its corresponding media profile. A *Reference Renderer*:

a) Is specified along with the media profile.

b) Supports binaural and loudspeaker based rendering.

c) Has a standardized implementable description documented either in 3GPP or in an external SDO.

d) Supports diegetic and non-diegetic content.

e) Has a Motion to Sound latency characterized according to the method defined in TS 26.260 [15].

f) Has a Loudness characterized according to the method defined in TS 26.260 [15].

g) Provides a suitable subjective quality level characterized by the *Rendering Test* (see clause 4.5.1.6).

h) Provides an interface to specify the set of HRTFs used for binaural rendering.

NOTE: The *Reference Renderer* could be an external renderer following the properties defined above.

4.5.1.2 Common Informative Binaural Renderer (CIBR)

The CIBR is a binaural renderer defined for the purposes of the Renderer Test in TS 26.259 [16]. The CIBR:

a) Supports binaural rendering.

b) Supports diegetic and non-diegetic content.

c) Has a Motion to Sound latency characterized according to the method defined in TS 26.260 [15].

d) Has a Loudness characterized according to the method defined in TS 26.260 [15].

e) Is intended to provide a quality comparison point for the *Reference Renderer* (see clause 4.5.1.1) and any *External Renderer* (see clause 4.5.1.3).

The CIBR consists of four components. The first three are currently available as VST audio plugins:

1) The *"ESD to HOA"* component, which receives a set of audio input signals in an Equivalent Spatial Domain (ESD) representation and converts them into a set of audio output signals in the HOA domain (ACN/SN3D format). The ESD representation corresponds to the immersive audio content rendered at a set of pre-determined virtual loudspeaker locations (Fliege Points) The ESD to HOA conversion is accomplished using the "AmbiX Decoder" plugin (https://github.com/kronihias/ambix) with the appropriate conversion matrices specified in TS 26.260 [15] clause 4.1 (Fliege Points).

2) The *"Sound Field rotation"* component, which performs rotation of the soundfield in the HOA domain. The sound field rotation is accomplished with the "AmbiX Soundfield Rotator" plugin (https://github.com/kronihias/ambix) using a connected headtracking device.

3) The *"HOA to Binaural"* component, which performs the binaural rendering of the HOA signals. The HOA to binaural conversion is accomplished with the "Google Resonance Monitoring" plugin (https://github.com/resonance-audio/resonance-audio-daw-tools), which supports up to 3rd order HOA.

4) A *"Diegetic/Non-Diegetic content mixer"*. The non-diegetic signals are directly mixed at the headphone output.

Note that the CIBR may introduce spatial and or timbral quality changes to the rendered objects and channel based-audio signals (ESD loudspeaker inputs).

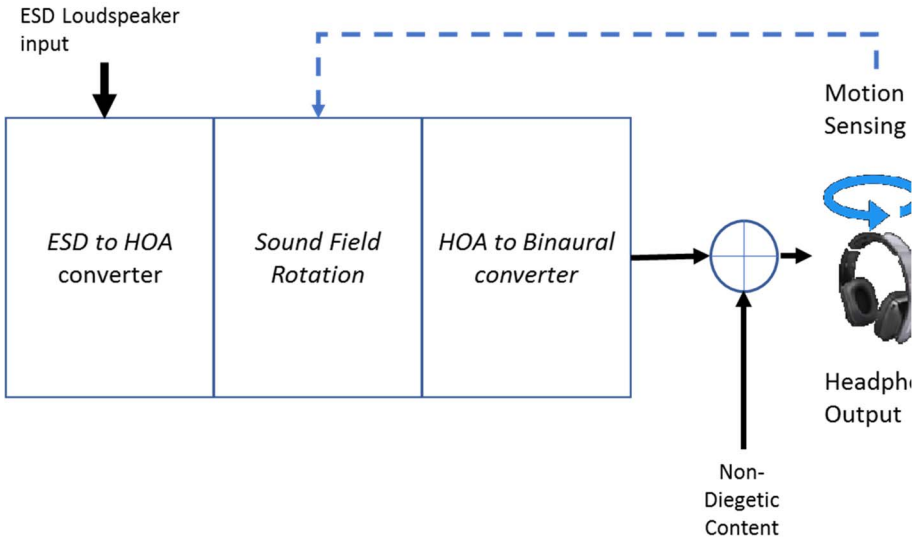

Figure 4.5-1: Block diagram of Common Informative Binaural Renderer

4.5.1.3 External Renderer

The primary purpose of the *External Renderer* is to enable alternatives to the *Reference Renderer*. There may be several *External Renderers* for a given media profile.

An *External Renderer*:

a) Supports binaural and/or loudspeaker based rendering.

b) Can be the *Reference Renderer* associated to the Audio Media Profile.

c) Does not require a standardized implementable description.

d) Exposes an *External Renderer API* (see clause 4.5.1.5) and/or the *Common Renderer API* (see clause 4.5.1.4) for connecting it to an audio decoder.

e) Supports diegetic and non-diegetic content.

f) Has a Motion to Sound latency documented according to the method defined in TS 26.260 [15].

g) Has a Loudness documented according to the method defined in TS 26.260 [15].

h) Provides a suitable subjective quality level characterized according to the *Rendering Test* (see clause 4.5.1.6) with additional comparison with the *Reference Renderer*.

i) Provides an interface to specify the set of HRTFs used for binaural rendering if applicable.

4.5.1.4 Common Renderer API

The purpose of the *Common Renderer API* is to enable the use of an *External Renderer* that can support all VRStream media profiles. The *Common Renderer API*:

a) Is normative.

b) Has a standardized implementable description in 3GPP technical specifications or by reference.

4.5.1.5 External Renderer API

The purpose of the *External Renderer API* is to enable the use of an External Renderer. The External Renderer API:

a) Has a standardized implementable description in 3GPP technical specifications or by reference.

b) Provides the necessary information to connect a VRStream media profile with an *External Renderer*.

4.5.1.6 Rendering Test

The purpose of the Rendering Test is to characterize the Quality of Experience (QoE) when using the *Reference Renderer* or *External Renderer*.

The *Rendering Test*:

a) Is defined in TS 26.259 [16] clause 6.

b) Characterizes media profile performance with *Reference Renderer* or *External Renderer*.

c) Assesses performance for multiple audio quality attributes and overall quality.

5 Video

5.1 Video Operation Points

5.1.1 Definition of Operation Point

For the purpose to define interfaces to a conforming video decoder, video operation points are defined. In this case the following definitions hold:

- **Operation Point:** A collection of discrete combinations of different content formats including spatial and temporal resolutions, colour mapping, transfer functions, VR specific rendering metadata, etc. and the encoding format.
- **Receiver:** A receiver that can decode and render any bitstream that is conforming to a certain Operation Point.
- **Bitstream:** A video bitstream that conforms to a video encoding format and certain Operation Point including VR rendering metadata.

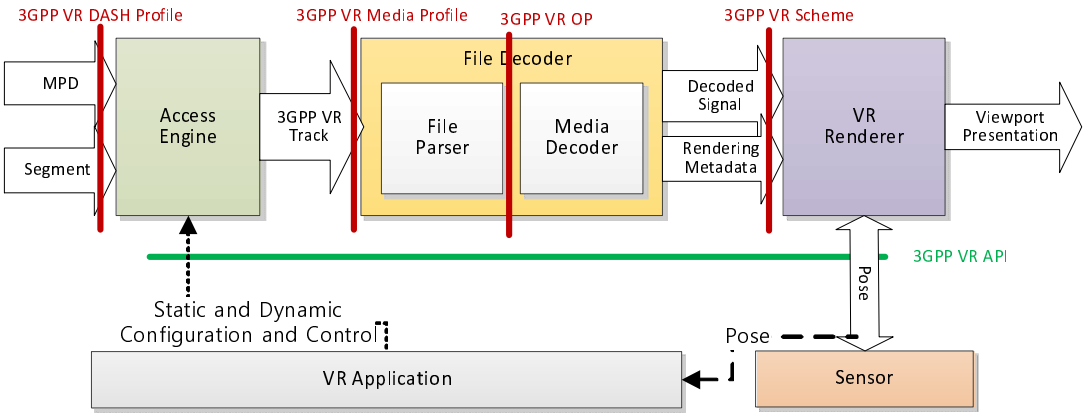

Figure 5.1-1: Video Operation Points

This clause focuses on the interoperability point to a media decoder as indicated in Figure 5.1-1. This clause does not deal with the access engine and file parser which addresses aspects how the video bitstream is delivered.

In all video operation points, the VR Presentation can be rendered using a single media decoder which provides decoded signals and rendering metadata by decoding relevant SEI messages.

5.1.2 Parameters of Visual Operation Point

This clause defines the potential parameters of Visual Operation Points. This includes the video decoder profile and levels with additional restrictions, conventional video signal parameters and VR rendering metadata. The requirements are defined from the perspective of the video decoder and renderer.

Parameters for a Visual Operation Point include:

- Codec, Profile and level requirements
- Restrictions of regular video parameters, typically expressed in the Video Usability information
- Usage and restrictions of VR rendering metadata

5.1.3 Operation Point Summary

The present document defines several operation points for different target applications and scenarios. In particular, two legacy operation points are defined that use existing video codecs H.264/AVC and H.265/HEVC to enable distribution of up to 4K full 360 mono video signals up to 60 Hz by using simple equirectangular projection.

In addition, one operation for each codec is defined that enables enhanced features, in particular stereo video, up to 8K mono, higher frame rates and HDR.

Table 5.1-1 summarizes the Operation Points, the detailed definitions are defined in the remainder of clause 5.1 where 3k refers to 2880 × 1440 pixels, 4k to 4096 × 2048 pixels, 6k to 6144 × 3072 pixels and 8k to 8192 × 4096 pixels (expressed in luminance pixel width × luminance pixel height).

Note: The Table only provides an informative high-level summary and is not considered to be complete. The specification text in the remainder of clause 5.1 refines the table and takes precedence over any information documented in the table.

Restrictions on source formats such as resolution and frame rates, content generation and encoding guidelines are provided in Annex A.

Table 5.1-1: High-level Summary of Operation Points

Operation Point name	Decoder	Bit depth	Typical Original Spatial Resolution	Frame Rate	Colour space format	Transfer Characteristics	Projection	Rotation	RWP	Stereo
Basic H.264/AVC	H.264/AVC HP@L5.1	8	Up to 4k	Up to 60 Hz	BT.709	BT.709	ERP w/o padding	No	No	No
Main H.265/HEVC	H.265/HEVC MP10@L5.1	8, 10	Up to 6k in mono and 3k in stereo	Up to 60 Hz	BT.709 BT.2020	BT.709	ERP w/o padding	No	Yes	Yes
Flexible H.265/HEVC	H.265/HEVC MP10@L5.1	8, 10	Up to 8k in mono and 3k in stereo	Up to 120 Hz	BT.709 BT.2020	BT.709, BT.2100 PQ	ERP w/o padding CMP	No	Yes	Yes

VR Rendering metadata in the Operation Points is carried in SEI messages. Receivers are expected to be able to process the VR metadata carried in SEI messages. However, the same VR metadata may be duplicated on system-level. In this case, the Receiver may rely on the system level processing to extract the relevant VR Rendering metadata rather than extracting this from the SEI message.

5.1.4 Basic H.264/AVC

5.1.4.1 General

This operation point targets simple deployments and legacy receivers at basic quality. A full 360-degree video signal with equirectangular projection following the 3GPP reference system may be provided to the decoding and rendering

system for immediate decoding and rendering. Note that this operation point enables to distribute 4k video at regular frame rates and 3k video at higher frame rates.

Restricted coverage is supported as well, but only in a basic and backward-compatible fashion.

A Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation point shall conform to the requirements in the remainder of clause 5.1.4.

A receiver conforming to the 3GPP VR Basic H.264/AVC Operation point shall support decoding and rendering a Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation point. Detailed receiver requirements are provided in the remainder of clause 5.1.4.

5.1.4.2 Profile and level

A Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation point shall conform to H.264/AVC Progressive High Profile Level 5.1 [5] for H.264/AVC with the following additional restrictions and requirements:

- the maximum VCL Bit Rate is constrained to be 120Mbps with `cpbBrVclFactor` and `cpbBrNalFactor` being fixed to be 1250 and 1500, respectively.
- the bitstream does not contain more than 10 slices per picture.

NOTE: High Profile for H.264/AVC excludes Flexible macro-block order, Arbitrary slice ordering, Redundant slices, Data partition.

Hence, for a Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation point, the following applies:

- The `profile_idc` shall be set to 100 indicating the High profile.
- The `constrain_set0_flag`, `constrain_set1_flag`, `constrain_set2_flag` and `constrain_set3_flag` shall all be set to 0.
- The value of `level_idc` shall not be greater than `51` (corresponding to the level 5.1) and should indicate the lowest level to which the Bitstream conforms.

5.1.4.3 Aspect Ratios and Spatial resolutions

Picture aspect ratio 2:1 should be used for the encoded picture.

The spatial resolution of the original format in equirectangular projection (ERP) should be one of the following (expressed in luminance pixel width × luminance pixel height):

- 4096 × 2048, 3840 × 1920, 3072 × 1536, 2880 × 1440, 2048 × 1024.

The spatial resolution of the distribution format should be one of the following (expressed in luminance pixel width × luminance pixel height):

- 3840 × 1920, 2880 × 1440, 1920 × 960, 1440 × 720, 960 × 480.
- 4096 × 2048, 3072 × 1536, 2048 × 1024, 1536 × 768, 1024 × 512.

NOTE: Distribution formats do not exceed the native resolution of the Operation Point, but they may be subsampled in order to optimize distribution or adapt to the viewing conditions.

A Receiver conforming to the 3GPP VR Basic H.264/AVC Operation Point shall be capable of decoding and rendering Bitstreams that contain spatial resolutions as above.

5.1.4.4 Colour information

A Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation Point shall use Recommendation ITU-R BT.709 [3] colorimetry. Hence, in the VUI, the colour parameter information shall be present, i.e.

- `video_signal_type_present_flag` value and `colour_description_present_flag` value shall be set to 1.

- The `colour_primaries` value, the `transfer_characteristics` value and the `matrix_coefficients` value in the Video Usability Information shall all be set to 1.

A Receiver conforming to the 3GPP VR Basic H.264/AVC Operation Point shall be capable of decoding and rendering Bitstreams that use Recommendation ITU-R BT.709 [3] colorimetry according to the bitstream requirements documented above.

5.1.4.5 Frame rates

A Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation Point shall have one of the following frame rates: 24; 25; 30; 24/1001; 30/1001; 50; 60; 60/1001 Hz.

The profile and level constraints of H.264/AVC Progressive High Profile Level 5.1 require careful balance of the permitted frame rates and spatial resolutions. Table 5.1-2 provides the permitted combinations of spatial resolutions and frame rates.

Table 5.1-2: Permitted combinations of spatial resolutions and frame rates

Spatial Resolution	Permitted Frame Rates
4096 × 2048	24; 25; 30; 24/1001; 30/1001 Hz
3840 × 1920	24; 25; 30; 24/1001; 30/1001 Hz
3072 × 1536	24; 25; 30; 24/1001; 30/1001; 50 Hz
2880 × 1440	24; 25; 30; 24/1001; 30/1001; 50; 60; 60/1001 Hz
2048 × 1024	24; 25; 30; 24/1001; 30/1001; 50; 60; 60/1001 Hz

In the VUI, the timing information may be present:

- If the timing information is present, i.e. the value of `timing_info_present_flag` is set to 1, then the values of `num_units_in_tick` and `time_scale` shall be set according to the frame rates allowed above. The timing information present in the video Bitstream should be consistent with the timing information signalled at the system level.

- The frame rate shall not change between two RAPs. `fixed_frame_rate_flag` value shall be set to 1.

A Receiver conforming to the 3GPP VR Basic H.264/AVC Operation Point shall be capable of decoding and rendering Bitstreams that use frame rates according to the bitstream requirements documented above.

5.1.4.6 Random access point

For H.264/AVC random access point (RAP) definition refer to TS 26.116 [12], clause 4.4.1.1.

RAPs shall be present in the Bitstream at least once every 5 seconds. It is recommended that RAPs occur in the video Bitstream on average at least every 2 seconds. The time interval between successive RAPs is measured as the difference between their respective decoding time values.

5.1.4.7 Sequence parameter set

The following restrictions apply to the active Sequence Parameter Set (SPS):

- `gaps_in_frame_num_value_allowed_flag` value shall be set to 0.

- The Video Usability Information shall be present in the active Sequence Parameter Set. The `vui_parameter_present_flag` shall be set to 1.

- The source video format shall be progressive. `frame_mbs_only_flag` shall be set to 1 for every picture of the Bitstream.

5.1.4.8 Video usability information

In addition to the previous constraints on the VUI on colour information in clause 5.1.4.4 and on frame rates in clause 5.1.4.5, this clause contains further requirements.

The aspect ratio information shall be present, i.e.

- The `aspect_ratio_present_flag` value shall be set to 1.
- The `aspect_ratio_idc` value shall be set to 1 indicating a square pixel format.

There are no requirements on output timing conformance for H.264/AVC decoding (Annex C of [5]). The Hypothetical Reference Decoder (HRD) parameters, if present, should be ignored by the Receiver.

5.1.4.9 Omni-directional Projection Format

This operation point uses equirectangular projection, such the video is automatically rendered in the 3GPP reference system. This is enabled by using the MPEG metadata on equirectangular projection.

A Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation Point shall include the equirectangular projection SEI message (`payloadType` equal to 150) at every RAP. The `erp_guard_band_flag` shall be set to 0.

A Receiver conforming to the 3GPP VR Basic H.264/AVC Operation Point shall be able to process the information contained on equirectangular projection SEI message (`payloadType` equal to 150) with `erp_guard_band_flag` shall be set to 0.

5.1.4.10 Restricted Coverage

This operation point permits the decoding and rendering of restricted coverage video signals in a rudimentary way. In this case it is expected that pixels that are projected to a non-covered region are included in the full image, but are visually differentiated from the covered region, for example using black, grey or white color.

Application or system-based signalling may support signalling the coverage region.

5.1.4.11 Other VR Metadata

For a Bitstream conforming to the 3GPP VR Basic H.264/AVC Operation Point:

- the equirectangular projection SEI message (`payloadType` equal to 150) with `erp_guard_band_flag` not set to 0 shall not be present,
- the sphere rotation SEI message (`payloadType` equal to 154) shall not be present,
- the region-wise packing SEI message (`payloadType` equal to 155) shall not be present,
- the frame-packing arrangement SEI message (`payloadType` equal to 45) shall not be present.

5.1.4.12 Receiver Compatibility

Receivers conforming to the 3GPP VR Basic H.264/AVC Operation Point shall support decoding and displaying 3GPP VR Basic H.264/AVC Operation Point Bitstreams.

Receivers conforming to the 3GPP VR Basic H.264/AVC Operation Point shall support all Receiver requirements in clause 5.1.4.

5.1.5 Main H.265/HEVC

5.1.5.1 General

This operation targets enhanced 360 video decoding and rendering of H.265/HEVC video for VR applications. Among others, this operation point supports among others rendering of:

- 4K mono video at up to 60 Hz frame rates
- 3K stereoscopic video at up to 60 Hz frame rates

- Higher than 4K resolutions for restricted coverage
- Rendering of certain viewports in higher quality than others beyond 4K
- extended colour space and SDR transfer characteristics

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation point shall conform to the requirements in the remainder of clause 5.1.5.

A Receiver conforming to the 3GPP VR Main H.265/HEVC Operation point shall support decoding and rendering a Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation point. Detailed receiver requirements are provided in the remainder of clause 5.1.5.

5.1.5.2 Profile and level

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation point shall conform to H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [6].

Hence, for a Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation point shall comply with the following restrictions:

- The `general_profile_idc` shall be set to 2 indicating the Main10 profile.
- The `general_tier_flag` shall be set to 0 indicating the Main tier.
- The value of `level_idc` shall not be greater than 153 (corresponding to the Level 5.1) and should indicate the lowest level to which the Bitstream conforms.

5.1.5.3 Bit depth

Bitstreams conforming to the 3GPP VR Main H.265/HEVC Operation point shall be encoded with either 8 or 10 bit precision:

- `bit_depth_luma_minus8` = 0 or 2 (8 or 10 bits respectively)
- `bit_depth_chroma_minus8` = `bit_depth_luma_minus8`

Receivers conforming to the 3GPP VR Main H.265/HEVC Operation Point shall support 8 bit and10 bit precision.

5.1.5.4 Spatial Resolutions

Due to the options provided in this operation point, additional original format may be considered that can then be decoded and rendered by a Receiver conforming to this operation point. Recommended original formats beyond those specified in clause 5.1.4.3 for equirectangular projection (ERP) are:

- Mono formats: 6144 × 3072, 5880 × 2880
- Stereo formats with resolution for each eye: 3840 × 1920, 2880 × 1440, 2048 × 1024

If original signals are beyond the maximum permitted resolution of the video codec, then the region-wise packing needs to be applied to generate suitable distribution formats.

The distribution formats are more flexible as additional VR metadata as defined in the remainder of clause 5.1.5 may be used. However, for the distribution formats, all requirements of H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [5] shall apply to the decoded texture signal.

According to H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [6], the maximum luminance width and height does not exceed 8,444 pixels. In addition to the H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [6] constraints, a Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation point, the decoded texture signal shall in addition:

- not exceed the luminance width of 8192 pixels, and
- not exceed the luminance height of 8192 pixels.

A Receiver conforming to the 3GPP VR Main H.265/HEVC Operation Point shall be capable of decoding and rendering Bitstreams with a decoded texture signal of maximum luminance width of 8192 pixels a, maximum luminance height of 8192 pixels and the overall profile/level constraints.

5.1.5.5 Colour information and Transfer Characteristics

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation Point shall use either Recommendation ITU-R BT.709 [3] colorimetry or Recommendation ITU-R BT.2020 [4] colorimetry in non-constant luminance for standard dynamic range (SDR).

Specifically, in the VUI, the colour parameter information shall be present, i.e.:

- `video_signal_type_present_flag` value and `colour_description_present_flag` value shall be set to 1.
- If BT.709 [3] is used, it shall be signalled by setting `colour_primaries` to the value 1, `transfer_characteristics` to the value 1 and `matrix_coeffs` to the value 1.
- If BT.2020 [4] and SDR is used, it shall be signalled by setting `colour_primaries` to the value 9, `transfer_characteristics` to the value 14 and `matrix_coeffs` to the value 9.

A Receiver conforming to the 3GPP VR Main H.265/HEVC Operation Point shall be capable of decoding and rendering according to any of the two above configurations.

5.1.5.6 Frame rates

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation Point shall have one of the following frame rates: 24; 25; 30; 24/1001; 30/1001; 50; 60; 60/1001 Hz.

Selected combinations of frame rates with other source parameters are provided in Annex A.2.2.2.

In the VUI, the timing information may be present:

- If the timing information is present, i.e. the value of `vui_timing_info_present_flag` is set to 1, then the values of `vui_num_units_in_tick` and `vui_time_scale` shall be set according to the frame rates allowed in this clause. The timing information present in the video Bitstream should be consistent with the timing information signalled at the system level.
- The frame rate shall not change between two RAPs. `fixed_frame_rate_flag` value, if present, shall be set to 1.

There are no requirements on output timing conformance for H.265/HEVC decoding (Annex C of [6]). The Hypothetical Reference Decoder (HRD) parameters, if present, should be ignored by the Receiver.

A Receiver conforming to the 3GPP VR Main H.265/HEVC Operation Point shall be capable of decoding and rendering Bitstreams that use frame rates according to the bitstream requirements documented above.

5.1.5.7 Random access point

For H.265/HEVC random access point (RAP) definition refer to TS 26.116 [12], clause 4.5.1.2.1.

RAPs shall be present in the Bitstream at least once every 5 seconds. It is recommended that RAPs occur in the video Bitstream on average at least every 2 seconds. The time interval between successive RAPs is measured as the difference between their respective decoding time values.

If viewport adaptation is offered, then RAPs should occur even more frequently to enable transitioning across these viewport-optimized bitstreams.

5.1.5.8 Video and Sequence Parameter Sets

Receivers conforming to the 3GPP VR Main H.265/HEVC Operation Point should ignore the content of all Video Parameter Sets (VPS) NAL units [9] as defined in Recommendation ITU-T H.265 / ISO/IEC 23008-2 [6].

The following restrictions apply to the active Sequence Parameter Set (SPS):

- The Video Usability Information (VUI) shall be present in the active Sequence Parameter Set. The `vui_parameters_present_flag` shall be set to 1.
- The chroma sub-sampling shall be 4:2:0, `chroma_format_idc` value shall be set to 1.
- The source video format shall be progressive, i.e.:
  - The `general_progressive_source_flag` shall be set to 1,
  - The `general_interlaced_source_flag` shall be set to 0,
  - The `general_frame_only_constraint_flag` shall be set to 1.

Receivers conforming to the 3GPP VR Main H.265/HEVC Operation Point shall support Bitstreams with the restrictions on the SPS defined above.

5.1.5.9 Video usability information

In addition to the previous constraints on the VUI on colour information in clauses 5.1.5.5 and 5.1.5.6, this clause contains further requirements.

The aspect ratio information shall be present, i.e.:

- The `aspect_ratio_present_flag` value shall be set to 1.
- The `aspect_ratio_idc` value shall be set to 1 indicating a square pixel format.

There are no requirements on output timing conformance for H.265/HEVC decoding (Annex C of [6]). The Hypothetical Reference Decoder (HRD) parameters, if present, should be ignored by the Receiver.

5.1.5.10 Omni-directional Projection Formats

This operation point permits using either equirectangular projection following the MPEG metadata specifications, such the video is automatically rendered in the 3GPP reference system.

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation Point shall include at every RAP the equirectangular projection SEI message (`payloadType` equal to 150) with the `erp_guard_band_flag` set to 0.

5.1.5.11 Restricted Coverage

This operation point permits to distribute content with less than 360 degree coverage in an encoding optimized manner by the use of region-wise packing.

It is recommended that the number of pixels that are projected to non-covered regions are minimized in the decoded texture signal. If this is applied and not the full 360 video is encoded, the region-wise packing SEI message (`payloadType` equal to 155) shall be included in the bitstream to signal the encoded regions of the 360 video. If present, it shall be present in a H.265/HEVC RAP.

Application or system-based signalling may support signalling the exact coverage region in the spherical coordinates.

5.1.5.12 Viewport-Optimized Content

This operation point permits the use of region-wise packing, for example to optimize the spatial resolution of specific viewports. For some example usage and settings, refer to Annex A.2.

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation Point may include the region-wise packing SEI message (`payloadType` equal to 155). If present, it shall be present in a H.265/HEVC RAP.

A Receiver conforming to the 3GPP VR Main H.265/HEVC Operation Point shall be able to process the region-wise packing SEI message (`payloadType` equal to 155).

5.1.5.13 Frame packing arrangement

A Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation Point may include the frame packing arrange SEI message (`payloadType` equal to 45). If present, then the following settings shall apply:

- The SEI message is present in a H.265/HEVC RAP.
- The value of `frame_packing_arrangement_cancel_flag` is equal to 0.
- The value of `frame_packing_arrangement_type` is equal to 4.
- The value of `quincunx_sampling_flag` is equal to 0.
- The value of `spatial_flipping_flag` is equal to 0.
- The value of `field_views_flag` is equal to 0.
- The value of `frame0_grid_position_x` is equal to 0.
- The value of `frame0_grid_position_y` is equal to 0.
- The value of `frame1_grid_position_x` is equal to 0.
- The value of `frame1_grid_position_y` is equal to 0.

A Receiver conforming to the 3GPP VR Main H.265/HEVC Operation Point shall process the frame packing arrangement SEI (`payloadType` equal to 45) with settings restrictions as above. If processing is supported, then the Receiver shall render the viewport indicated by the message.

5.1.5.14 Other VR Metadata

For a Bitstream conforming to the 3GPP VR Main H.265/HEVC Operation Point:

- the sphere rotation SEI message (`payloadType` equal to 154) shall not be present.
- any frame-packing arrangement SEI message (`payloadType` equal to 45) that does not conform to an SEI message defined in clause 5.1.5.13 shall not be present.

5.1.5.15 Receiver Compatibility

Receivers conforming to the 3GPP VR Main H.265/HEVC Operation Point shall support decoding and displaying 3GPP VR Main H.265/HEVC Operation Point Bitstreams.

Receivers conforming to the 3GPP VR Main H.265/HEVC Operation Point shall support all Receiver requirements in clause 5.1.5. Specifically, receivers conforming to the 3GPP VR Main H.265/HEVC Operation Point shall support decoding and rendering Bitstreams that include the following VR rendering metadata:

- the region-wise packing SEI message (for details see clauses 5.1.5.11 and 5.1.5.12)
- the equirectangular projection SEI message (for details see clause 5.1.5.10)
- the frame-packing arrangement SEI message (for details see clause 5.1.5.13)
- any combinations of those

5.1.6 Flexible H.265/HEVC

5.1.6.1 General

This operation targets enhanced 360 video decoding and rendering of H.265/HEVC video for VR applications. Among others, this operation point supports rendering of:

- 4K mono video at up to 120 Hz frame rates

- 3K stereoscopic video at up to 60 Hz frame rates
- Higher than 4K resolutions for restricted coverage
- Rendering of certain viewports in higher quality than others beyond 4K
- ERP and CMP projection
- SDR and HDR transfer characteristics

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation point shall conform to the requirements in the remainder of clause 5.1.6.

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation point shall support decoding and rendering a Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation point. Detailed receiver requirements are provided in the remainder of clause 5.1.6.

5.1.6.2 Profile and level

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation point shall conform to H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [6].

Hence, for a Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation point shall comply with the following restrictions:

- The `general_profile_idc` shall be set to 2 indicating the Main10 profile.
- The `general_tier_flag` shall be set to 0 indicating the Main tier.
- The value of `level_idc` shall not be greater than 153 (corresponding to the Level 5.1) and should indicate the lowest level to which the Bitstream conforms.

5.1.6.3 Bit depth

Bitstreams conforming to the 3GPP VR Flexible H.265/HEVC Operation point shall be encoded with either 8 or 10 bit precision:

- `bit_depth_luma_minus8` = 0 or 2 (8 or 10 bits respectively)
- `bit_depth_chroma_minus8` = `bit_depth_luma_minus8`

Receivers conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall support 8 bit and10 bit precision.

5.1.6.4 Spatial Resolutions

Due to the options provided in this operation point, additional original format may be considered that can then be decoded and rendered by a Receiver conforming to this operation point. Recommended original formats beyond those specified in clause 5.1.5.4 for equirectangular projection (ERP) are:

- Mono formats: 8192 × 4096

This operation point permits the distribution of ERP signals directly as well as the conversion of ERP signals to cube-map (CMP) projection. A conversion operation is provided in Annex A.2.3. Typical original cubemap format, either generated by conversion or provided by the content provider, that are suitable for this operation point are listed as follows:

- Mono Formats: 6144x4096, 4608x3072, 4320x2880, 3072x2048, 2880x1920, 2304x1536, 2160x1440
- Stereo Formats with resolution for each eye: 4320x2880, 3072x2048, 2880x1920, 2304x1536, 2160x1440

If original signals are beyond the maximum permitted resolution of the video codec, then region wise packing needs to be applied to generate suitable distribution formats.

The distribution formats are more flexible as additional VR metadata as defined in the remainder of clause 5.1.6 may be used. However, for the distribution formats, all requirements of H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [5] shall apply to the decoded texture signal.

According to H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 [6], the maximum luminance width and height does not exceed 8,444 pixels. However, for improved interoperability, for a Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation point, the decoded texture signal:

- shall not exceed the luminance width of 8192 pixels, and
- shall not exceed the luminance height of 8192 pixels.

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall be capable of decoding and rendering Bitstreams with a decoded texture signal of maximum luminance width of 8192 pixels and maximum luminance height of 8192 pixels.

5.1.6.5 Colour information and Transfer Characteristics

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall use either Recommendation ITU-R BT.709 [3] colorimetry or Recommendation ITU-R BT.2020 [4] colorimetry in non-constant luminance for standard dynamic range (SDR). For High Dynamic Range (HDR), BT.2020 [4] colorimetry in non-constant luminance and Perceptual Quantization (PQ) electro-optical transfer function (EOTF) as defined in Recommendation ITU-R BT.2100 [11] are used.

Specifically, in the VUI, the colour parameter information shall be present, i.e.:

- `video_signal_type_present_flag` value and `colour_description_present_flag` value shall be set to 1.
- If BT.709 [3] is used, it shall be signalled by setting `colour_primaries` to the value 1, `transfer_characteristics` to the value 1 and `matrix_coeffs` to the value 1.
- If BT.2020 [4] and SDR is used, it shall be signalled by setting `colour_primaries` to the value 9, `transfer_characteristics` to the value 14 and `matrix_coeffs` to the value 9.
- If BT.2020 [4] and ITU-R BT.2100 [11] are used in HDR, it shall be signalled by setting `colour_primaries` to the value 9, `transfer_characteristics` to the value 16 and `matrix_coeffs` to the value 9.

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall be capable of decoding and rendering according to any of the three above configurations.

SEI messages for HDR metadata signalling may be used. The requirements and recommendations for Bitstreams and Receivers as documented in TS 26.116 [12], clause 4.5.5.7 also apply for the 3GPP VR Flexible H.265/HEVC Operation Point.

5.1.6.6 Frame rates

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall have one of the following frame rates: 24; 25; 30; 24/1001; 30/1001; 50; 60; 60/1001, 90, 100, 120 Hz.

Selected combinations of frame rates with other source parameters are provided in Annex A.2.2.2.

In the VUI, the timing information may be present:

- If the timing information is present, i.e. the value of `vui_timing_info_present_flag` is set to 1, then the values of `vui_num_units_in_tick` and `vui_time_scale` shall be set according to the frame rates allowed in this clause. The timing information present in the video Bitstream should be consistent with the timing information signalled at the system level.
- The frame rate shall not change between two RAPs. `fixed_frame_rate_flag` value, if present, shall be set to 1.

There are no requirements on output timing conformance for H.265/HEVC decoding (Annex C of [6]). The Hypothetical Reference Decoder (HRD) parameters, if present, should be ignored by the Receiver.

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall be capable of decoding and rendering Bitstreams that use frame rates according to the bitstream requirements documented above.

5.1.6.7 Random access point

For H.265/HEVC random access point (RAP) definition refer to TS 26.116 [12], clause 4.5.1.2.1.

RAPs shall be present in the Bitstream at least once every 5 seconds. It is recommended that RAPs occur in the video Bitstream on average at least every 2 seconds. The time interval between successive RAPs is measured as the difference between their respective decoding time values.

If viewport adaptation is offered, then RAPs should occur even more frequently to enable transitioning across these viewport-optimized bitstreams.

5.1.6.8 Video and Sequence Parameter Sets

Receivers conforming to the 3GPP VR Flexible H.265/HEVC Operation Point should ignore the content of all Video Parameter Sets (VPS) NAL units [9] as defined in Recommendation ITU-T H.265 / ISO/IEC 23008-2 [6].

The following restrictions apply to the active Sequence Parameter Set (SPS):

- The Video Usability Information (VUI) shall be present in the active Sequence Parameter Set. The `vui_parameters_present_flag` shall be set to 1.
- The chroma sub-sampling shall be 4:2:0, `chroma_format_idc` value shall be set to 1.
- The source video format shall be progressive, i.e.:
  - The `general_progressive_source_flag` shall be set to 1,
  - The `general_interlaced_source_flag` shall be set to 0,
  - The `general_frame_only_constraint_flag` shall be set to 1.

Receivers conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall support Bitstreams with the restrictions on the SPS defined above.

5.1.6.9 Video usability information

In addition to the previous constraints on the VUI on colour information in clauses 5.1.6.5 and 5.1.6.6, this clause contains further requirements.

The aspect ratio information shall be present, i.e.:

- The `aspect_ratio_present_flag` value shall be set to 1.
- The `aspect_ratio_idc` value shall be set to 1 indicating a square pixel format.

There are no requirements on output timing conformance for H.265/HEVC decoding (Annex C of [6]). The Hypothetical Reference Decoder (HRD) parameters, if present, should be ignored by the Receiver.

5.1.6.10 Omni-directional Projection Formats

This operation point permits using either equirectangular projection or cubemap projection following the MPEG metadata specifications, such the video is automatically rendered in the 3GPP reference system.

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall include at every RAP either:

- the equirectangular projection SEI message (`payloadType` equal to 150) with the `erp_guard_band_flag` set to 0, or

- the cubemap projection SEI message (`payloadType` equal to 151).

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall be able to process the equirectangular projection SEI message (`payloadType` equal to 150) and the cubemap projection SEI message (`payloadType` equal to 151).

5.1.6.11 Restricted Coverage

This operation point permits to distribute content with less than 360 degree coverage in an encoding optimized manner by the use of region-wise packing.

It is recommended that the number of pixels that are projected to non-covered regions are minimized in the decoded texture signal. If this is applied and not the full 360 video is encoded, the region-wise packing SEI message (`payloadType` equal to 155) shall be included in the bitstream to signal the encoded regions of the 360 video. If present, it shall be present in a H.265/HEVC RAP.

Application or system-based signalling may support signalling the exact coverage region in the spherical coordinates.

5.1.6.12 Viewport-Optimized Content

This operation point permits the use of region-wise packing, for example to optimize the spatial resolution of specific viewports. For some example usage and settings, refer to Annex A.2.

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation Point may include the region-wise packing SEI message (`payloadType` equal to 155). If present, it shall be present in a H.265/HEVC RAP.

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall be able to process the region-wise packing SEI message (`payloadType` equal to 155).

5.1.6.13 Frame packing arrangement

A Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation Point may include the frame packing arrange SEI message (`payloadType` equal to 45). If present, then the following settings shall apply:

- The SEI message is present in a H.265/HEVC RAP.
- The value of `frame_packing_arrangement_cancel_flag` is equal to 0.
- The value of `frame_packing_arrangement_type` is equal to 4.
- The value of `quincunx_sampling_flag` is equal to 0.
- The value of `spatial_flipping_flag` is equal to 0.
- The value of `field_views_flag` is equal to 0.
- The value of `frame0_grid_position_x` is equal to 0.
- The value of `frame0_grid_position_y` is equal to 0.
- The value of `frame1_grid_position_x` is equal to 0.
- The value of `frame1_grid_position_y` is equal to 0.

A Receiver conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall process the frame packing arrangement SEI (`payloadType` equal to 45) with settings restrictions as above. If processing is supported, then the Receiver shall render the viewport indicated by the message.

5.1.6.14 Other VR Metadata

For a Bitstream conforming to the 3GPP VR Flexible H.265/HEVC Operation Point:

- the sphere rotation SEI message (`payloadType` equal to 154) shall not be present.

- any frame-packing arrangement SEI message (`payloadType` equal to 45) that does not conform to an SEI message defined in clause 5.1.6.13 shall not be present.

5.1.6.15 Receiver Compatibility

Receivers conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall support decoding and displaying 3GPP VR Main H.265/HEVC Operation Point Bitstreams and 3GPP VR Flexible H.265/HEVC Operation Point Bitstreams.

Receivers conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall support all Receiver requirements in clause 5.1.6. Specifically, receivers conforming to the 3GPP VR Flexible H.265/HEVC Operation Point shall support decoding and rendering Bitstreams that include the following VR rendering metadata:

- the region-wise packing SEI message (for details see clauses 5.1.6.11 and 5.1.6.12),
- the equirectangular projection SEI message (for details see clause 5.1.6.10),
- the cubemap projection SEI message (for details see clause 5.1.6.10),
- the frame-packing arrangement SEI message (for details see clause 5.1.6.13),
- any combinations of those.

5.2 Video Media Profiles

5.2.1 Introduction and Overview

This clause defines the media profiles for video. Media profiles include specification on the following:

- Elementary stream constraints based on the video operation points defined in clause 5.1.
- File format encapsulation constraints and signalling including capability signalling. The defines to a 3GPP VR Track as defined above.
- DASH Adaptation Set constraints and signalling including capability signalling. This defines a DASH content format profile.

Table 5.2-1 provides an overview of the Media Profiles in defined in the remainder of clause 5.3.2.

Table 5.2-1 Video Media Profiles

Media Profile	Operation Point	Sample Entry	DASH Integration
Basic Video	Basic H.264/AVC	`resv` `avc1`	Single Adaptation Set Single Representation streaming
Main Video	Main H.265/HEVC	`resv` `hvc1`	Single or Multiple independent Adaptation Sets offered Single Representation streaming
Advanced Video	Flexible H.265/HEVC	`resv` `hvc1, hvc2`	Single or Multiple dependent Adaptation Sets offered Single or Multiple representation streaming

NOTE: Advanced Video Profile Receivers are expected to playback content conforming to the Main Video Media Profile.

5.2.2 Basic Video Media Profile

5.2.2.1 Overview

The Basic Video Media Profile permits to download and stream elementary streams for VR content generated according to the H.264/AVC Basic Operation Point as defined in clause 5.1.4. This enables reuse of the `avc1` sample entry as for example also used in the TV Video Profiles in TS 26.116 [12]. It also permits to reuse streaming the VR video content in an adaptive manner by offering multiple switchable Representations in a single Adaptation Set in a DASH MPD.

For content generation guidelines for this media profile refer to Annex A.2.3.

5.2.2.2 File Format Signaling and Encapsulation

3GP VR Tracks conforming to this media profile used in the context of the specification shall conform to ISO BMFF [17] with the following further requirements:

- The bitstream included on the track shall comply to the Bitstream requirements and recommendations for the Basic H.264/AVC Operation Point as defined in clause 5.1.4.
- The sample entry type of each sample entry of the track shall be equal to `'resv'`.
- The `scheme_type` value of `SchemeTypeBox` in the `RestrictedSchemeInfoBox` shall be `'podv'`, and all instances of `CompatibleSchemeTypeBox` defined in ISO/IEC 23090-2 [13] in the same `RestrictedSchemeInfoBox` shall include at least the `scheme_type` value `'erpv'`.
- The untransformed sample entry type shall be equal to `'avc1'`.

NOTE: If a file decoder experiences issues in the playback of the VR Track with the restricted sample `'resv'`, but the application is able to control the rendering according to the VR rendering metadata, then the untransformed sample entry could be used to initialize the decoding process for the file decoder.

- The Track Header Box (`'tkhd'`) shall obey the following constraints:
  - The `width` and `height` fields for a visual track shall specify the track's visual presentation size as fixed-point 16.16 values expressed in on a uniformly sampled grid (commonly called square pixels) (of the decoded texture signal)
  - The Video Media Header (`'vmhd'`) shall obey the following constraints:
    - The value of the `version` field shall be set to '0'.
    - The value of the `graphicsmode` field shall be set to '0'.
    - The value of the `opcolor` field shall be set to {'0', '0', '0'}.
  - The Sample Description Box (`'stsd'`) obeys the following constraints:
    - A visual sample entry shall be used.
    - The box shall include a NAL Structured Video Parameter Set.
    - `width` and `height` field shall correspond to the cropped horizontal and vertical sample counts provided in the Sequence Parameter Set of the track.
    - It shall contain a Decoder Configuration Record which signals the Profile, Level, and other parameters of the video track.
    - It shall contain `AVCConfigurationBox` which signals the Profile, Level, Bit depth, and other parameters conforming to the bitstream constraints specified in clause 5.1.4.
  - The Colour Information Box (`'colr'`) should be present. If present, it shall signal the `colour_primaries`, `transfer_characteristics` and `matrix_coeffs` applicable to all the bitstreams associated with this sample entry.
  - The `ProjectionFormatBox` with `projection_type` equal to 0 as defined in ISO/IEC 23090-2 [13] should be present in the sample entry applying to the sample containing the picture.
  - It shall not contain the `RegionWisePackingBox` and `StereoVideoBox.`
  - If the content contained in the Bitstream in the track does not cover the entire sphere, the `CoverageInformationBox` as defined in ISO/IEC 23090-2 [13] should be present. If present, only a single region may be signaled and the following restrictions apply:
    - The `coverage_shape_type` shall be set to 1.

- The num_regions value shall be set to 1.

- The view_idc_presence_flag shall be set to 0.

- The default_view_idc shall be set to 0.

If 3GP VR Tracks conforming to the constraints of this media profile, the '3vrb' ISO brand should be set as a compatible_brand in the File Type Box ('ftyp').

5.2.2.3 DASH Integration

5.2.2.3.1 Definition

If all Representations in an Adaptation Set conform to the requirements in clause 5.2.2.3.2 and the Adaptation Set conforms to the requirements in clause 5.2.2.3.3, then the @profiles parameter in the Adaptation Set may signal conformance to this Operation Point by using "urn:3GPP:vrstream:mp:video:basic".

5.2.2.3.2 Additional Restrictions for DASH Representations

If a VR Track conforming to this media profile is included in a DASH Representation, the Representation use movie fragments and therefore, the following additional requirements apply:

- The Media Header Box ('mdhd') shall obey the following constraints:

  - The value of the duration field shall be set to '0'.

    - The value of the duration field in the Movie Header Box ('mvhd') shall be set to a value of '0'.

- The Sample Table Box ('stbl') shall obey the following constraints:

  - The entry_count field of the Sample-to-Chunk Box ('stsc') shall be set to '0'.

  - Both the sample_size and sample_count fields of the Sample Size Box ('stsz') box shall be set to zero ('0'). The sample_count field of the Sample Size Box ('stz2') box shall be set to zero ('0'). The actual sample size information can be found in the Track Fragment Run Box ('trun') for the track.

  NOTE: This is because the Movie Box ('moov') contains no media samples.

  - The entry_count field of the Chunk Offset Box ('stco') shall be set to '0'.

- The Track Header Box ('tkhd') shall obey the following constraints:

  - The value of the duration field shall be set to '0'.

- Movie Fragment Header Boxes ('mfhd') shall contain sequence_number values that are sequentially numbered starting with the number 1 and incrementing by +1, sequenced by movie fragment storage and presentation order.

- Any Segment Index Box ('sidx'), if present, shall obey the additional constraints:

  - The timescale field shall have the same value as the timescale field in the Media Header Box ('mdhd') within the same track; and

  - The reference_ID field shall be set to the track_ID of the ISO Media track as defined in the Track Header Box ('tkhd').

  - The Segment Index shall describe the entire file and only a single Segment Index Box shall be present.

For all Representation in an Adaptation Set, the following shall apply:

- The identical coverage information shall be present on all Representations in one Adaptation Set.

- The frame rates of all Representations in one Adaptation Set shall be identical.

5.2.2.3.3 DASH Adaptation Set Constraints

For a video Adaptation Set, the following constraints apply:

- The `@codecs` parameter shall be present on Adaptation Set level and shall signal the maximum required capability to decode any Representation in the Adaptation Set. The `@codecs` parameter should be signalled on the Representation level if different from the one on Adaptation Set level.
- The attributes `@maxWidth` and `@maxHeight` shall be present. They are expected be used to signal the original projected source content format. This means that they may exceed the actual largest size of any coded Representation in one Adaptation Set.
- The `@width` and `@height` shall be signalled for each Representation (possibly defaulted on Adaptation Set level) and shall match the values of the maximum width and height in the Sample Description box of the contained Representation.
- The Chroma Format may be signalled. If signalled:
  - An Essential or Supplemental Descriptor shall be used to signal the value by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegB:cicp:MatrixCoefficients` as defined ISO/IEC 23001-8 [10] and the `@value` attribute according to Table 4 of ISO/IEC 23001-8 [10]. The values shall match the values set in the VUI.
  - The signalling shall be on Adaptation Set level.
- The Color Primaries and Transfer Function may be signalled. If signalled:
  - An Essential or Supplemental Descriptor shall be used to signal the value by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegB:cicp:ColourPrimaries` and `urn:mpeg:mpegB:cicp:TransferCharacteristics` as defined ISO/IEC 23001-8 [10] and the `@value` attribute according to Table 4 of ISO/IEC 23001-8 [10]. The values shall match the values set in the VUI.
  - The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.
- The `@frameRate` should be signalled on Adaptation Set level.
- Random Access Points shall be signalled by `@startsWithSAP` set to `1` or `2`.
- a Supplemental Descriptor should be used to signal the projection by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegI:omaf:2017:pf` as defined ISO/IEC 23090-2 [13] and the `omaf:@projection_type` attribute set to 0.
- If the `CoverageInformationBox` is present, a Supplemental Descriptor should be used to signal the value by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegI:omaf:2017:cc` as defined ISO/IEC 23090-2 [13] and shall match the information provided in the `CoverageInformationBox`. Specifically,
  - the **`cc`**`@shape_type` shall be present and be set to 1.
  - the **`cc`**`@view_idc_presence_flag` shall not be present.
  - exactly one **`cc.CoverageInfo`** element shall be present.
    - any **`cc.CoverageInfo`** attribute that is not `centre_azimuth`, `centre_elevation`, `azimuth_range` and `elevation_range`, shall not be present.
  - The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.
- The `FramePacking` element shall not be present.
- The `@profiles` parameters may be present to signal the constraints for the Adaptation Set.

5.2.3 Main Video Media Profile

5.2.3.1 Overview

The Main Video Media Profile permits to download and stream elementary streams for VR content generated according to the H.265/HEVC Main Operation Point as defined in clause 5.1.5. This enables reuse of the hvc1 sample entry as for example also used in the TV Video Profiles in TS 26.116 [12]. It also permits to reuse streaming the VR video content in an adaptive manner by offering multiple switchable Representations in a single Adaptation Set in a DASH MPD. Furthermore, this profile enables that multiple Video Adaptation Sets are offered for the same content, each encoded for a preferred viewport. Multiple Viewpoints may be signaled, for example expressing different type of content or different camera positions.

For content generation guidelines for this media profile refer to Annex A.2.3.2.

5.2.3.2 File Format Signaling and Encapsulation

3GP VR Tracks conforming to this media profile used in the context of the specification shall conform to ISO BMFF [17] with the following further requirements:

- The included in the video track shall comply to the Bitstream requirements and recommendations for the Main.265/HEVC Operation Point as defined in clause 5.1.5 with the additional constraints
  - the region-wise packing SEI message (payloadType equal to 155). if present in any H.265/HEVC RAP, shall be present in any H.265/HEVC RAP and shall be identical for all H.265/HEVC RAP.
- The sample entry type of each sample entry of the track shall be equal to `'resv'`.
- The `scheme_type` value of `SchemeTypeBox` in the `RestrictedSchemeInfoBox` shall be `'podv'`, and all instances of `CompatibleSchemeTypeBox` defined in ISO/IEC 23090-2 [13] in the same `RestrictedSchemeInfoBox` shall include at least one of the `scheme_type` values `'erpv'` and `'ercm'`.
- The untransformed sample entry type shall be equal to `'hvc1'`.

NOTE: If a file decoder experiences issues in the playback of the VR Track with the restricted sample `'resv'`, but the application is able to control the rendering according to the VR rendering metadata, then the untransformed sample entry could be used to initialize the decoding process for the file decoder.

- The Track Header Box (`'tkhd'`) shall obey the following constraints:
  - The `width` and `height` fields for a visual track shall specify the track's visual presentation size as fixed-point 16.16 values expressed in on a uniformly sampled grid (commonly called square pixels) (of the decoded texture signal)
  - The Video Media Header (`'vmhd'`) shall obey the following constraints:
    - The value of the `version` field shall be set to '0'.
    - The value of the `graphicsmode` field shall be set to '0'.
    - The value of the `opcolor` field shall be set to {'0', '0', '0'}.
  - The Sample Description Box (`'stsd'`) obeys the following constraints:
    - A visual sample entry shall be used.
    - The box shall include at least one Sequence Parameter Set NAL unit.
    - `width` and `height` field shall correspond to the cropped horizontal and vertical sample counts provided in the Sequence Parameter Set of the track.
    - It shall contain a Decoder Configuration Record which signals the Profile, Level, and other parameters of the video track.

- The Colour Information Box ('colr') should be present. If present, it shall signal the colour_primaries, transfer_characteristics and matrix_coeffs applicable to all the bitstreams associated with this sample entry.

- The ProjectionFormatBox with projection_type equal to 0 as defined in ISO/IEC 23090-2 [13] shall be present in the sample entry applying to the sample containing the picture.

- If the content contained in the Bitstream in the track does not cover the entire sphere, the CoverageInformationBox as defined in ISO/IEC 23090-2 [13] should be present. If present, only a single region may be signaled and the following restrictions apply:

  - The coverage_shape_type shall be set to 1, i.e. the sphere region is specified by two azimuth circles and two elevation circles.

  - The num_regions value shall be set to 1.

  - The view_idc_presence_flag shall be set to 0.

  - The default_view_idc shall be set to 0 or 3.

- If the content contained in the Bitstream in the track includes the region-wise packing SEI message (payloadType equal to 155), then the RegionWisePackingBox as defined in ISO/IEC 23090-2 [17] shall be present. It shall signal the same information that is included in the region-wise packing SEI message(s) in the elementary stream.

- If the content contained in the Bitstream in the track does includes the frame packing arrangement SEI message (payloadType equal to 45) in the video stream, the StereoVideoBox shall be present in the sample entry applying to the sample containing the picture. When StereoVideoBox is present, it shall signal the frame packing format that is included in the frame packing arrangement SEI message(s) in the elementary stream.

If 3GP VR Tracks conforming to the constraints of this media profile, the '3vrm' ISO brand should be set as a compatible_brand in the File Type Box ('ftyp').

5.2.3.3 DASH Integration

5.2.3.3.1 Definition

If all Representations in an Adaptation Set conform to the requirements in clause 5.2.3.3.2 and the Adaptation Set conforms to the requirements in clause 5.2.3.3.3, then the @profiles parameter in the Adaptation Set may signal conformance to this Operation Point by using "urn:3GPP:vrstream:mp:video:main".

Clause 5.2.3.3.4 defines Adaptation Set Ensembles for viewport-optimized offering.

5.2.3.3.2 Additional Restrictions for DASH Representations

If a VR Track conforming to this media profile is included in a DASH Representation, the Representation use movie fragments and therefore, the following additional requirements apply:

- The Media Header Box ('mdhd') shall obey the following constraints:

  - The value of the duration field shall be set to '0'.

  - The value of the duration field in the Movie Header Box ('mvhd') shall be set to a value of '0'

- The Sample Table Box ('stbl') shall obey the following constraints:

  - The entry_count field of the Sample-to-Chunk Box ('stsc') shall be set to '0'.

  - Both the sample_size and sample_count fields of the Sample Size Box ('stsz') box shall be set to zero ('0'). The sample_count field of the Sample Size Box ('stz2') box shall be set to zero ('0'). The actual sample size information can be found in the Track Fragment Run Box ('trun') for the track.

NOTE: This is because the Movie Box ('moov') contains no media samples.

- The entry_count field of the Chunk Offset Box ('stco') shall be set to '0'.

- The Track Header Box ('tkhd') shall obey the following constraints:
  - The value of the duration field shall be set to '0'.

- Movie Fragment Header Boxes ('mfhd') shall contain sequence_number values that are sequentially numbered starting with the number 1 and incrementing by +1, sequenced by movie fragment storage and presentation order.

- Any Segment Index Box ('sidx'), if present, shall obey the additional constraints:
  - The timescale field shall have the same value as the timescale field in the Media Header Box ('mdhd') within the same track; and
  - The reference_ID field shall be set to the track_ID of the ISO Media track as defined in the Track Header Box ('tkhd').
  - The Segment Index shall describe the entire file and only a single Segment Index Box shall be present.

5.2.3.3.3 DASH Adaptation Set Constraints

For all Representation in an Adaptation Set, the following shall apply:

- The identical coverage information shall be present on all Representations in one Adaptation Set, both on ISO BMFF and elementary stream level.
- The frame rates of all Representations in one Adaptation Set shall be identical.
- The identical region-wise packing information shall be present all Representations in one Adaptation Set, both on ISO BMFF and elementary stream level.
- The identical stereoscopic information shall be present all Representations in one Adaptation Set, both on ISO BMFF and elementary stream level.

For an Adaptation Set, the following constraints apply:

- The @codecs parameter shall be present on Adaptation Set level and shall signal the maximum required capability to decode any Representation in the Adaptation Set. The @codecs parameter should be signalled on the Representation level if different from the one on Adaptation Set level.
- The attributes @maxWidth and @maxHeight shall be present. They are expected be used to signal the used format prior to encoding. This means that they may exceed the actual largest size of any coded Representation in one Adaptation Set.
- The @width and @height shall be signalled for each Representation (possibly defaulted on Adaptation Set level) and shall match the values of the maximum width and height in the Sample Description box of the contained Representation.
- The Chroma Format may be signalled. If signalled:
  - An Essential or Supplemental Descriptor shall be used to signal the value by setting the @schemeIdURI attribute to urn:mpeg:mpegB:cicp:MatrixCoefficients as defined ISO/IEC 23001-8 [10] and the @value attribute according to Table 4 of ISO/IEC 23001-8 [10]. The values shall match the values set in the VUI.
  - The signalling shall be on Adaptation Set level.
- The Color Primaries and Transfer Function may be signalled. If signalled:
  - An Essential or Supplemental Descriptor shall be used to signal the value by setting the @schemeIdURI attribute to urn:mpeg:mpegB:cicp:ColourPrimaries and urn:mpeg:mpegB:cicp:TransferCharacteristics as defined ISO/IEC 23001-8 [10] and the

@value attribute according to Table 4 of ISO/IEC 23001-8 [10]. The values shall match the values set in the VUI.

- The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.

- The @frameRate shall be signalled on Adaptation Set level.

- Random Access Points shall be signalled by @startsWithSAP set to 1 or 2.

- A Supplemental Descriptor should be used to signal the projection by setting the @schemeIdURI attribute to urn:mpeg:mpegI:omaf:2017:pf as defined ISO/IEC 23090-2 [13] and the omaf:@projection_type attribute set to 0.

- If the CoverageInformationBox is present then the Coverage information should be signaled on Adaptation Set. If signalled

  - a Supplemental Descriptor shall be used to signal the value by setting the @schemeIdURI attribute to urn:mpeg:mpegI:omaf:2017:cc as defined ISO/IEC 23090-2 [13] and shall match the information provided in the CoverageInformationBox. Specifically:

    - The **cc**@shape_type shall be present and be set to 1.

    - The **cc**@view_idc_presence_flag shall not be present.

    - Exactly one **cc.CoverageInfo** element shall be present.

    - Any **cc.CoverageInfo** attribute that is not centre_azimuth, centre_elevation, azimuth_range and elevation_range, shall not be present.

  - The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.

- If the StereoVideoBox is present then the stereo information should be signaled on Adaptation Set. If signalled

  - A **FramePacking** descriptor shall be used to signal the value by setting the @schemeIdURI attribute to urn:mpeg:mpegB:cicp:VideoFramePackingType as defined ISO/IEC 23008-1 [10] and the @value attribute shall be set to 4.

  - The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.

5.2.3.3.4 Adaptation Set Ensembles for Viewport-Optimized offering

5.2.3.3.4.1 Introduction

If multiple Adaptation Sets are offered for the same content in order to permit seamless switching across Representations for a different Viewports, each offered in a different Adaptation Set, then this forms an Ensemble of Adaptation Sets. Note that switching across viewports is not a DASH client functionality, but it is enabled by possible access to the pose and/or viewport information by the DASH client using the 3GPP VR API as shown in Figure 4.6.

5.2.3.3.4.2 Definition and Adaptation Set Signalling

An Ensemble is defined as by Adaptation Sets with a **Viewpoint** Descriptor for which the value of the @schemeIdURI is prefixed as urn:3GPP:vrstream:ve and the actual value is urn:3GPP:vrstream:ve:<id> with <id> an unsigned integer that is identical for all Adaptation Sets in one Ensemble. By using different ids, multiple ensembles may be defined, each defining a different content (for example different camera angles). The value of @value of the descriptor, if present, is either

- a single unsigned integer value that is different for each Adaptation Set in the Ensemble. If this is present, then the spherical region-wise quality ranking (SRQR) descriptor for which the value of the `@schemeIdURI` is prefixed as `urn:mpeg:mpegI:omaf:2017:srqr` shall be present in the each Adaptation Set, or
- a tuple of integer values, separated by a white-spaces. The semantics and order are as follows:
  - `centre_azimuth:` Specifies the azimuth of the centre point of the sphere region in units of 2^{-16} degrees relative to the 3GPP coordinate system for which this Ensemble has been optimized.
  - `centre_elevation:` Specifies the elevation of the centre point of the sphere region in units of 2^{-16} degrees relative to the 3GPP coordinate system for which this Ensemble has been optimized.

  the spherical region-wise quality ranking (SRQR) descriptor for which the value of the `@schemeIdURI` is prefixed as `urn:mpeg:mpegI:omaf:2017:srqr` may additionally be present for additional information.

If the `@value` attribute is not present, then this Adaptation Set is not optimized for any Viewport. At most one adaptation set without the `@value` not present shall be present.

One Adaptation Set of one Ensemble shall be signalled as the main content. Signaling as main content shall be done by using the Role descriptor with `@schemeIdUri`="`urn:mpeg:dash:role:2011`" and `@value="main".` If for the main Ensemble an Adaptation Set is present for which the `@value` of the Viewpoint descriptor is not present, then this should be signalled as the main Adaptation Set.

The content should be offered such that within an Ensemble, if multiple Adaptation Sets with different centre points are signalled, the one is preferred which has the minimum square distance to actual Viewport center.

5.2.3.3.4.3 Representation Constraints in an Ensemble

For all Representations in an Ensemble, the following shall apply:

- The identical coverage information shall be present on all Representations in one Ensemble, both on ISO BMFF and elementary stream level.
- The frame rates of all Representations in one Ensemble shall be identical.
- The identical stereoscopic information shall be present all Representations in one Ensemble, both on ISO BMFF and elementary stream level.

5.2.3.3.4.4 Adaptation Set Constraints in an Ensemble

For all Adaptation Sets in an Ensemble, the following shall apply:

- The `@codecs` parameter shall be identical for all Adaptation Sets in one Ensemble.
- The Chroma Format shall be identical for all Adaptation Sets in one Ensemble.
- The Color Primaries and Transfer Function shall be identical for all Adaptation Sets in one Ensemble.
- The `@frameRate` shall be identical for all Adaptation Sets in one Ensemble.
- Segments and subsegments shall be aligned, i.e. `@segmentAlignment` or `@subSegmentAlignment` shall be present and shall signal the same unsigned integer value for all Adaptation Sets in an Ensemble.
- Coverage information shall be identical for all Adaptation Sets in one Ensemble.

5.2.4 Advanced Video Media Profile

5.2.4.1 Overview

This Profile permits to download and stream elementary streams for VR content generated according to the Flexible H.265/HEVC operation point as defined in clause 5.1.6. It also allows unconstrained use of rectangular region-wise packing and monoscopic and stereoscopic spherical video up to 360 degrees are supported. With the presence of region-wise packing, the resolution or quality of the omnidirectional video could be emphasized in certain regions, e.g.,

according to the user's viewing orientation. In addition, the untransformed sample entry type 'hvc2' is allowed, making it possible to use extractors and get a conforming HEVC bitstream when tile-based streaming is used.

5.2.4.2 File Format Signaling and Encapsulation

When a track is the only track in a file, `compatible_brands` containing a brand equal to `'3vra'` in `FileTypeBox` indicates that the track conforms to this media profile. When a file contains multiple tracks, `compatible_brands` containing a brand equal to `'3vra'` in `FileTypeBox` indicates that at least one of the tracks conforms to this media profile.

- The video track shall be indicated to conform to this media profile through one or both of `FileTypeBox` and `TrackTypeBox`.
- At least one sample entry type of each sample entry of the track shall be equal to `'resv'`.
- The `scheme_type` value of `SchemeTypeBox` in the `RestrictedSchemeInfoBox` shall be `'podv'`, and of all instances of `CompatibleSchemeTypeBox` defined in ISO/IEC 23090-2 [13] in the same `RestrictedSchemeInfoBox` shall include at least one of the `scheme_type` values `'erpv'` and `'ercm'`.
- The untransformed sample entry type shall be equal to `'hvc1'` or `'hvc2'`.
- When the untransformed sample entry type is `'hvc2'`, the track shall include one or more `'scal'` track references.
- `LHEVCConfigurationBox` shall not be present in `VisualSampleEntry`.
- `HEVCConfigurationBox` in `VisualSampleEntry` shall be added such that it does not contradict to the Bitstream requirements of the Flexible H.265/HEVC operation point in clause 5.1.6.
- The `track_not_intended_for_presentation_alone` flag of the TrackHeaderBox may be used to indicate that a track is not intended to be presented alone.
- The Track Header Box (`'tkhd'`) shall obey the following constraints:
  - The `width` and `height` fields for a visual track shall specify the track's visual presentation size as fixed-point 16.16 values expressed in on a uniformly sampled grid (commonly called square pixels) (of the decoded texture signal)
- The Video Media Header (`'vmhd'`) shall obey the following constraints:
  - The value of the `version` field shall be set to '0'.
  - The value of the `graphicsmode` field shall be set to '0'.
  - The value of the `opcolor` field shall be set to {'0', '0', '0'}.
- The Sample Description Box (`'stsd'`) obeys the following constraints:
  - A visual sample entry shall be used.
  - The box shall include a NAL Structured Video Parameter Set.
  - `width` and `height` field shall correspond to the cropped horizontal and vertical sample counts provided in the Sequence Parameter Set of the track.
  - It shall contain a Decoder Configuration Record which signals the Profile, Level, and other parameters of the video track.
- The Colour Information Box (`'colr'`) should be present. If present, it shall signal the `colour_primaries`, `transfer_characteristics` and `matrix_coeffs` applicable to all the bitstreams associated with this sample entry.
- A `ProjectionFormatBox` as defined in ISO/IEC 23090-2 [13] shall be present in the sample entry with `projection_type` equal to 0 or 1.

- If the content contained in the Bitstream in the track does not cover the entire sphere, the CoverageInformationBox as defined in ISO/IEC 23090-2 [13] should be present.
- If the video content contained in the Bitstream in the track is a subset of the entire video content carried in the file and the CoverageInformationBox as defined in ISO/IEC 23090-2 [13] is present, the following restrictions apply:
  - If the equirectangular projection is used then,
    - The `coverage_shape_type` shall be set to 1, i.e. the sphere region is specified by two azimuth circles and two elevation circles.
    - The `num_regions` value shall be set to 1.
  - If the cubemap projection is used, then one of the two following options applies:
    a) The `coverage_shape_type` shall be set to 1, i.e. the sphere region is specified by two azimuth circles and two elevation circles and the `num_regions` value shall be set to 1, or
    b) The `coverage_shape_type` shall be set to 0, i.e. the sphere region is specified by four great circles.
  - The `view_idc_presence_flag` shall be set to 0.
  - The `default_view_idc` shall be set to 0 or 3.
- If the content contained in the Bitstream in the track includes the region-wise packing SEI message (`payloadType` equal to 155), then the `RegionWisePackingBox` as defined in ISO/IEC 23090-2 [17] shall be present. It shall signal the same information that is included in the region-wise packing SEI message(s) in the elementary stream.
- If the content contained in the Bitstream in the track includes the frame packing arrangement SEI message (`payloadType` equal to 45) in the video stream, the `StereoVideoBox` shall be present in the sample entry applying to the sample containing the picture. When `StereoVideoBox` is present, it shall signal the frame packing format that is included in the frame packing arrangement SEI message(s) in the elementary stream.

5.2.4.3 DASH Integration

5.2.4.3.1 Definition

If all Representations in an Adaptation Set conform to the requirements in clause 5.2.4.3.2 and the Adaptation Set conforms to the requirements in clause 5.2.4.3.3, then the `@profiles` parameter in the Adaptation Set may signal conformance to this Operation Point by using "`urn:3GPP:vrstream:mp:video:advanced`".

5.2.4.3.2 Additional Restrictions for DASH Representations

If a VR Track conforming to this media profile is included in a DASH Representation, the Representation use movie fragments and therefore, the following additional requirements apply:

- The value of the duration field in the Media Header Box (`'mdhd'`) shall be set to a value of '0'.
- The value of the duration field in the Movie Header Box (`'mvhd'`) shall be set to a value of '0'.
- The value of the duration field in the Track Header Box (`'tkhd'`) shall be set to a value of '0'.
- Movie Fragment Header Boxes (`'mfhd'`) may contain sequence_number values that are not sequentially numbered.
- Any Segment Index Box (`'sidx'`), if present, shall obey the additional constraints:
  - the timescale field shall have the same value as the timescale field in the Media Header Box ('mdhd') within the same track;

- the reference_ID field shall be set to the track_ID of the ISO Media track as defined in the Track Header Box (`'tkhd'`).

- The Sample Table Box (`'stbl'`) shall obey the following constraints:

  - The `entry_count` field of the Sample-to-Chunk Box (`'stsc'`) shall be set to '0'.

  - Both the `sample_size` and `sample_count` fields of the Sample Size Box ('stsz') box shall be set to zero ('0'). The `sample_count` field of the Sample Size Box (`'stz2'`) box shall be set to zero ('0'). The actual sample size information can be found in the Track Fragment Run Box ('trun') for the track.

NOTE 1: This is because the Movie Box (`'moov'`) contains no media samples.

- The `entry_count` field of the Chunk Offset Box (`'stco'`) shall be set to '0'.

- The same projection format shall be used on all Representations in one Adaptation Set.

- The same frame packing format shall be used on all Representations in one Adaptation Set.

- The same coverage information shall be used on all Representations in one Adaptation Set.

- The same spatial resolution shall be used on all Representations in one Adaptation Set.

- When `@dependencyId` is used, the values of profiles of the respective dependent and complementary Representations shall be the same.

When the MPD contains a Representation with a track for which the untransformed sample entry type is equal to `'hvc2'`, the following applies:

- Either the Representations carrying a track conforming to the media profile track constraints with the untransformed sample entry type equal to `'hvc2'` shall contain `@dependencyId` listing all dependent Representations that carry a track conforming to the media profile track constraints with the untransformed sample entry type equal to `'hvc1'` or a Preselection property descriptor shall be present and constrained as follows:

  - The Main Adaptation Set shall contain a Representation carrying a track conforming to the media profile track constraints with the untransformed sample entry type equal to `'hvc2'`.

  - The Partial Adaptation Sets shall contain Representations each carrying a track conforming to the media profile track constraints with the untransformed sample entry type equal to `'hvc1'`.

NOTE 2: When using the Preselection property descriptor, the number of Representations for carrying tracks with the untransformed sample entry type equal to `'hvc2'` is typically smaller than when using `@dependencyId`. However, the use of `@dependencyId` might be needed for encrypted video tracks.

- The Initialization Segment of the Representation that contains `@dependencyId` or belongs to the Main Adaptation Set is constrained as follows:

  - Tracks conform to the media profile track constraints.

  - The track corresponding to the untransformed sample entry type equal to `'hvc2'` refers to the tracks indicated in the TrackReferenceBox of the Initialization Segment.

NOTE 3: When Preselection is used, the `sequence_number` integer values are not required to be processed and therefore the concatenation of the Subsegments (of the different Representations of the Adaptation Sets of a Preselection) in any order results in a conforming file.

NOTE 4: The conforming Segment sequence formed on the basis of the Preselection property descriptor or by resolving `@dependencyId` attribute(s) as specified in ISO/IEC 23009-1 [18] and the `track_ID` value of the track with the untransformed sample entry type equal to `'hvc2'` produces the HEVC bitstream which conforms to H.265/HEVC Flexible Operation Point.

When switching or accessing Representations at each segment or subsegment is relevant, the following DASH profiles include sufficient constraints:

- ISO Base Media File Format Live profile: `urn:mpeg:dash:profile:isoff-live:2011`
- ISO Base Media File Format Main profile: `urn:mpeg:dash:profile:isoff-main:2011`

When low latency considerations are relevant, the following DASH profiles provide tools to support efficient low latency services:

- ISO Base Media File Format On Demand profile: `urn:mpeg:dash:profile:isoff-on-demand:2011`
- ISO Base Media File Format Broadcast TV profile: `urn:mpeg:dash:profile:isoff-broadcast:2015`

5.2.4.3.3 DASH Adaptation Set Constraints

For all Representation in an Adaptation Set, the following shall apply:

- The identical coverage information shall be present on all Representations in one Adaptation Set, both on ISO BMFF and elementary stream level.
- The frame rates of all Representations in one Adaptation Set shall be identical.
- The identical region-wise packing information shall be present all Representations in one Adaptation Set, both on ISO BMFF and elementary stream level.
- The identical stereoscopic information shall be present all Representations in one Adaptation Set, both on ISO BMFF and elementary stream level.

For an Adaptation Set, the following constraints apply:

- The `@codecs` parameter shall be present on Adaptation Set level and shall signal the maximum required capability to decode any Representation in the Adaptation Set. The @codecs parameter should be signalled on the Representation level if different from the one on Adaptation Set level.
- The attributes `@maxWidth` and `@maxHeight` shall be present. They are expected be used to signal the decoded texture format of the original signal. This means that they may exceed the actual largest size of any coded Representation in one Adaptation Set.
- The `@width` and `@height` shall be signalled for each Representation (possibly defaulted on Adaptation Set level) and shall match the values of the maximum width and height in the Sample Description box of the contained Representation.
- The Chroma Format may be signalled. If signalled:
  - An Essential or Supplemental Descriptor shall be used to signal the value by setting the `@schemeIdURI` attribute to urn:mpeg:mpegB:cicp:MatrixCoefficients as defined ISO/IEC 23001-8 [10] and the `@value` attribute according to Table 4 of ISO/IEC 23001-8 [10]. The values shall match the values set in the VUI.
  - The signalling shall be on Adaptation Set level.
- The Color Primaries and Transfer Function may be signalled. If signalled:
  - An Essential or Supplemental Descriptor shall be used to signal the value by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegB:cicp:ColourPrimaries` and urn:mpeg:mpegB:cicp:TransferCharacteristics as defined ISO/IEC 23001-8 [10] and the `@value` attribute according to Table 4 of ISO/IEC 23001-8 [10]. The values shall match the values set in the VUI.

  The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.
- The `@frameRate` shall be signalled on Adaptation Set level.
- Random Access Points shall be signalled by `@startsWithSAP` set to 1 or 2.

- An Essential Descriptor shall be used to signal the projection by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegI:omaf:2017:pf` as defined ISO/IEC 23090-2 [13] and the `omaf:@projection_type` attribute set to 0 or 1.
- If the `CoverageInformationBox` is present, then the Coverage information should be signaled on Adaptation Set. If signalled:
  - A Supplemental Descriptor shall be used to signal the value by setting the `@schemeIdURI` attribute to `urn:mpeg:mpegI:omaf:2017:cc` as defined ISO/IEC 23090-2 [13] and shall match the information provided in the CoverageInformationBox. Specifically,
    - the **`cc`**`@shape_type` shall be present and be set to 0 or 1.
    - the **`cc`**`@view_idc_presence_flag` shall not be present.
    - exactly one **`cc.CoverageInfo`** element shall be present.
    - any **`cc.CoverageInfo`** attribute that is not `centre_azimuth, centre_elevation, azimuth_range` and `elevation_range`, shall not be present.
  - The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.
- If the `StereoVideoBox` is present, then the stereo information should be signaled on Adaptation Set. If signalled:
  - a FramePacking descriptor shall be used to signal the value by setting the @schemeIdURI attribute to `urn:mpeg:mpegB:cicp:VideoFramePackingType` as defined ISO/IEC 23008-1 [10] and the `@value` attribute shall be set to 4.
  - The signalling shall be on Adaptation Set level only, i.e. the value shall not be different for different Representations in one Adaptation Set.
- The following applies for the use of `@mimeType`:
  - `@mimeType` of the Main Adaptation Set shall include the profiles parameter `'3vra'`.
  - When Preselection is used, the value of profiles of the main Adaptation Set shall be the same as the value of profiles of its partial Adaptation Sets.
- When Preselection is used, the following applies:
  - The value of `@subsegmentAlignment` in the Main Adaptation Set shall be an unsigned integer and equal to the value of `@subsegmentAlignment` of the each associated Partial Adaptation Set.
  - The value of `@segmentAlignment` in the Main Adaptation Set shall be an unsigned integer and equal to the value of `@segmentAlignment` of the each associated Partial Adaptation Set.

5.2.4.3.4 Adaptation Set Constraints for Viewport Selection

If multiple Adaptation Sets are offered for the same content which have emphasized quality regions for different viewports, in order to provide signaling information for switching across Viewports, the spherical region-wise quality ranking (SRQR) descriptor for which the value of the `@schemeIdURI` is prefixed as `urn:mpeg:mpegI:omaf:2017:srqr` shall be present in the each Adaptation Set with following restrictions:

- The **`sphRegionQuality`**`@view_idc_presence_flag` shall be set to 0.
- The **`sphRegionQuality`**`@default_view_idc` shall be set to 0 or 3.
- The value of **`sphRegionQualityll.qualityInfo`**`@quality_ranking` shall be greater than 0.

For all Representations in multiple Adaptation Sets for switching accross Viewports, the following shall apply:

- The identical coverage information shall be present on all Representations, both on ISO BMFF and elementary stream level.
- The frame rates of all Representations in Adaptation Sets shall be identical.
- The identical stereoscopic information shall be present all Representations, both on ISO BMFF and elementary stream level.

For all Adaptation Sets with SRQR descriptors for switching across Viewports, the following shall apply:

- The `@codecs` parameter shall be identical for all Adaptation Sets.
- The Chroma Format shall be identical for all Adaptation Sets.
- The Color Primaries and Transfer Function shall be identical for all Adaptation Sets.
- The `@frameRate` shall be identical for all Adaptation Sets.
- Segments and subsegments shall be aligned, i.e. `@segmentAlignment` or `@subSegmentAlignment` shall be present and shall signal the same unsigned integer value for all Adaptation Sets.
- Coverage information shall be identical for all Adaptation Sets.

6 Audio

6.1 Audio Operation Points

6.1.1 Definition of Operation Point

For the purpose to define interfaces to a conforming audio decoder, audio operation points are defined. In this case the following definitions hold:

- **Operation Point:** A collection of discrete combinations of different content formats and VR specific rendering metadata, etc. and the encoding format.
- **Receiver:** A receiver that can decode and render any bitstream that is conforming to a certain Operation Point.
- **Bitstream:** A audio bitstream that conforms to an audio format.

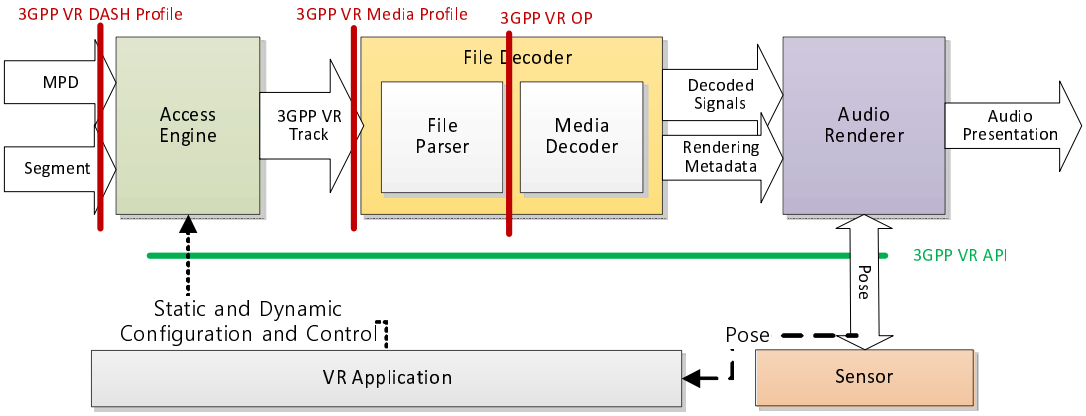

Figure 6.1-1: Audio Operation Points

This clause focuses on the interoperability point to a media decoder as indicated in Figure 6.1-1. This clause does not deal with the access engine and file parser which addresses aspects how the audio bitstream is delivered.

In all audio operation points, the VR Presentation can be rendered using a single media decoder which provides decoded PCM signals and rendering metadata to the audio renderer.

6.1.2 Parameters of Audio Operation Point

This clause defines the potential parameters of Audio Operation Points. This includes the detailed audio decoder requirements and audio rendering metadata. The requirements are defined from the perspective of the audio decoder and renderer.

Parameters for an Audio Operation Point include:

- the audio decoder that the bitstream needs to conform to,
- the mandated or permitted rendering data that is included in the audio bitstream.

6.1.3 Summary of Audio Operation Points

Table 6.1-1 provides an informative overview of the Audio Operating Points. The detailed, normative specification for each audio operating point is subsequently provided in the referenced clause.

Table 6.1-1: Overview of OMAF operation points for audio (informative)

Operation Point	Codec	Configuration		Max Sampling Rate	Clause
3GPP MPEG-H Audio	MPEG-H Audio	Low Complexity Profile, Level 1,2 or 3		48 kHz	6.1.4

6.1.4 3GPP MPEG-H Audio Operation Point

6.1.4.1 Overview

The 3GPP MPEG-H Audio Operation Point fulfills the requirements to support 3D audio and is specified in ISO/IEC 23090-2 [13], clause 10.2.2. Channels, Objects and First/Higher-Order Ambisonics (FOA/HOA) are supported, as well as combinations of those. The Operation Point is based on MPEG-H 3D Audio [19].

A bitstream conforming to the 3GPP MPEG-H Audio Operation Point shall conform to the requirements in clause 6.1.4.2.

A receiver conforming to the 3GPP MPEG-H Audio Operation Point shall support decoding and rendering a Bitstream conforming to the 3GPP MPEG-H Audio Operation Point. Detailed receiver requirements are provided in clause 6.1.4.3.

6.1.4.2 Bitstream requirements

The audio stream shall comply with the MPEG-H 3D Audio Low Complexity (LC) Profile, Levels 1, 2 or 3 as defined in ISO/IEC 23008-3, clause 4.8 [19]. The values of the `mpegh3daProfileLevelIndication` for LC Profile Levels 1, 2 and 3 are `"0x0B"`, `"0x0C"` and `"0x0D"`, respectively, as specified in ISO/IEC 23008-3 [19], clause 5.3.2.

Audio encapsulation shall be done according to ISO/IEC 23090-2 [12], clause 10.2.2.2.

All Low Complexity Profile and Levels restrictions specified in ISO/IEC 23008-3 [19], clause 4.8.2 shall apply. The constraints on input and output configurations are provided in *Table 3 — "Levels and their corresponding restrictions for the Low Complexity Profile"*, of ISO/IEC 23008-3 [19]. This includes the following for Low Complexity Profile Level 3:

- Maximum number of core coded channels (in compressed data stream): 32,
- Maximum number of decoder processed core channels: 16,

- Maximum number of loudspeaker output channels: 12
- Maximum number of decoded objects: 16
- Maximum HOA order: 6

MPEG-H Audio sync samples contain Immediate Playout Frames (IPFs), as specified in ISO/IEC 23008-3, clause 20.2 [19] and shall follow the requirements specified in ISO/IEC 23090-2 [12], clause 10.2.2.3.1.

6.1.4.3 Receiver requirements

6.1.4.3.1 General

A receiver supporting the 3GPP MPEG-H Audio Operation Point shall fulfill all requirements specified in the remainder of clause 6.1.4.3.

6.1.4.3.2 Decoding process

The receiver shall be capable of decoding MPEG-H Audio LC Profile Level 1, Level 2 and Level 3 bitstreams as specified in ISO/IEC 23008-3, clause 4.8 [19] with the following relaxations:

- The Immersive Renderer defined in ISO/IEC 23008-3 [19], clause 11 is optional.
- The carriage of generic data defined in ISO/IEC 23008-3 [19], clause 14.7 is optional and thus MHAS packets of the type PACTYP_GENDATA are optional and the decoder may ignore packets of this type.

The decoder shall read and process MHAS packets of the following types in accordance with ISO/IEC 23008-3 [19], clause 14:

PACTYP_SYNC,
PACTYP_MPEGH3DACFG,
PACTYP_AUDIOSCENEINFO,
PACTYP_AUDIOTRUNCATION,
PACTYP_MPEGH3DAFRAME,
PACTYP_USERINTERACTION,
PACTYP_LOUDNESS_DRC,
PACTYP_EARCON,
PACTYP_PCMCONFIG, and
PACTYP_PCMDATA.

The decoder may read and process MHAS packets of the following types:

PACTYP_SYNCGAP,
PACTYP_BUFFERINFO,
PACTYP_MARKER and
PACTYP_DESCRIPTOR.

Other MHAS packets may be present in an MHAS elementary stream, but may be ignored.

The Earcon metadata shall be processed and applied as described in ISO/IEC 23008-3 [19], clause 28.

6.1.4.3.3 Random Access

The audio decoder is able to start decoding a new audio stream at every random access point (RAP). As defined in clause 6.1.4.2, the sync sample (RAP) contains the configuration information (PACTYP_MPEGH3DACFG and PACTYP_AUDIOSCENEINFO) that is used to initialize the audio decoder. After initialization, the audio decoder reads encoded audio frames (PACTYP_MPEGH3DAFRAME) and decodes them.

To optimize startup delay at random access, the information from the MHAS PACTYP_BUFFERINFO packet should be taken into account. The input buffer should be filled at least to the state indicated in the MHAS PACTYP_BUFFERINFO packet before starting to decode audio frames.

NOTE 1: It may be necessary to feed several audio frames into the decoder before the first decoded PCM output buffer is available, as described in ISO/IEC 23008-3 [19], clause 5.5.6.3 and clause 22.

It is recommended that, at random access into an audio stream, the receiving device performs a 100ms fade-in on the first PCM output buffer that it receives from the audio decoder.

NOTE 2: The MPEG-H 3D Audio Codec can output the original input samples without any inherent fade-in behavior. Thus, the receiving device needs to appropriately handle potential signal discontinuities, resulting from the original input signal, by fading in at random access into an audio stream.

6.1.4.3.4 Configuration change

If the decoder receives an MHAS stream that contains a configuration change, the decoder shall perform a configuration change according to ISO/IEC 23008-3 [19], clause 5.5.6. The configuration change can, for instance, be detected through the change of the `MHASPacketLabel` of the packet `PACTYP_MPEGH3DACFG` compared to the value of the `MHASPacketLabel` of previous MHAS packets.

If MHAS packets of type `PACTYP_AUDIOTRUNCATION` are present, they shall be used as described in ISO/IEC 23008-3 [19], clause 14.

The Access Unit that contains the configuration change and the last Access Unit before the configuration change may contain a truncation message (`PACTYP_AUDIOTRUNCATION`) as defined in ISO/IEC 23008-3 [19], clause 14. The MHAS packet of type `PACTYP_AUDIOTRUNCATION` enables synchronization between video and audio elementary streams at program boundaries. When used, sample-accurate splicing and reconfiguration of the audio stream are possible.

6.1.4.3.5 MPEG-H Multi-stream Audio

The receiver shall be capable of simultaneously receiving at least 3 MHAS streams. The MHAS streams can be simultaneously decoded or combined into a single stream prior to the decoder, by utilizing the field `mae_bsMetaDataElementIDoffset` in the Audio Scene Information as described in ISO/IEC 23008-3 [19], clause 14.6.

6.1.4.3.6 Rendering requirements

6.1.4.3.6.1 General

The 3GPP MPEG-H Audio Operation Point builds on the MPEG-H 3D Audio codec, which includes rendering to loudspeakers, binaural rendering and also provides an interface for external rendering. Legacy binaural rendering using fixed loudspeaker setups can be supported by using loudspeaker feeds as output of the decoder.

6.1.4.3.6.2 Rendering to Loudspeakers

Rendering to loudspeakers shall be done according to ISO/IEC 23008-3 [19] using the interface for local loudspeaker setup and rendering as defined in ISO/IEC 23008-3 [19], clause 17.3.

NOTE: ISO/IEC 23008-3 [19] specifies rendering to predefined loudspeaker setups as well as rendering to arbitrary setups.

6.1.4.3.6.3 Binaural Rendering of MPEG-H 3D Audio

6.1.4.3.6.3.1 General

MPEG-H 3D Audio specifies methods for binauralizing the presentation of immersive content for playback via headphones, as is needed for omnidirectional media presentations. MPEG-H 3D Audio specifies a normative interface for the user's viewing orientation and permits low-complexity, low-latency rendering of the audio scene to any user orientation.

The binaural rendering of MPEG-H 3D Audio shall be applied as described in ISO/IEC 23008-3 [19], clause 13 according to the Low Complexity Profile and Levels restrictions for binaural rendering specified in ISO/IEC 23008-3 [19], clause 4.8.2.2.

6.1.4.3.6.3.2 Head Tracking Interface

For binaural rendering using head tracking the `useTrackingMode` flag in the `BinauralRendering()` syntax element shall be set to 1, as described in ISO/IEC 23008-3 [19], clause 17.4. This flag defines if a tracker device is connected and the binaural rendering shall be processed in a special headtracking mode, using the scene displacement values (yaw, pitch and roll).

The values for the scene displacement data shall be sent using the interface for scene displacement data specified in ISO/IEC 23008-3 [19], clause 17.9. The syntax of `mpegh3daSceneDisplacementData()` interface provided in ISO/IEC 23008-3 [19], clause 17.9.3 shall be used.

6.1.4.3.6.3.3 Signaling and processing of diegetic and non-diegetic audio

The metadata flag `fixedPosition` in `SignalGroupInformation()` indicates if the corresponding audio signals are updated during the processing of scene-displacement angles. In case the flag is equal to one, the positions of the corresponding audio signals are not updated during the processing of scene displacement angles.

Channel groups for which the flag `gca_directHeadphone` is set to "1" in the `mpegh3da_getChannelMetadata()sytax element` are routed to left and right output channel directly and are excluded from binaural rendering using scene displacement data (non-diegetic content). Non-diegetic content may have stereo or mono format. For mono, the signal is mixed to left and right headphone channel with a gain factor of 0.707.

6.1.4.3.6.3.4 HRIR/BRIR Interface processing

The interface for binaural room impulse responses (BRIRs) specified in ISO/IEC 23008-3 [19], clause 17.4 shall be used for external BRIRs and HRIRs. The HRIR/BRIR data for the binaural rendering can be fed to the decoder by using the syntax element `BinauralRendering()`. The number of BRIR/HRIR pairs in each BRIR/HRIR set shall correspond to the number indicated in the relevant level-dependent row in Table 9 - *"The binaural restrictions for the LC profile"* of ISO/IEC 23008-3 [19] according to the Low Complexity Profile and Levels restrictions in ISO/IEC 23008-3 [19], clause 4.8.2.2.

The measured BRIR positions are passed to the `mpegh3daLocalSetupInformation()`, as specified in ISO/IEC 23008-3 [19], clause 4.8.2.2. Thus, all renderer stages are set to the target layout that is equal to the transmitted channel configuration. As one BRIR is available per regular input channel, the Format Converter can be passed through in case regular input channel positions are used. Preferably, the BRIR measurement positions for standard target layouts 2.0, 5.1, 10.2 and 7.1.4 should be provided.

6.1.4.3.6.4 Rendering with External Binaural Renderer

MPEG-H 3DA provides the output interfaces for the delivery of un-rendered channels, objects, and HOA content and associated metadata as specified in clause 6.1.4.3.6.5. External binaural renderers can connect to this interface e.g. for playback of head-tracked audio via headphones. An example of such external binaural renderer that connects to the external rendering interface of MPEG-H 3DA is specified in Annex B.

6.1.4.3.6.5 External Renderer Interface

ISO/IEC 23008-3 [19], clause 17.10 specifies the output interfaces for the delivery of un-rendered channels, objects, and HOA content and associated metadata. For connecting to external renderers, a receiver shall implement the interfaces for object output, channel output and HOA output as specified in ISO/IEC 23008-3 [19], clause 17.10, including the additional specification of production metadata defined in ISO/IEC 23008-3 [19], clause 27. Any external renderer should apply the metadata provided in this interface and related audio data in the same manner as if MPEG-H internal rendering is applied:

- Correct handling of loudness-related metadata in particular with the aim of preserving intended target loudness
- Preserving artistic intent, such as applying transmitted Downmix and HOA Rendering matrices correctly
- Rendering spatial attributes of objects appropriately (position, spatial extent, etc.)

NOTE: The external example binaural renderer in Annex B only handles a subset of the parameters to illustrate the use of the output interface. Alternative external binaural renderers are expected to apply and handle the metadata provided in this interface and related audio data in the same manner as if internal rendering is applied.

In this interface the PCM data of the channels and objects interfaces is provided through the decoder PCM buffer, which first contains the regular rendered PCM signals (e.g. 12 signals for a 7.1+4 setup). Subsequently $n_{chan,out}$ additional signals carry the PCM data of the originally transmitted channel representation. These are followed by $n_{obj,ou}$ signals carrying the PCM data of the un-rendered output objects. Then additional signals carry the $n_{HOA,out}$ HOA audio PCM data which number is indicated in the HOA metadata interface via the HOA order (e.g. 16 signals for HOA order 3). The HOA audio PCM data in the HOA output interface is provided in the so-called Equivalent Spatial Domain (ESD) representation. The conversion from the HOA domain into the ESD representation and vice versa is described in ISO/IEC 23008-3 [19], Annex C.5.1.

The metadata for channels, objects, and HOA is available once per frame and their syntax is specified in `mpegh3da_getChannelMetadata()`, `mpegh3da_getObjectAudioAndMetadata()`, and `mpegh3da_getHoaMetadata()` respectively. The metadata and PCM data shall be aligned for an external renderer to match each metadata element with the respective PCM frame.

6.2 Audio Media Profiles

6.2.1 Introduction and Overview

This clause defines the media profiles for audio. Media profiles include specification on the following:

- Elementary stream constraints based on the audio operation points defined in clause 6.1.
- File format encapsulation constraints and signalling including capability signalling. The defines to a 3GPP VR Track as defined above.
- DASH Adaptation Set constraints and signalling including capability signalling. This defines a DASH content format profile.

Table 6.2-1 provides an overview of the Media Profiles in defined in the remainder of clause 6.2.

Table 6.2-1 Audio Media Profiles

Media Profile	Operation Point	Sample Entry	DASH Integration
OMAF 3D Audio Baseline Media Profile	3GPP MPEG-H Audio Operation Point	`mhm1` `mhm2`	

6.2.2 OMAF 3D Audio Baseline Media Profile

6.2.2.1 Overview

MPEG-H 3D Audio [19] specifies coding of immersive audio material and the storage of the coded representation in an ISO BMFF track. The MPEG-H 3D Audio decoder has a constant latency, see *Table 1 — "MPEG-H 3DA functional blocks and internal processing domain"*, of ISO/IEC 23008-3 [19]. With this information, content authors could synchronize audio and video portions of a media presentation, e.g. ensuring lip-synch.

ISO BMFF integration for this profile is provided following the requirements and recommendations in ISO/IEC 23090-2 [12], clause 10.2.2.3.

6.2.2.2 File Format Signaling and Encapsulation

6.2.2.2.1 General

3GP VR Tracks conforming to this media profile used in the context of the specification shall conform to the ISO BMFF [17] with the following further requirements:

- The audio track shall comply to the Bitstream requirements and recommendations for the Operation Point as defined in clause 6.1.4.
- The sample entry 'mhm1' shall be used for encapsulation of MHAS packets into ISO BMFF files, per ISO/IEC 23008-3 [19], clause 20.6.
- All ISO Base Media File Format constraints specified in ISO/IEC 23090-2 [12], clause 10.2.2.3 shall apply.
- ISO BMFF Tracks shall be encoded following the requirements in ISO/IEC 23090-2 [12], clause 10.2.2.3.1.

6.2.2.2.2 Configuration change constraints

A configuration change takes place in an audio stream when the content setup or the Audio Scene Information changes (e.g., when changes occur in the channel layout, the number of objects etc.), and therefore new PACTYP_MPEGH3DACFG and PACTYP_AUDIOSCENEINFO packets are required upon such occurrences. A configuration change usually happens at program boundaries, but it may also occur within a program.

Configuration change constraints specified in ISO/IEC 23090-2 [12], clause 10.2.2.3.2 shall apply.

6.2.2.3 Multi-stream constraints

The multi-stream-enabled MPEG-H Audio System is capable of handling Audio Programme Components delivered in several different elementary streams (e.g., the main MHAS stream containing one complete audio main, and one or more auxiliary MHAS streams, containing different languages and audio descriptions). The MPEG-H Audio Metadata information (MAE) allows the MPEG-H Audio Decoder to correctly decode several MHAS streams.

The sample entry 'mhm2' shall be used in cases of multi-stream delivery, i.e., the MPEG-H Audio Scene is split into two or more streams for delivery as described in ISO/IEC 23008-3 [19], clause 14.6. All constraints for file formats using the sample entry 'mhm2' specified in ISO/IEC 23090-2 [12], clause 10.2.2.3.3 shall apply.

6.2.2.3a Additional Restrictions for DASH Representations

DASH Integration is provided following the requirements and recommendations in ISO/IEC 23090-2 [12], clause B.2.1. All constraints in ISO/IEC 23090-2 [12], clause B.2.1 shall apply.

6.2.2.4 DASH Adaptation Set Constraints

6.2.2.4.1 General

An instantiation of an OMAF 3D Audio Baseline Profile in DASH should be represented as one Adaptation Set. If so the Adaptation Set should provide the following signalling according to ISO/IEC 23090-2 [12] and ISO/IEC 23008-3 [19], clause 21 as shown in Table 6.2-2.

Table 6.2-2: MPEG-H Audio MIME parameter according to RFC 6381 and ISO/IEC 23008-3

Codec	MIME type	codecs parameter	profiles	ISO BMFF Encapsulation
MPEG-H Audio LC Profile Level 1	audio/mp4	mhm1.0x0B	'oabl'	ISO/IEC 23008-3
MPEG-H Audio LC Profile Level 2	audio/mp4	mhm1.0x0C	'oabl'	ISO/IEC 23008-3
MPEG-H Audio LC Profile Level 3	audio/mp4	mhm1.0x0D	'oabl'	ISO/IEC 23008-3
MPEG-H Audio LC Profile Level 1, multi-stream	audio/mp4	mhm2.0x0B	'oabl'	ISO/IEC 23008-3
MPEG-H Audio LC Profile Level 2, multi-stream	audio/mp4	mhm2.0x0C	'oabl'	ISO/IEC 23008-3
MPEG-H Audio LC Profile Level 3, multi-stream	audio/mp4	mhm2.0x0D	'oabl'	ISO/IEC 23008-3

Mapping of relevant MPD elements and attributes to MPEG-H Audio as well as the Preselection Element and Preselection descriptor are specified in ISO/IEC 23090-2 [12], clause B.2.1.2.

6.2.2.4.2 DASH Adaptive Bitrate Switching

MPEG-H 3D Audio enables seamless bitrate switching in a DASH environment with different Representations (i.e., bit streams encoded at different bitrates) of the same content, i.e., those Representations are part of the same Adaptation Set.

If the decoder receives a DASH Segment of another Representation of the same Adaptation Set, the decoder shall perform an adaptive switch according to ISO/IEC 23008-3 [19], clause 5.5.6.

7 Metadata

7.1 Presentation without Pose Information to 2D Screens

In several devices, the VR sensor providing pose information may not be available or may be disabled by the user and the presentation of the VR360 presentation is on a 2D screen. In this case, the receiver needs to rely on other information to determine a proper rendering of the VR360 presentation that is to be presented at a specific media time.

For this purpose, a VR Media Presentation may include the recommended viewport metadata. The recommended viewport metadata may be encapsulated in a timed metadata track in either a file or a DASH representation.

Receivers presenting on a 2D screen and not implementing a viewport sensor should implement the recommended viewport processing to process the recommended viewport metadata and to render VR video or audio accordingly.

Receivers implementing a viewport sensor may implement the recommended viewport processing.

If the viewport sensor is not implemented at the receiver, or if the viewport sensor is disabled (permanently or temporarily), the receiver should process the recommended viewport metadata, if present.

If the viewport metadata is provided in the VR Media Presentation and if processing is supported and applied, then the Receiver shall render the viewport indicated metadata.

8 VR Presentation

8.1 Definition

A VR presentation provides an omnidirectional audio-visual experience. A 3GPP VR Presentation is a VR Presentation for which each of the VR Tracks contained in a VR Presentation are aligned to the 3GPP DOF Reference System as defined in clause 4 and are time-synchronized.

8.2 3GPP VR File

A 3GPP VR Presentation may be provided in an ISO BMFF conforming file. A 3GPP VR File is defined as a file that conforms to ISO BMFF [17] and for which at least two tracks are present whereby:

- at least one track conforms to a 3GPP VR Track according to a video media profile defined in clause 5,
- at least one track conforms to a 3GPP VR Track according to an audio media profile defined in clause 6.

Conformance to a 3GPP VR File may be signalled with a compatibility brand `'3gvr'`.

8.3 3GPP VR DASH Media Presentation

A 3GPP VR Presentation may be provided in DASH Media Presentation. A 3GPP VR DASH Media Presentation is defined as a DASH Media Presentation that conforms to 3GP DASH and for which at least two Adaptation Sets are present whereby

- at least one Adaptation Set conforms to an Adaptation Set for a video media profile defined in clause 5,
- at least one Adaptation Set conforms to an Adaptation Set for an audio media profile defined in clause 6.

Conformance to a 3GPP VR File may be signalled with an MPD `@profiles` parameter`'urn:3gpp:vrstream:presentation'`.

9 VR Metrics

9.1 General

VR metrics is a functionality where the client collects specific quality-related metrics during a session. These collected metrics can then be reported back to a network side node for further analysis. The metric functionality is based on the QoE metrics concept in 3GP-DASH [8], but further extended to also cover VR-specific metrics. A VR client supporting VR metrics shall support all metrics listed in clause 9.3, and shall handle metric configuration and reporting as specified in clauses 9.4 and 9.5.

9.2 VR Client Reference Architecture

9.2.1 Architecture

The client reference architecture for VR metrics, shown below in Figure 9.2.1-1, is based on the client architecture in Figure 4.3-1. It also contains a number of observation points where specific metric-related information can be made available to the Metrics Collection and Computation (MCC) function. The MCC can use and combine information from the different observation points to calculate more complex metrics.

Note that these observation points are only defined conceptually, and might not always directly interface to the MCC. For instance, an implementation might relay information from the actual observation points to the MCC via the VR application. It is also possible that the MCC is not separately implemented, but simply included as an integral part of the VR application.

Also note that in this version of this specification not all of the described observation points are necessarily used to produce VR metrics.

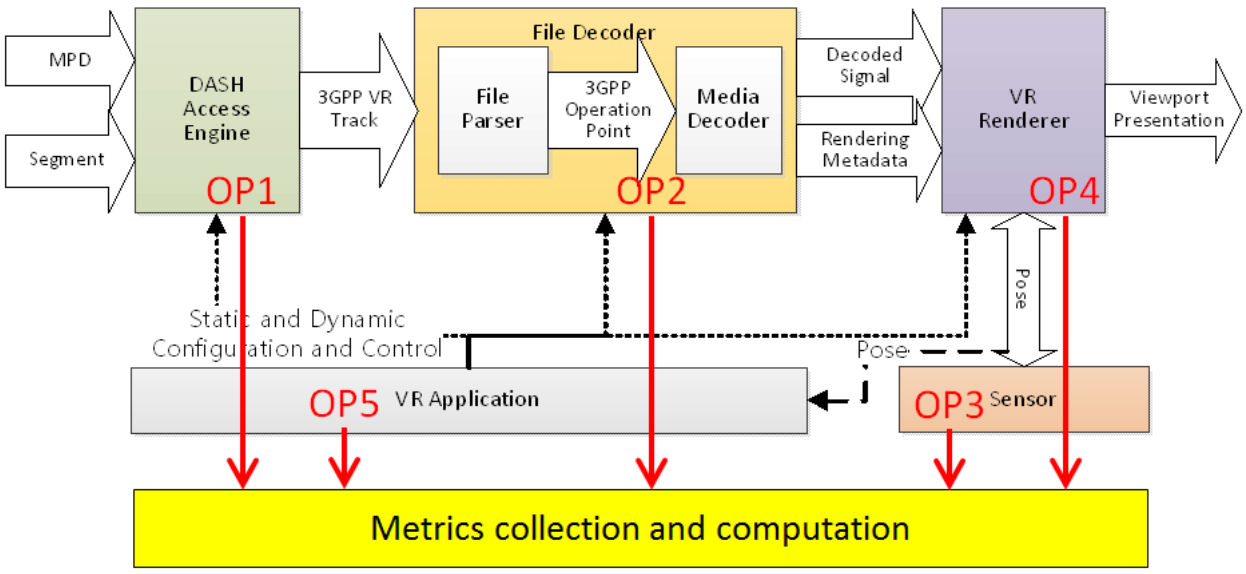

Figure 9.2.1-1: Client reference architecture for VR metrics

9.2.1 Observation Point 1

The access engine fetches the MPD, constructs and issues segment requests for relevant adaptation sets or preselections as ordered by the VR application, and receives segments or parts of segments. It may also adapt between different representations due to changes in available bitrate. The access engine provides a conforming 3GPP VR track to the file decoder.

The interface from the access engine towards MCC is referred to as observation point 1 (OP1) and is defined to monitor:

- A sequence of transmitted network requests, each defined by its transmission time, contents, and the TCP connection on which it is sent
- For each network response, the reception time and contents of the response header and the reception time of each byte of the response body
- The projection/orientation metadata carried in network manifest file if applicable
- The reception time and intended playout time for each received segment

9.2.2 Observation Point 2

The file decoder processes the 3GPP VR Track and typically includes a file parser and a media decoder. The file parser processes the file or segments, extracts elementary streams, and parses the metadata, if present. The processing may be supported by dynamic information provided by the VR application, for example which tracks to choose based on static and dynamic configurations. The media decoder decodes media streams of the selected tracks into the decoded signals. The file decoder outputs the decoded signals and metadata which is used for rendering.

The interface from the file decoder towards MCC is referred to as observation point 2 (OP2) and is defined to monitor:

- Media resolution
- Media codec
- Media frame rate
- Media projection, such as region wise packing, region wise quality ranking, content coverage
- Mono vs. stereo 360 video
- Media decoding time

9.2.3 Observation Point 3

The sensor extracts the current pose according to the user's head and/or eye movement and provides it to the renderer for viewport generation. The current pose may also be used by the VR application to control the access engine on which adaptation sets or preselections to fetch.

The interface from the sensor towards MCC is referred to as observation point 3 (OP3) and is defined to monitor:

- Head pose
- Gaze direction
- Pose timestamp
- Depth

9.2.4 Observation Point 4

The VR Renderer uses the decoded signals and rendering metadata, together with the pose and the knowledge of the horizontal/vertical field of view, to determine a viewport and render the appropriate part of the video and audio signals.

The interface from the media presentation towards MCC is referred to as observation point 4 (OP4) and is defined to monitor:

- The media type
- The media sample presentation timestamp
- Wall clock counter
- Actual presentation viewport
- Actual presentation time
- Actual playout frame rate
- Audio-to-video synchronization
- Video-to-motion latency
- Audio-to-motion latency

9.2.5 Observation Point 5

The VR application manages the complete device, and controls the access engine, the file decoder and the rendering based on media control information, the dynamic user pose, and the display and device capabilities.

The interface from the VR application towards MCC is referred to as observation point 5 (OP5) and is defined to monitor:

- Display resolution
- Max display refresh rate
- Field of view, horizontal and vertical
- Eye to screen distance
- Lens separation distance
- OS support, e.g. OS type, OS version

9.3 Metrics Definitions

9.3.1 General

As the VR metrics functionality is based on the DASH QoE metrics [8], all metrics already defined in [8] are valid also for a VR client. Thus the following sub-clauses only define additional VR-related metrics.

9.3.2 Comparable quality viewport switching latency

The comparable quality viewport switching latency metric reports the latency and the quality-related factors when viewport movement causes quality degradations, such as when low-quality background content is briefly shown before the normal higher-quality is restored. Note that this metric is only relevant if the Advanced Video Media profile and region-wise packing is used. Also note that the metric currently does not report factors related to foveated rendering.

The viewport quality is represented by two factors; the quality ranking (QR) value, and the pixel resolution of one or more regions within the viewport. The resolution is defined by the orig_width and orig_height values in ISO/IEC 23090-2 [13] in SRQR (Spherical-Region Quality Ranking) or 2DQR (2-Dimensional Quality Ranking). The resolution corresponds to the monoscopic projected picture from which the packed region covering the viewport is extracted.

In order to determine whether two viewports have a comparable quality, if more than one quality ranking region is visible inside the viewport, the aggregated viewport quality factors are calculated as the area-weighted average for QR and the area-weighted (effective) pixel resolution, respectively.

For instance, if 60% of the viewport is from a region with QR=1, Res=3840 x 2160, and 40% is from a region with QR=2, Res=960 x 540, then the average QR is 0.6 x 1 + 0.4 x 2, and the effective pixel resolution is 0.6 x 3840 x 2160 + 0.4 x 960 x 540 (also see Annex D.1 for more examples).

If the viewport is moved so that the current viewport includes at least one new quality ranking region (i.e. a quality ranking region not included in the previous viewport), a switch event is started. The list of quality factors related to the last evaluated viewport quality before the switch are assigned to the `firstViewport` log entry. The start time of the switch is also set to the time of the last evaluated viewport before the switch.

The end time for the switch is defined as when both the weighted average QR and the effective resolution for the viewport reach values comparable to the ones before the switch. A value is comparable if it is not more than QRT% (QR threshold) or ERT% (effective resolution threshold) worse than the corresponding values before the switch. If comparable values are not achieved within N milliseconds, a timeout occurs (for instance if an adaptation to a lower bitrate occurs, and the viewport never reaches comparable quality).

Note that smaller QR values and larger resolution values are better. For instance, QRT=5% would require a weighted average QR value equal or smaller than 105% of the weighted average QR before the switch, but ERT=5% would require an effective resolution value equal or larger than 95% of the effective resolution before the switch.

The list of quality factors related to the viewport which fulfills both thresholds are assigned to the `secondViewport` log entry, and the latency (end time minus start time) is assigned to the `latency` log entry. In case of a timeout, this is indicated under the `cause` log entry.

During the duration of the switch the worst evaluated viewport is also stored, and assigned to the `worstViewport` log entry. The worst viewport is defined as the viewport with the worst relative weighted average QR or relative effective resolution, as compared to the values before the switch.

If a new viewport switching event occurs (e.g. yet another new region becomes visible) before an ongoing switch event has ended, only the N milliseconds timeout is reset. The ongoing measurement process continues to evaluate the viewport quality until a comparable viewport quality value is achieved (or a timeout occurs).

The observation points needed to calculate the metrics are:

- OP2 File Decoder: SRQR/2DQR information
- OP3 Sensor: Gaze information
- OP4 VR Renderer: Start of switch event detection (alternatively, region coverage information from SRQR/2DQR can be used as strict rendering pixel-exactness is not required)

- OP5 VR Application: Field-of-view information of the device

The accuracy of the measured latency depends on how the client implements the viewport switching monitoring. As this might differ between clients, the client shall report the estimated `accuracy`.

The thresholds QRT, ERT, and the timeout N, can be specified during metrics configuration (see clause 9.4) as attributes within parenthesis, e.g. "CompQualLatency(QRT=3.5,ERT=6.8,N=900)". If a threshold or the timout is not specified the client shall use appropriate default values.

The data type `ViewportDataType` is defined in Table 9.3.2-1 below, and identifies the direction and coverage of the viewport.

Table 9.3.2-1: ViewportDataType

Key		Type	Description
`ViewportDataType`		`Object`	
	`centre_azimuth`	`Integer`	Specifies the azimuth of the centre of the viewport in units of 2^{-16} degrees. The value shall be in the range of $-180 * 2^{16}$ to $180 * 2^{16} - 1$, inclusive.
	`centre_elevation`	`Integer`	Specifies the elevation of the centre of the viewport in units of 2^{-16} degrees. The value shall be in the range of $-90 * 2^{16}$ to $90 * 2^{16}$, inclusive.
	`centre_tilt`	`Integer`	Specifies the tilt angle of the viewport in units of 2^{-16} degrees. The value shall be in the range of $-180 * 2^{16}$ to $180 * 2^{16} - 1$, inclusive.
	`azimuth_range`	`Integer`	Specifies the azimuth range of the viewport through the centre point of the viewport, in units of 2^{-16} degrees.
	`elevation_range`	`Integer`	Specifies the elevation range of the viewport through the centre point of the viewport, in units of 2^{-16} degrees.

The data type `Viewport-Item` is defined as shown in Table 2. `Viewport-Item` is an `Object` which identifies a viewport and quality-related factors for the region(s) covered by the viewport.

Table 9.3.2-2: ViewportItem

Key				Type	Description
`ViewportItem`				`Object`	
	`Position`			`ViewportDataType`	Identifies the viewport
	`QualityLevels`			List	List of different quality levels regions within the viewport
		`Coverage`		`Float`	Percentage of the viewport area covered by this region
		`QR`		`Integer`	Quality ranking (QR) value of this region
		`Resolution`		`Object`	Resolution for this region
			`Width`	`Integer`	Horizontal resolution for this region
			`Height`	`Integer`	Vertical resolution for this region

The comparable quality viewport switching latency metric is specified in Table 9.3.2-1 below.

Table 9.3.2-1: Comparable quality viewport switching latency metric

Key	Type	Description
CompQualLatency	List	List of comparable quality viewport switching latencies
Entry	Object	
firstViewport	ViewportItem	Specifies information about the first viewport
secondViewport	ViewportItem	Specifies information about the second viewport
worstViewport	ViewportItem	Specifies information about the worst viewport seen during the switch duration
time	Real-Time	Wall-clock time when the switch started
Mtime	Media-Time	Media presentation time when the switch started.
Latency	Integer	Specifies the switching delay in milliseconds.
Accuracy	Integer	Specifies the estimated accuracy of the latency metric in milliseconds
Cause	List	Specifies a list of possible causes for the latency.
Entry	Object	
code	Enum	A possible cause for the latency. The value is equal to one of the following: - 0: Segment duration - 1: Buffer fullness - 2: Availability of comparable quality segment - 3: Timeout

9.3.3 Rendered viewports

The rendered viewports metric reports a list of viewports that have been rendered during the media presentation.

The client shall evaluate the current viewport gaze every X ms and potentially add the viewport to the rendered viewport list. To enable frequent viewport evaluations without necessarily increasing the report size too much, consecutive viewports which are close to each other may be grouped into clusters, where only the average cluster viewport data is reported. Also, clusters which have too short durations may be excluded from the report.

The viewport clustering is controlled by an angular distance threshold D. If the center (i.e. the azimuth and the elevation) of the current viewport is closer than the distance D to the current cluster center (i.e. the average cluster azimuth and elevation), the viewport is added to the cluster. Note that the distance is only compared towards the current (i.e. last) cluster, not to any earlier clusters which might have been created.

If the distance to the cluster center is instead equal to or larger than D, a new cluster is started based on the current viewport, and the average old cluster data and the start time and duration for the old cluster is added to the viewport list.

Before reporting a viewport list, a filtering based on viewport duration shall be done. Each entry in the viewport list is first assigned an "aggregated duration" equal to the duration of that entry. Then, for each entry E, the other entries in the viewport list are checked. The duration for a checked entry is added to the aggregated duration for entry E, if the checked entry is both less than T ms away from E, and closer than the angular distance D from E.

After all viewport entries have been evaluated and have received a final aggregated duration, all viewport entries with an aggregated duration of less than T are deleted from the viewport list (and thus not reported). Note that the aggregated duration is only used for filtering purposes, and not itself included in the viewport list reports.

Some examples of metric calculation are shown in Annex D.2.

The observation points needed to calculate the metrics are:

- OP3 Sensor: Gaze information

- OP5 P5 VR Application: Field-of-view information of the device

The viewport sample interval X (in ms), the distance threshold D (in degrees), and the duration threshold T (in ms) can be specified during metrics configuration as attributes within parenthesis, e.g. "RenderedViewports(X=50,D=15,T=1500)". Note that if no clustering or duration filtering is wanted, the D and T thresholds can be set to 0 (e.g. specifying "RenderedViewports(X=1000,D=0,T=0)" will just log the viewport every 1000 ms). If no sample interval or thresholds values are specified the client shall use appropriate default values.

The rendered viewports metric is specified in Table 9.3.3-1.

Table 9.3.3-1 Rendered viewports metric

Key			Type	Description
`RenderedViewports`			`List`	List of rendered viewports
	`Entry`		`Object`	
		`startTime`	`Media-Time`	Specifies the media presentation time of the first played out media sample when the viewport cluster indicated in the current entry is rendered starting from this media sample.
		`duration`	`Integer`	The time duration, in units of milliseconds, of the continuously presented media samples when the viewport cluster indicated in the current entry is rendered starting from the media sample indicated by `startTime`. "Continuously presented" means that the media clock continued to advance at the playout speed throughout the interval.
		`viewport`	`ViewportDataType`	Indicates the average region of the omnidirectional media corresponding to the viewport cluster being rendered starting from the media sample time indicated by `startTime`.

9.3.4 VR Device information

This metric contains information about the device, and is logged at the start of each session and whenever changed (for instance if the rendered field-of-view for the device is adjusted). If an individual metric cannot be logged, its value shall be set to 0 (zero) or to the empty string.

The observation point needed to report the metrics is:

- OP5 VR Application: Device Information

Table 9.3.4-1: Device information

Key	Type	Description
VrDeviceInformation	List	A list of device information objects.
Entry	Object	A single object containing new device information.
start	Real-Time	Wall-clock time when the device information was logged.
mstart	Media-Time	The presentation time at which the device information was logged.
deviceIdentifier	String	The brand, model and version of the device.
horizontalResolution	Integer	The horizontal display resolution, per eye, in pixels.
verticalResolution	Integer	The vertical display resolution, per eye, in pixels.
horizontalFoV	Integer	Maximum horizontal field-of-view, per eye, in degrees.
verticalFoV	Integer	Maximum vertical field-of-view, per eye, in degrees.
renderedHorizontalFoV	Integer	Current rendered horizontal field-of-view, per eye, in degrees.
renderedVerticalFoV	Integer	Current rendered vertical field-of-view, per eye, in degrees.
refreshRate	Integer	Display refresh rate, in Hz

9.4 Metrics Configuration and Reporting

9.4.1 Configuration

Metrics configuration is done according to clauses 10.4 and 10.5 in DASH [8], but can also include any metrics defined in clause 9.3.

9.4.2 Reporting

Metrics reporting is done according to clause 10.6 in DASH [8], with the type QoeReportType extended to handle the additional VR-specific metrics according the the XML schema in clause 9.4.3. In this version of the specification the element *vrMetricSchemaVersion* shall be set to 1.

9.4.3 Reporting Format

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    targetNamespace="urn:3gpp:metadata:2020:VR:metrics"
    xmlns:hsd="urn:3gpp:metadata:2011:HSD:receptionreport"
    xmlns="urn:3gpp:metadata:2020:VR:metrics" elementFormDefault="qualified">

    <xs:complexType name="VrQoeReportType">
        <xs:complexContent>
            <xs:extension base="QoeReportType">
                <xs:sequence>
                    <xs:element name="vrMetric" type="VrMetricType"
                        minOccurs="0" maxOccurs="unbounded"/>
                    <xs:element name="vrMetricSchemaVersion" type="unsignedInt"/>
                    <xs:any namespace="##other" processContents="lax"
                        minOccurs="0" maxOccurs="unbounded"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
        <xs:anyAttribute processContents="skip"/>
    </xs:complexType>

    <xs:complexType name="VrMetricType">
        <xs:choice maxOccurs="unbounded">
            <xs:element name="compQualLatency" type="CompQualLatencyType"
                maxOccurs="unbounded"/>
            <xs:element name="renderedViewports" type="RenderedViewportsType"
                maxOccurs="unbounded"/>
            <xs:element name="vrDeviceInformation" type="VrDeviceInformationType"
                maxOccurs="unbounded"/>
            <xs:any namespace="##other" processContents="lax"
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:choice>
```

```
        <xs:anyAttribute processContents="skip"/>
    </xs:complexType>

    <xs:complexType name="CompQualLatencyType">
        <xs:sequence>
            <xs:element name="firstViewport" type="ViewportItem"/>
            <xs:element name="secondViewport" type="ViewportItem"/>
            <xs:element name="worstViewport" type="ViewportItem"/>
            <xs:element name="time" type="xs:dateTime"/>
            <xs:element name="mtime" type="xs:duration"/>
            <xs:element name="latency" type="xs:unsignedInt"/>
            <xs:element name="accuracy" type="xs:unsignedInt"/>
            <xs:element name="cause" type="unsignedInt" minoccurs="0" maxoccurs="unbounded"/>
            <xs:any namespace="##other" processContents="lax"
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:anyAttribute processContents="skip"/>
    </xs:cmplexType>

    <xs:complexType name="RenderedViewportsType">
        <xs:sequence>
            <xs:element name="startTime" type="xs:duration"/>
            <xs:element name="duration" type="xs:unsignedInt"/>
            <xs:element name="viewport" type="ViewportDataType"/>
            <xs:any namespace="##other" processContents="lax"
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:anyAttribute processContents="skip"/>
    </xs:cmplexType>

    <xs:complexType name="VrDeviceInformationType">
        <xs:sequence>
            <xs:element name="start" type="xs:dateTime"/>
            <xs:element name="mstart" type="xs:duration"/>
            <xs:element name="deviceIdentifier" type=cs:string/>
            <xs:element name="horizontalResolution" type=cs:unsignedInt/>
            <xs:element name="verticalResolution" type=cs:unsignedInt/>
            <xs:element name="horizontalFoV" type=cs:unsignedInt/>
            <xs:element name="verticalFoV" type=cs:unsignedInt/>
            <xs:element name="renderedHorizontalFoV" type=cs:unsignedInt/>
            <xs:element name="renderedVerticalFoV" type=cs:unsignedInt/>
            <xs:element name="refreshRate" type=cs:unsignedInt/>
            <xs:any namespace="##other" processContents="lax"
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:anyAttribute processContents="skip"/>
    </xs:cmplexType>

    <xs:complexType name="ViewportItem">
        <xs:sequence>
            <xs:element name="position" type="ViewportDataType"/>
            <xs:element name="qualityLevel" type="QualityLevelEntry" maxOccurs="unbounded"/>
            <xs:any namespace="##other" processContents="lax"
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:anyAttribute processContents="skip"/>
    </xs:complexType>

    <xs:complexType name="ViewportDataType">
        <xs:sequence>
            <xs:element name="centreAzimuth" type="xs:unsignedInt"/>
            <xs:element name="centreElevation" type="xs:unsignedInt"/>
            <xs:element name="centreTilt" type="xs:unsignedInt"/>
            <xs:element name="azimuthRange" type="xs:unsignedInt"/>
            <xs:element name="elevationRange" type="xs:unsignedInt"/>
            <xs:any namespace="##other" processContents="lax"
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:anyAttribute processContents="skip"/>
    </xs:cmplexType>

    <xs:complexType name="QualityLevelEntry">
        <xs:sequence>
            <xs:element name="coverage" type="xs:double"/>
            <xs:element name="qr" type="xs:unsignedInt"/>
            <xs:element name="width" type="xs:unsignedInt"/>
            <xs:element name="height" type="xs:unsignedInt"/>
            <xs:any namespace="##other" processContents="lax"
```

```
                minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:anyAttribute processContents="skip"/>
    </xs:complexType>

</xs:schema>
```

Annex A (informative): Content Generation Guidelines

A.1 Introduction

This clause collects information that supports the generation of VR Content following the details in the present document. Video and audio related aspects are collected. For additional details and background also refer to TR 26.918 [2].

A.2 Video

A.2.1 Overview

This clause collects information on that support the generation of video bitstreams that conform to operation points and media profile in the present document.

A.2.2 Decoded Texture Signal Constraints

A.2.2.1 General

Due to the restrictions to use a single decoder, the decoded texture signals require to follow the profile and level constraints of the decoder. Generally, this requires a careful balance of the permitted frame rates, stereo modes, spatial resolutions, and usage of region wise packing for different resolutions and coverage restrictions. Details on preferred settings such as frame rates and spatial resolutions are for example discussed in TR 26.918 [2].

This clause provides a summary of restrictions for the different operation points defined in the present document.

A.2.2.2 Constraints for Main and Flexible H.265/HEVC Operation Point

The profile and level constraints of H.265/HEVC Main-10 Profile Main Tier Profile Level 5.1 require careful balance of the permitted frame rates, stereo modes, spatial resolutions, and usage of region wise packing for different resolutions and coverage restrictions. If the decoded texture signal is beyond the Profile and Level constraints, then a careful adaptation of the signal is recommended to fulfil the constraints.

This clause provides a brief overview of potential signal constraints and possible adjustments.

Table A.2-1 provides selected permitted combinations of spatial resolutions, frame rates and stereo modes assuming full coverage and no region-wise packing applied. Note that fractional frame rates are excluded for better readability. Note that the Main H.265/HEVC Operation Point only allows frame rates up to 60 Hz.

Table A.2-1 Selected permitted combinations of spatial resolutions and frame rates

Spatial resolution per eye	Stereo	Permitted Frame Rates in Hz
4096 × 2048	None	24; 25; 30; 50; 60
3840 × 1920	None	24; 25; 30; 50; 60
3072 × 1536	None	24; 25; 30; 50; 60; 90; 100
2880 × 1440	None	24; 25; 30; 50; 60; 90; 100; 120
2048 × 1024	None	24; 25; 30; 50; 60; 90; 100; 120
2880 × 1440	TaB	24; 25; 30; 50; 60
2048 × 1024	TaB	24; 25; 30; 50; 60; 90; 100; 120

Table A.2-2 provides the maximum percentage of high-resolution area that can be encoded assuming that the low-resolution area is encoded in 2k resolution covering the full 360 degree area, i.e. using 2048 × 1024 or 1920 ×

960 and full coverage is provided for different frame rates. Note also that a viewport typically covers about 12-25% of a full 360 video. Note that fractional frame rates are excluded for better readability.

Table A.2-2 Maximum Percentage of high-resolution area when assuming that the low-resolution area is encoded in 2k resolution, i.e. using 2048 × 1024 or 1920 × 960 and full coverage is provided for different frame rates

Spatial resolution per eye in VP	Spatial resolution per eye in non-VP	Stereo	Frame Rates in Hz							
			24	25	30	50	60	90	100	120
6144 × 3072	2048 × 1024	None	9.29%	9.29%	9.29%	9.29%	9.29%	5.24%	4.43%	3.21%
4096 × 2048	2048 × 1024	None	21.67%	21.67%	21.67%	21.67%	21.67%	12.22%	10.33%	7.50%
3840 × 1920	1920 × 960	None	100.00%	100.00%	100.00%	100.00%	100.00%	74.12%	63.38%	47.26%
6144 × 3072	2048 × 1024	TaB	3.21%	3.21%	3.21%	3.21%	3.21%	1.19%	0.79%	0.18%
4096 × 2048	2048 × 1024	TaB	7.50%	7.50%	7.50%	7.50%	7.50%	2.78%	1.83%	0.42%
3840 × 1920	1920 × 960	TaB	47.26%	47.26%	47.26%	47.26%	47.26%	20.40%	15.02%	6.96%

Table A.2-3 provides the maximum percentage of coverage area that can be encoded assuming that the remaining pixels are not encoded for different frame rates. Note that fractional frame rates are excluded for better readability.

Table A.2-3 Maximum Percentage of coverage area for different frame rates

Spatial resolution per eye	Stereo	Frame Rates in Hz							
		24	25	30	50	60	90	100	120
6144 × 3072	None	47.22%	47.22%	47.22%	47.22%	47.22%	31.48%	28.33%	23.61%
4096 × 2048	None	100.00%	100.00%	100.00%	100.00%	100.00%	70.83%	63.75%	53.13%
3840 × 1920	None	100.00%	100.00%	100.00%	100.00%	100.00%	80.59%	72.53%	60.44%
6144 × 3072	TaB	23.61%	23.61%	23.61%	23.61%	23.61%	15.74%	14.71%	11.81%
4096 × 2048	TaB	53.13%	53.13%	53.13%	53.13%	53.13%	35.42%	31.88%	26.56%
3840 × 1920	TaB	60.44%	60.44%	60.44%	60.44%	60.44%	40.30%	36.27%	30.22%

A.2.3 Conversion of ERP Signals to CMP

A.2.3.1 General

The 3D XYZ coordinate system as shown in Figure A.1 can be used to describe the 3D geometry of ERP and CMP projection format representations. Starting from the center of the sphere, X axis points toward the front of the sphere, Z axis points toward the top of the sphere, and Y axis points toward the left of the sphere.

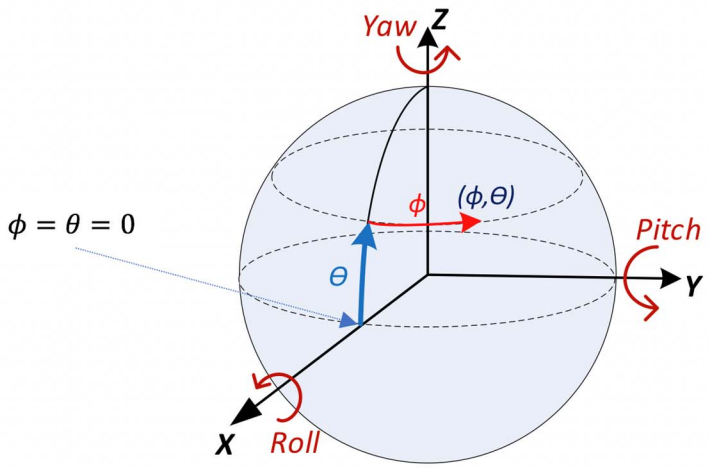

Figure A.1: 3D XYZ coordinate definition

NOTE: The text in this Annex is based on JVET output document (JVET-H1004: Algorithm descriptions of projection format conversion and video quality metrics in 360Lib (Version 5)) with simplifications to only two projection types which are used in the present document and further fixes regarding misalignments. C++ implementation of the concepts described by this annex is available in 360Lib Software: available at: https://jvet.hhi.fraunhofer.de/svn/svn_360Lib.

The coordinate system is specified for defining the sphere coordinates azimuth (ϕ) and elevation (θ) for identifying a location of a point on the unit sphere. The azimuth ϕ is in the range [−π, π], and elevation θ is in the range [−π/2, π/2], where π is the ratio of a circle's circumference to its diameter. The azimuth (ϕ) is defined by the angle starting from X axis in counter-clockwise direction as shown in Figure A.1. The elevation (θ) is defined by the angle from the equator toward Z axis as shown in Figure A.1. The (X, Y, Z) coordinates on the unit sphere can be evaluated from (ϕ, θ) using following equations:

$$X = \cos(\theta)*\cos(\phi)$$

$$Y = \cos(\theta)*\sin(\phi)$$

$$Z = \sin(\theta)$$

Inversely, the longitude and latitude (ϕ, θ) can be evaluated from (X, Y, Z) coordinates using:

$$\phi = \tan^{-1}(Y/X)$$

$$\theta = \sin^{-1}(Z/(\sqrt{X^2+Y^2+Z^2}))$$

A 2D plane coordinate system is defined for each face in the 2D projection plane. Where Equirectangular Projection (ERP) has only one face, Cubemap Projection (CMP) has six faces. In order to generalize the 2D coordinate system, a face index is defined for each face in the 2D projection plane. Each face is mapped to a 2D plane associated with one face index.

A.2.3.2 Equirectangular Projection (ERP)

Equirectangular mapping is the most commonly used mapping from spherical video to a 2D texture signal. The mapping is bijective, i.e. it may be expressed in both directions and is illustrated in Figure A.2.

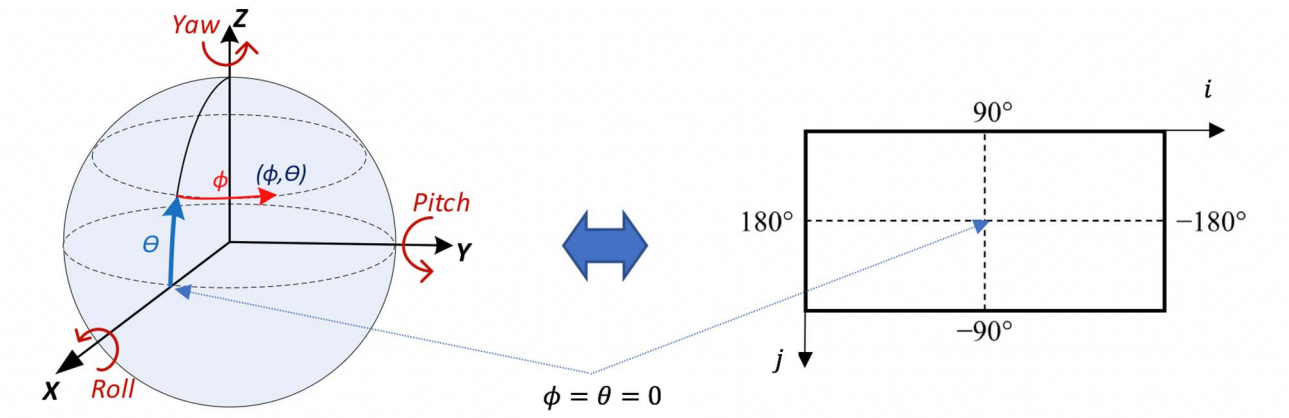

Figure A.2: Mapping of spherical video to a 2D texture signal

ERP has only one face and the face index f for ERP is always set to 0. The sphere coordinates (ϕ, θ) for a sample location (i, j), in degrees, are given by the following equations:

$$\phi = (0.5 - i/\text{pictureWidth})*360$$

$$\theta = (0.5 - j/\text{pictureHeight})*180$$

Finally, (X, Y, Z) can be calculated from the equations given above.

A.2.3.3 Cubemap Projection (CMP)

Figure A.3 shows the CMP projection with 6 square faces, labelled as PX, PY, PZ, NX, NY, NZ (with "P" standing for "positive" and "N" standing for "negative"). Table A.2-4 specifies the face index values corresponding to each of the six CMP faces.

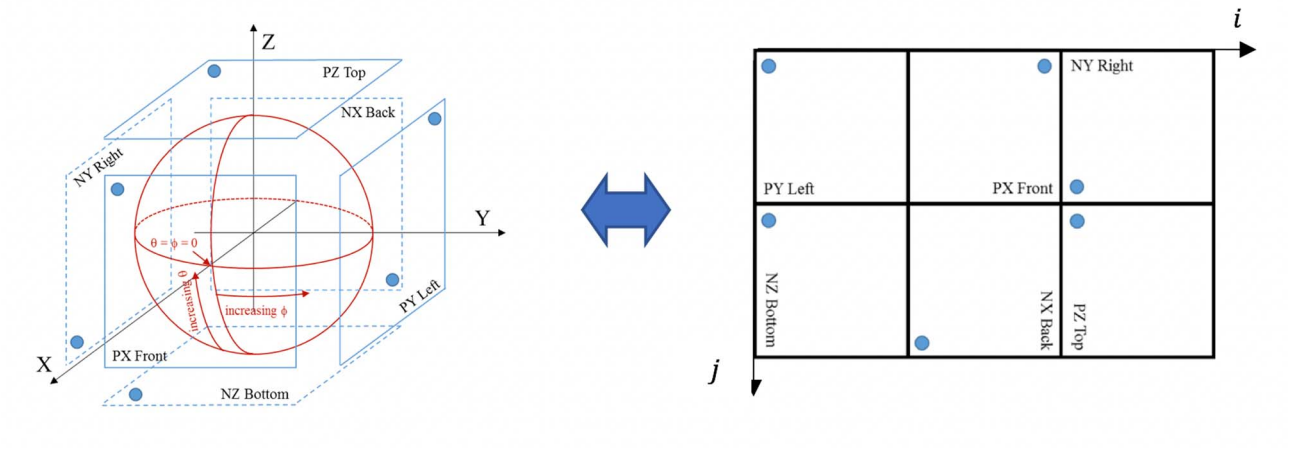

Figure A.3: Relation of the cube face arrangement of the projected picture to the sphere coordinates.

Table A.2-4: Face index of CMP

Face index	Face label	Notes
0	PX	Front face with positive X axis value
1	NX	Back face with negative X axis value
2	PY	Left face with positive Y axis value
3	NY	Right face with negative Y axis value
4	PZ	Top face with positive Z axis value
5	NZ	Bottom face with negative Z axis value

The 3D coordinates (X, Y, Z) are derived using following equations:

```
lw = pictureWidth / 3
lh = pictureHeight / 2
tmpHorVal = i − Floor( i ÷ lw ) * lw
tmpVerVal = j − Floor( j ÷ lh ) * lh
i′ = −( 2 * tmpHorVal ÷ lw ) + 1
j′ = −( 2 * tmpVerVal  ÷ lh ) + 1
w = Floor( i ÷ lw )
h = Floor( j ÷ lh )
if( w  = =  1  &&  h  = =  0 ) { // PX: positive x front face
   X = 1.0
   Y = i′
   Z = j′
} else if( w  = =  1  &&  h  = =  1 ) { // NX: negative x back face
   X = −1.0
   Y = −j′
   Z = −i′
} else if( w  = =  2  &&  h  = =  1 ) { // PZ: positive z top face
   X = −i′
   Y = −j′
   Z = 1.0
} else if( w  = =  0  &&  h  = =  1 ) { // NZ: negative z bottom face
   X = i′
   Y = −j′
   Z = −1.0
} else if( w  = =  0  &&  h  = =  0 ) { // PY: positive y left face
   X = −i′
   Y = 1.0
   Z = j′
} else { // ( w  = =  2  &&  h  = =  0 ), NY: negative y right face
   X = i′
   Y = −1.0
   Z = j′
}
```

A.2.3.4 Conversion between two projection formats

Denote (f_d, i_d, j_d) as a point (i_d, j_d) on face f_d in the destination projection format, and (f_s, i_s, j_s) as a point (i_s, j_s) on face f_s in the source projection format. Denote (X, Y, Z) as the corresponding coordinates in the 3D XYZ space. The conversion process starts from each sample position (f_d, i_d, j_d) on the destination projection plane, maps it to the corresponding (X, Y, Z) in 3D coordinate system, finds the corresponding sample position (f_s, i_s, j_s) on the source projection plane, and sets the sample value at (f_d, i_d, j_d) based on the sample value at (f_s, i_s, j_s).

Therefore, the projection format conversion process from ERP source format to CMP destination format is performed in the following three steps:

1) Map the destination 2D sampling point (f_d, i_d, j_d) to 3D space coordinates (X, Y, Z) based on the CMP format.

2) Map (X, Y, Z) from step 1 to 2D sampling point (f_0, i_s, j_s) based to the ERP format.

3) Calculate the sample value at (f_0, i_s, j_s) by interpolating from neighboring samples at integer positions on face f_0, and the interpolated sample value is placed at (f_d, i_d, j_d) in the destination projection format.

The above steps are repeated until all sample positions (f_d, i_d, j_d) in the destination projection format are filled. Note that (Step 1) and (Step 2) can be pre-calculated at the sequence level and stored as a lookup table, and only (Step 3) needs to be performed per sample position for each picture in order to render the sample values.

Annex B (informative): Example External Binaural Renderer

B.1 General

Binaural rendering allows 3D audio content to be played back via headphones. The rendering is performed as a fast convolution of point sound source streams in the 3D space with head-related impulse responses (HRIRs) or binaural room impulse responses (BRIRs) corresponding to the direction of incidence relative to the listener. HRIRs will be provided from an external source.

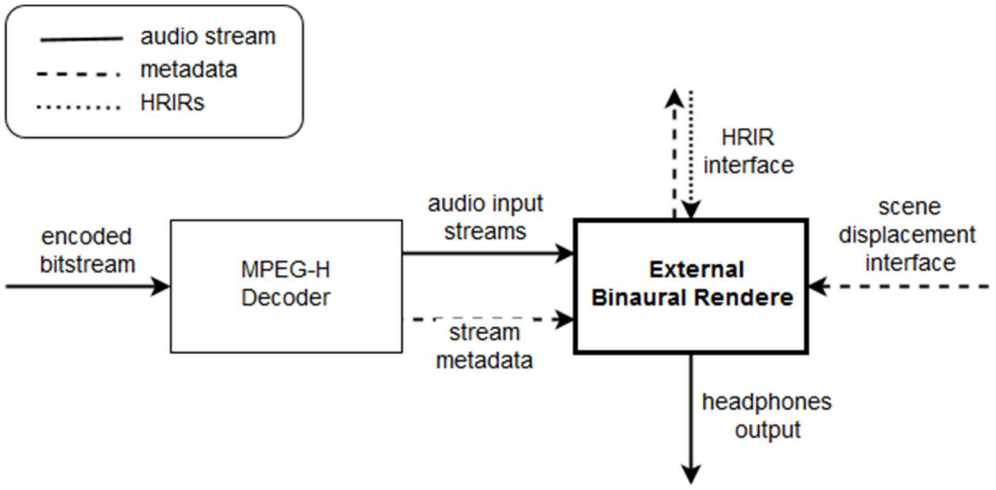

Figure B.1-1: High level overview of an external binaural renderer setup.

The renderer has three input interfaces (see Fig. B.1-1): the audio streams and metadata from the MPEG-H decoder, a head tracking interface for scene displacement information (for listener tracking), and a head-related impulse response (HRIR) interface providing binaural impulse responses for a given direction of incidence. The metadata as described in B.3, together with the scene displacement information, is used to construct a scene model, from which the renderer can infer the proper listener-relative point source positions.

The audio input streams may include Channel content, Object content, HOA content. The renderer performs pre-processing steps to translate the respective content type into several point sources that are then processed for binaural rendering. Channel groups and objects that are marked a non-diegetic in the metadata are excluded from any scene displacement processing.

B.2 Interfaces

B.2.1 Interface for Audio Data and Metadata

The example external binaural renderer has an interface for the input of un-rendered channels, objects, and HOA content and associated metadata. The syntax of this input interface follows the specification of the External Renderer Interface for MPEG-H 3D Audio to output un-rendered channels, objects, and HOA content and associated metadata according to clause 6.1.4.3.6.5.

The input PCM data of the channels and objects interfaces is provided through an input PCM buffer, which first contains $n_{chan,\ out}$ signals carry the PCM data of the channel content. These are followed by $n_{obj,\ out}$ signals carrying the PCM data of the un-rendered objects. Then additional signals carry the $n_{HOA,\ out}$ HOA data which number is indicated in the HOA metadata via the HOA order (e.g. 16 signals for HOA order 3). The HOA audio data in the HOA interface is provided in the ESD representation. The conversion from the HOA domain into the equivalent spatial domain representation and vice versa is described in ISO/IEC 23008-3 [19], Annex C.5.1.

The metadata for channels, objects, and HOA is received via the input interface once per frame and their syntax is specified in `mpegh3da_getChannelMetadata()`, `mpegh3da_getObjectAudioAndMetadata()`, and `mpegh3da_getHoaMetadata()` respectively, see ISO/IEC 23008-3, clause 17.10 [19]. The metadata and PCM data will be aligned to match each metadata element with the respective PCM frame.

B.2.2 Head Tracking Interface.

The external binaural renderer receives scene displacement values (yaw, pitch and roll) e.g. from an external head tracking device via the head tracking interface. The syntax is specified in `mpegh3daSceneDisplacementData()` as defined in ISO/IEC 23008-3 [19], clause 17.9.3.

B.2.3 Interface for Head-Related Impulse Responses

An interface is provided to specify the set of HRIRs used for the binaural rendering. These directional FIR filters will be input using the SOFA (Spatially Oriented Format for Acoustics) files format according to AES-69 [21]. The SimpleFreeFieldHRIR convention will be used, where binaural filters are indexed by polar coordinates (azimuth φ in radians, elevation ϕ in radians, and radius r in meters) relative to the listener.

B.3 Preprocessing

B.3.1 Channel Content

Channel input content is converted into a corresponding set of point sources with associated positions using the loudspeaker configuration data included in `mpegh3da_getChannelMetadata()` and the associated PCM data obtained via the interface specified in B.2.1

B.3.2 Object Content

Object input content is converted into corresponding point sources with associated positions using the metadata included in `mpegh3da_getObjectAudioAndMetadata()` and the associated PCM data obtained via the interface specified in clause B.2.1

B.3.3 HOA Content

As specified in clause B.2.1 HOA content is input in the ESD representation together with the metadata included in `mpegh3da_getHoaMetadata()`. As a pre-processing step, the ESD representation is first converted into HOA coefficients. All coefficients associated with HOA of order larger than three are discarded to limit the maximum computational complexity.

B.3.4 Non-diegetic Content

Channel groups for which the gca_directHeadphone flag is set in `mpegh3da_getChannelMetadata()` are routed to left and right output channel directly and are excluded from binaural rendering using scene displacement data (non-diegetic content). Non-diegetic content may have stereo or mono format. For mono, the signal is mixed to left and right headphone channel with a gain factor of 0.707.

For each channel group it has to be checked in the `mpegh3da_getChannelMetadata()` if the `gca_fixedChannelsPosition` flag is equal to 0 or 1. A channel group with an associated `'gca_fixedChannelsPosition` == 1' is included in the binaural rendering but excluded from the scene displacement processing according to clause B.4, i.e. its position is not updated.

For each object it has to be checked in the `mpegh3da_getObjectAudioAndMetadata()` if the goa_fixedPosition flag is equal to 0 or 1. An object with an associated `'goa_fixedPosition` == 1' is included in

the binaural rendering but excluded from the scene displacement processing according to clause B.4, i.e. its position is not updated.

B.4 Scene Displacement Processing

B.4.1 General

The position of each point source derived from the channels and objects input is represented by a 3-dimensional vector $\vec{s}_c$ in a Cartesian coordinate system. The scene displacement information is used to compute an updated version of the position vector $\vec{s}_c'$ as described in clause B.4.2. The position of point sources that result from non-diegetic channel groups with an associated `'gca_fixedChannelsPosition == 1'` or from non-diegetic objects with an associated `'goa_fixedPosition == 1'` (see clause B.3.4) is not updated, i.e. $\vec{s}_c'$ is equal to $\vec{s}_c$.

B.4.2 Applying Scene Displacement Information

The vector representation of a point source $\vec{s}_c$ is transformed to the listener-relative coordinate system by rotation based on the scene displacement values obtained via the head tracking interface. This is achieved by multiplying the position $\vec{s}_c$ with a rotation matrix calculated from the orientation of the listener:

$$\vec{s}_c' = T_{rot}\vec{s}_c$$

The determination of the rotation matrix T_{rot} is defined in ISO/IEC 23008-3 [19], Annex I.

For HOA content, the rotation matrix $T_{rot,hoa}$ suited for rotating the spherical harmonic representation is calculated as defined in ISO/IEC 23008-3 [19], Annex I. After the rotation, the HOA coefficients are transformed back into the ESD representation. Each ESD component is then converted to the corresponding point source with its associated positional information. For the ESD components the position information is fixed, i.e. $\vec{s}_c' = \vec{s}_c$, as the rotation due to scene displacement is performed in the spherical harmonic representation.

B.5 Headphone Output Signal Computation

B.5.1 General

The overall Scene Model is represented by the collection of all point sources with updated position $\vec{s}_c'$ obtained from the rotated channels, objects, and the ESD components as well as the non-diegetic channels and objects for which `'gca_fixedChannelsPosition == 1'` or `'goa_fixedPosition == 1'`. The overall number of point sources in the Scene Model is denoted with C.

B.5.2 HRIR Selection

The position $\vec{s}_c'$ of each point source in the listener-relative coordinate system is used to query a best-match HRIR pair from the set of available HRIRs. For lookup, the polar coordinates of the HRIR locations are transformed into the internally used cartesian coordinates and the closest-match available HRIR for a given point source position is selected. As no interpolation between different HRIRs is performed, HRIR datasets with sufficient spatial resolution should be provided.

B.5.3 Initialization

The HRIR filters used for binauralization are asynchronously partitioned and transformed into the frequency domain using a Fast FourierTransform (FFT). The necessary steps for each of the C HRIR filter pairs are as follows:

1) Uniformly partition the length N HRIR filter pairs $f_{c,L/R}(n)$ into $P = \lceil \mathrm{N}/\mathrm{B} \rceil$ filter partitions $f_{c,p,L/R}(n)$ of length B.

2) Zero-pad the filter partitions to length K.

3) Transform all filter partitions into the frequency domain using real-to-complex FFT to obtain the P frequency domain filter pairs $F_{c,p,L/R}(k)$, where k denotes the frequency index.

B.5.4 Convolution and Crossfade

Each audio block of a point source of the Scene Model is convolved with its selected HRIR filter pair for the left and right ear respectively. To reduce the computational complexity, a fast frequency domain convolution technique of uniformly partitioned overlap-save processing is useful for typical FIR filter lengths for HRIRs/BRIRs. The required processing steps are described in the following.

The following block processing steps are performed for each of the C point sources of the Scene Model:

a) Obtain a block of B new input samples $x_c(n)$ of the point source c .

b) Perform a real-to-complex FFT transforms of length K to obtain the frequency domain representation of the input $X_c(k)$.

c) Compute the frequency domain headphone output signal pair $Y_{c,L/R}(k)$ for the point source c by multiplying each HRIR frequency domain filter partition $F_{c,p,L/R}(k)$ with the associated frequency domain input block and adding the product results over all partitions.

d) K samples of the time domain output signal pair $y_{c,L/R}(n)$ are obtained from $Y_{c,L/R}(k)$ by performing a complex-to-real IFFT.

e) Only the last B output samples represent valid output samples. The $K - B$ samples before are time-aliased and are discarded.

f) In case of a HRIR filter exchange happens due to changes in the scene displacement, steps 3-5 are computed for both the current HRIR filter and the ones used in the previous block. A time-domain crossfade is performed over the B output samples obtained in step 5:

g) $y_{c,L/R}(n) = w_{in}(n)y_{c,L/R,current}(n) + w_{out}(n)y_{c,L/R,prev}(n)$

h) The crossfade envelopes are defined as

$$w_{in}(n) = \sin^2\left(\frac{n\pi}{2B}\right)$$

$$w_{out}(n) = \cos^2\left(\frac{n\pi}{2B}\right)$$

to preserve a constant power of the resulting output signal.

The crossfade operation define in step 6 is only applied to point sources of the Scene Model that have been generated from channel or object content. For HOA content, the crossfade is applied between the current and the previous rotation matrices $T_{rot,hoa}$ (see B.4.2).

B.5.5 Binaural Downmix

The rendered headphone output signal is computed as the sum over all binauralized point source signal pairs $y_{c,L/R}(n)$. In case that the metadata provided together with the audio data at the input interface (see X.3.1) includes gain values applicable to a specific channel group (`gca_channelGain` in `mpegh3da_getChannelMetadata()`) or objects (`goa_objectGainFactor` in `mpegh3da_getObjectAudioAndMetadata()`), these gain values g_c are applied to the corresponding binauralized point source signal $y_{c,L/R}(n)$ before the summation:

$$y_{L/R}(n) = \sum_{c=0}^{C-1} g_c y_{c,L/R}(n)$$

Finally, any additional non-binauralized non-diegetic audio input (`'gca_directHeadphone == 1'`, see B.3.4) is added time-aligned to the two downmix channels.

B.5.6 Complexity

The algorithmic complexity of the external binaural renderer using a fast convolution approach can be evaluated for the following computations:

Convolution (B.5.4)	1) RFFT: $C * F * K * log_2 K$ (with $F = 2.5$ as an estimated additional complexity factor for the FFT) 2) complex multiplications: $2 * P * C * \left\lceil \frac{K+1}{2} \right\rceil$ 3) complex additions: $2 * P * \lceil \frac{K+1}{2} \rceil$ 4) IRFFT: $2 * C * F * K * log_2 K$
Downmix (B.5.5)	1) real multiplications: $2 * C * B$ 2) real additions: $2 * C * B$
Filter Exchange and Crossfade (B.5.4)	1) RFFT: $2 * C * P * F * K * log_2 K$ 2) Time-domain crossfade (real multiplications): $4 * C * B$ 3) Time-domain crossfade (real additions): $2 * C * B$

Additional computations are required for scene displacement processing (see B.4).

The total complexity per output sample can be determined by adding the complexity estimation for convolution and downmix and dividing by the block length B. In blocks where a filter exchange is performed, items 2-4 from the convolution contribute two times to the overall complexity in addition to the time-domain crossfade multiplications and additions (filter exchange items 2 and 3). The partitioning and FFT for the filter exchange, as well as the scene displacement, can be performed independent of the input block processing.

B.5.7 Motion Latency

The Scene Model can be updated with arbitrary temporal precision, but the resulting HRIR exchange is only done at processing block boundaries of the convolution. With a standard block size of $B = 256$ samples at 48 kHz sampling rate, this leads to a maximum onset latency of 5.3 ms until there is an audible effect of a motion of sources or the listener. In the following block, a time-domain crossfade between the new and the previous filtered signal is performed (see Convolution/Initialization), so that a discrete, instantaneous motion is completed after a maximum of two convolution processing blocks (10.6 ms for 512 samples at 48 kHz sampling rate). Additional latency from head trackers, audio buffering, etc. is not considered.

The rotation of the HOA content is performed at a block boundary resulting in a maximum latency of one processing block, until a motion is completed.

Annex C (informative): Registration Information

C.1 3GPP Registered URIs

The clause documents the registered URIs in this specification following the process in http://www.3gpp.org/specifications-groups/34-uniform-resource-name-urn-list

Table C-1 lists all registered URN values as well as

- a brief description of its functionality;
- a reference to the specification or other publicly available document (if any) containing the definition;
- the name and email address of the person making the application; and
- any supplementary information considered necessary to support the application.

Table C-1: 3GPP Registered URNs

URN	Description	Reference	Contact	Remarks
urn:3GPP:vrstream:mp:video:basic	DASH profile identifier for VR Streaming Basic Video Media Profile	TS 26.118, clause 5.2.2.3	Thomas Stockhammer tsto@qti.qualcomm.com	none
urn:3GPP:vrstream:mp:video:main	DASH profile identifier for VR Streaming Main Video Media Profile	TS 26.118, clause 5.2.3.3	Thomas Stockhammer tsto@qti.qualcomm.com	none
urn:3GPP:vrstream:mp:video:advanced	DASH profile identifier for VR Streaming Advanced Video Media Profile	TS 26.118, clause 5.2.4.3	Thomas Stockhammer tsto@qti.qualcomm.com	none
urn:3GPP:vrstream:ve:<id>	Viewpoint Descriptor Scheme for Ensemble Signaling	TS 26.118, clause 5.2.3.3.4	Thomas Stockhammer tsto@qti.qualcomm.com	none
urn:3GPP:vrstream:presentation	DASH profile identifier for VR DASH Media Presentation	TS 26.118, clause 8.3	Thomas Stockhammer tsto@qti.qualcomm.com	none

Annex D (informative): VR metrics calculation examples

D.1 Comparable quality viewport switching latency

This sub-clause illustrates how the weighted average QR value and the effective resolution can be calculated.

The quality level of each region is determined with its respective quality ranking (QR) value. A viewport can be covered with multiple regions. A quality level value for the viewport can be derived as weighted average of the QR values of the regions covering the viewport. The weight of each region is defined as the percentage of the viewport area covered by the corresponding region. The viewport quality level can be calculated by the following equation.

$$QualityLevel(viewport) = \sum_{i=1}^{N} (QR[i] \times Coverage[i]/100)$$

N: Number of regions covering the viewport

$QR[i]$: QR value of i-th quality ranking region

$Coverage[i]$: The viewport coverage value (in percent) of i-th quality ranking region

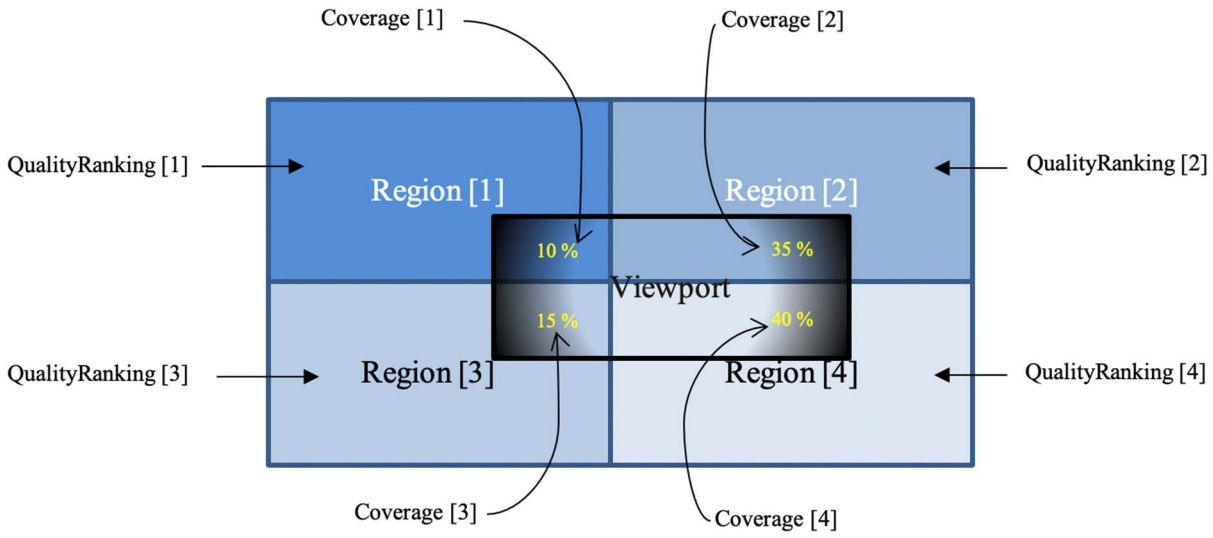

Figure D.1-1: An example of a viewport covered by four quality ranking 2D regions

Figure D.1-1 is an example of a viewport covered by four quality ranking regions. The quality of the viewport is equal to the weighted sum of the quality ranking value and the coverage percentage value of each quality ranking region.

The resolution of each region is determined by its respective width and height values in pixel which are available in the quality ranking box under the name orig_width and orig_height. Note that these values are already normalized to represent the full-sphere resolution you would get if the resolution of this region would be used for the full sphere.

The effective resolution (i.e. the total number of original pixels) for the content visible in the viewport can be derived as the weighted average of the resolution of each region covering the viewport. The weight of each

region is defined by the percentage of the viewport area covered by the corresponding region. The effective viewport resolution can be calculated by the following equation.

$$EffectiveResolution(viewport) = \sum_{i=1}^{N} (width[i] \times height[i] \times Coverage[i]/100)$$

N: Number of regions covering the viewport

$width[i]$: The width component of the original source pixel resolution for the i-th quality ranking region

$height[i]$: The height component of the original source pixel resolution resolution for the i-th quality ranking region

$Coverage[i]$: The viewport coverage value (in percent) of i-th quality ranking region

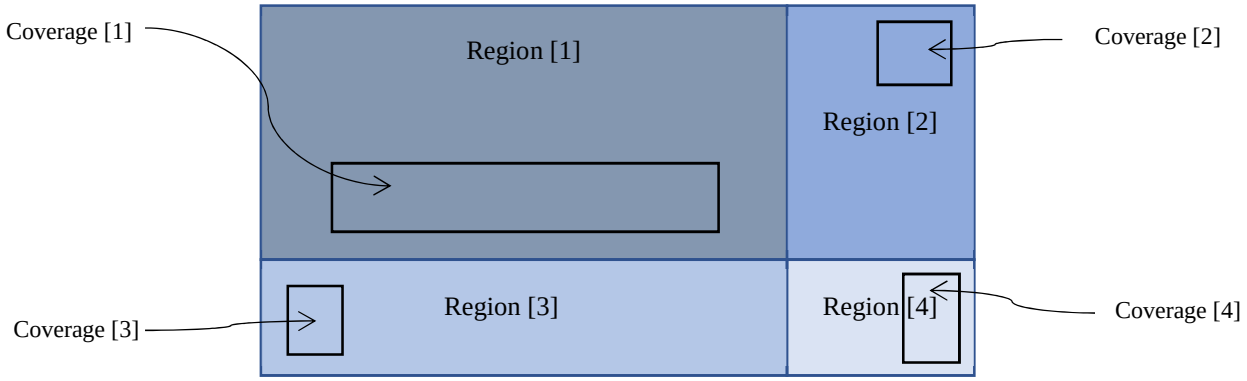

Figure D.1-2: An example of a source packed image with four quality ranking 2D regions with different resolutions

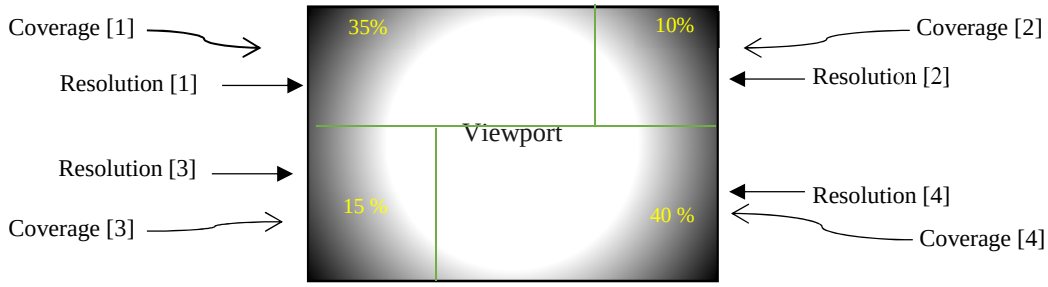

Figure D.1-3: An example of a viewport covered by four different quality ranking 2D regions

Figure D.1-2 is an example of a source with four regions with different resolution. Figure D.1-3 represents an example of a viewport which is covered by the four quality ranking 2D regions. The effective viewport resolution is equal to the weighted sum of the resolution for each quality-ranking 2D region and its corresponding viewport coverage percentage value.

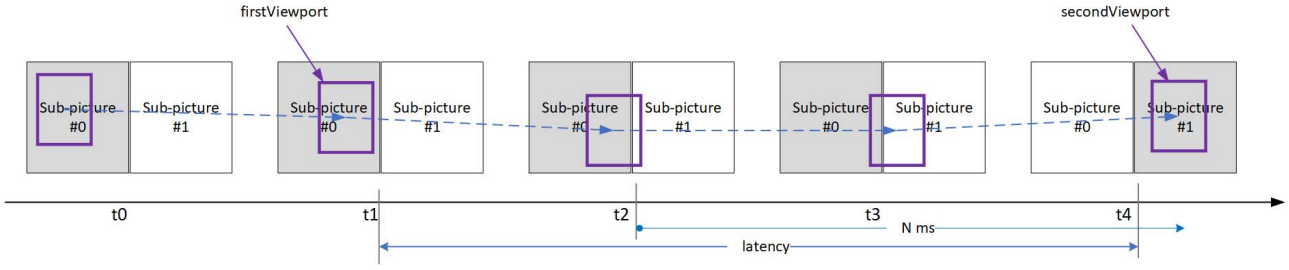

Figure D.1-4 – Comparable quality viewport switching latency measurement example

Figure D.1-4 presents an example of the metric measurement operation. The viewport quality is evaluated at time t0, and then again at time t1. The media playback module renders the high-resolution sub-picture #1 at time t1. The user viewing orientation is gradually changing from sub-pic#1 to sub-pic#2 as the time progresses.

At time t2, the media playback module starts to render the buffered low-quality representation of sub-pic#2 as the viewport moves into sub-picture #2. At time t2, the viewport quality drops in values as compared to the viewport quality at time t1, and a new sub-picutre (sub-pic #2) is rendered. A viewport switching event is identified at time t2.

The viewport quality values evaluated at t1 identifies the first viewport. The viewport position and viewport quality level list are assigned to the attributed `Position` and `QualityLevel` of the `firstViewportItem`.

An effective viewport resolution and viewport QR quality value for the new viewport that is comparable to that of the `firstViewportItem` after viewport switching time is logged at time t4. The new viewport position identifies the `Position` of the `secondViewportItem`. The corresponding `QualityLevel` list for the `secondViewportItem` is assigned.

The associated viewport values stored for the worst viewport quality during the switch is assigned to the field `Position` of the `worstViewportItem.` The corresponding `QualityLevel` list for the `worstViewportItem` is also assigned.

The comparable-quality viewport switching latency is measured as the time interval between the logged times for `firstViewportItem` (t1 in this example) and `secondViewportItem` (t4 in this example).

D.2 Rendered viewports

Figure D.2-1 illustrates an example of clustering and the associated viewports. The first three evaluated viewports are all with the distance D (indicated by the blue circle), and are thus assigned to the same cluster. Note that the cluster center moves a bit for each new viewport which is added to the cluster.

Viewport #4 is too far away from the center of cluster #1, and thus starts a new cluster, which eventually gathers three viewport members. Then viewport #7 is too far away from the center of cluster #2, and again starts a new cluster.

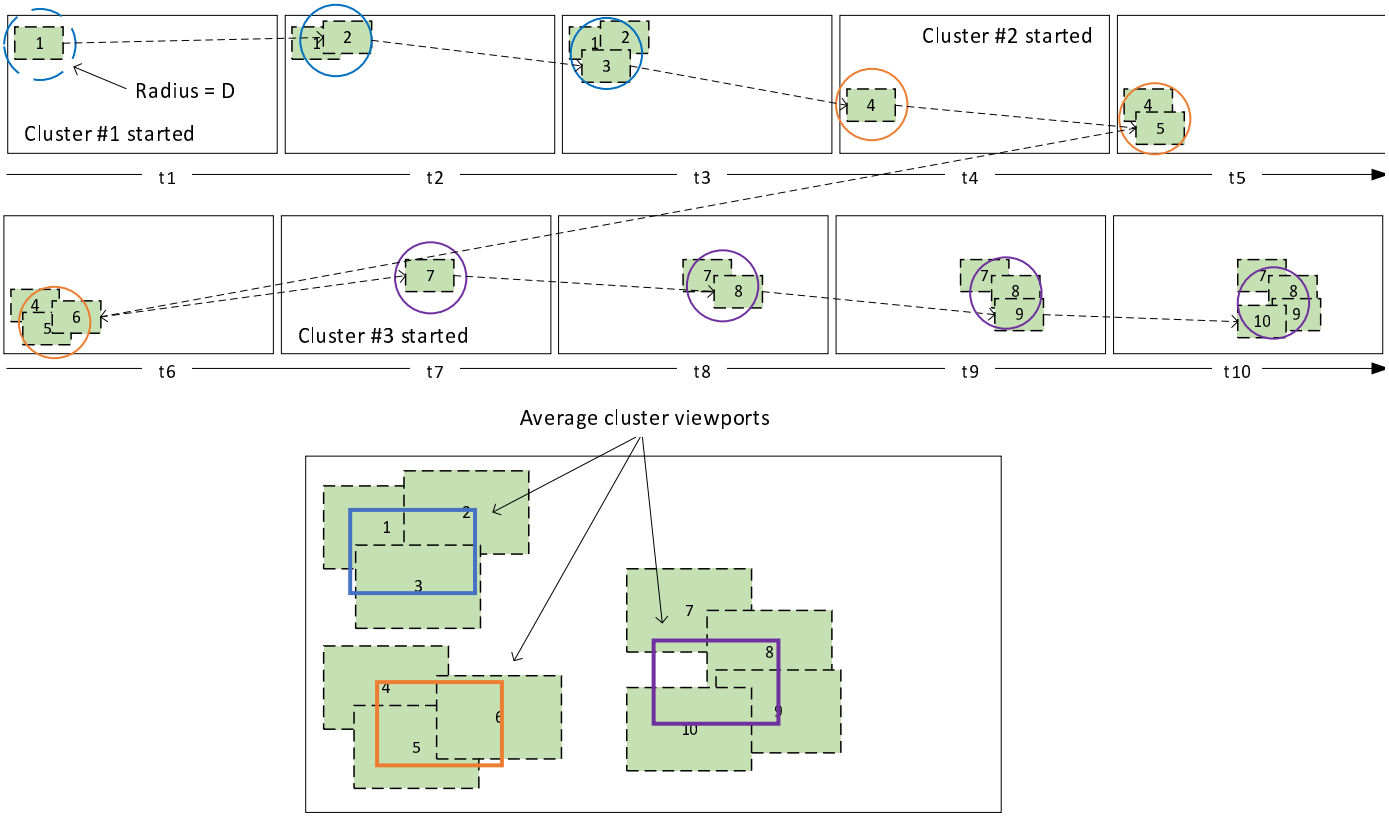

Figure D.2-1 Clustering example

For each cluster *j*, the final averaged viewport parameters can be derived as follows, assuming there are *N* viewports in the *j:th* cluster. Note that the center azimuth and tilt averaging also needs to handle the special case around -180/180 degrees, as some values might be positive (e.g. 176 degrees), while others might be negative (e.g. -178 degrees). This special case is not shown in the equations below.

Note also that the azimuth and elevation range (i.e. the visible coverage of the viewport) might often be the same for every viewport, unless the user explicitly changes the field-of-view for the device. For consistency, and to catch any during-session field-of-view changes, these two parameters should still be averaged.

$$average\_centre\_azimuth[j] = \sum_{i=1}^{N} centre\_azimuth[i] / N$$

$$average\_centre\_elevation[j] = \sum_{i=1}^{N} centre\_elevation[i] / N$$

$$average\_centre\_tilt[j] = \sum_{i=1}^{N} centre\_tile[i] / N$$

$$average\_azimuth\_range[j] = \sum_{i=1}^{N} azimuth\_range[i] / N$$

$$average\_elevation\_range[j] = \sum_{i=1}^{N} elevation\_range[i] / N$$

Figure D.2-2 below illustrates an example of the duration filtering. The user starts by looking at the upper left part of the media (viewports #1 to #3), then make a very brief glance to the right (viewport #4), and then moves back to the upper-left again (viewports #5 and #6). Then the user moves his gaze to the lower-right part (viewports #7 to #10).

Assume here that the duration T is set to 4 times the value of the viewport sample rate X, i.e. a cluster needs to have a duration corresponding to at least four viewports to be reported. Here four clusters are formed, but before filtering only cluster #4 would be reported. After filtering, clusters #1 and #3 are close enough both in time and distance to add to

each other's aggregated duration, so each of them will be be assigned an aggregated duration of 5, and thus be reported. Cluster #2, the quick gance up to the right, has too short duration and will not be reported.

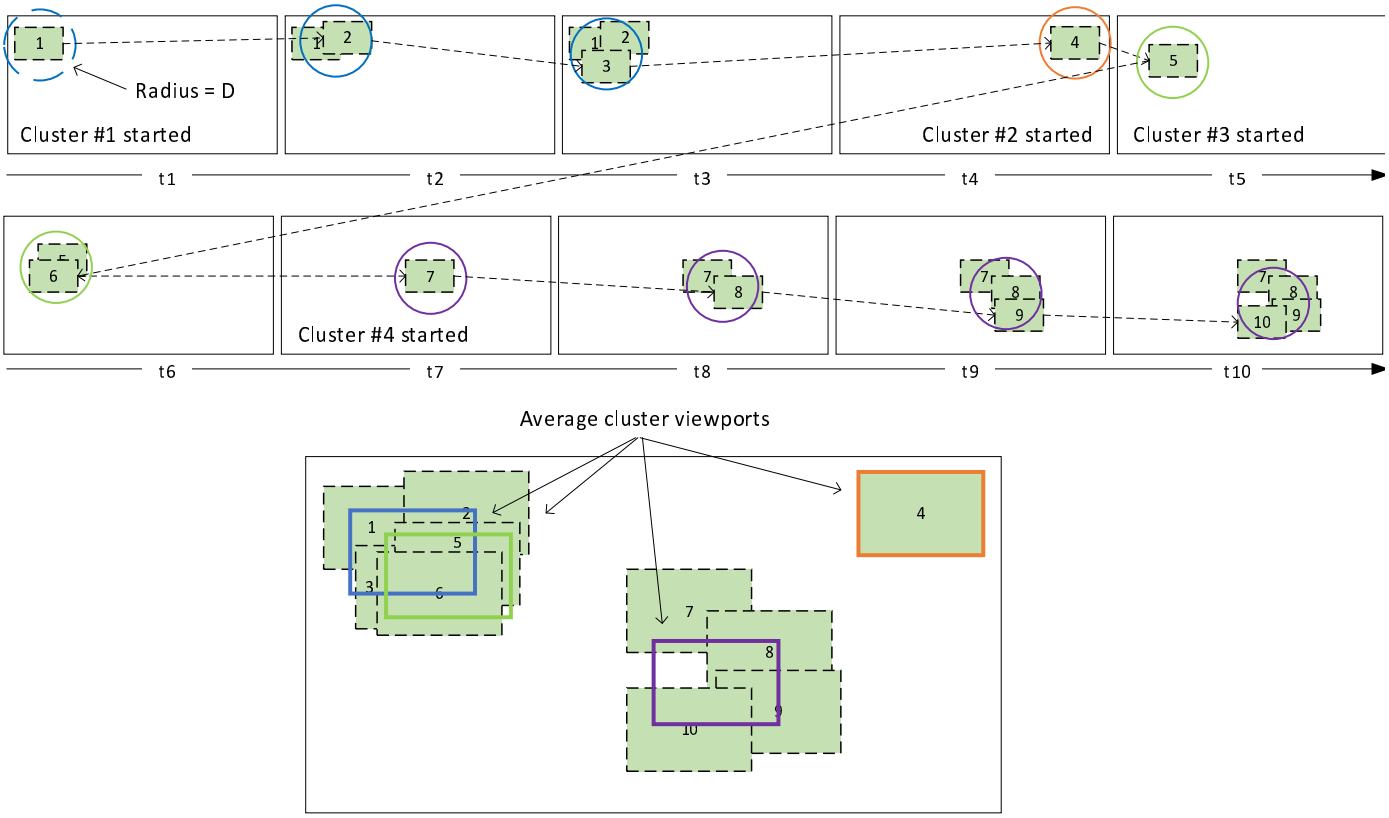

Figure D.2-2 Duration filtering example

Annex E (informative): Change history

Change history							
Date	**Meeting**	**TDoc**	**CR**	**Rev**	**Cat**	**Subject/Comment**	**New version**
2018-06	SA#80	SP-180270				Presented to TSG SA#80 (for information)	1.0.0
2018-09	SA#81	SP-180647				Presented to TSG SA#81 (for approval)	2.0.0
2018-09	SA#81					Approved at TSG SA#81	15.0.0
2018-12	SA#82	SP-180967	0001	2	F	Corrections to 26.118	15.1.0
2019-09	SA#85	SP-190649	0002	-	F	Correction of figure references	15.2.0
2020-03	SA#87-e	SP-200038	0003	-	B	Addition of Feature	16.0.0
2020-03	SA#87-e					Editorial Correction in Change History, Copyright etc	16.0.1
2020-03	Post SA#87-e					Minor Editorial changes	16.0.2
2020-12	SA#90-e	SP-200932	0005		A	Corrections to Video Operation Points	16.1.0
2021-01	Post SA#90-e					Update of History Table	16.1.1

History

Document history		
V16.1.1	January 2021	Publication