Live VR end-to-end workflows: real-life deployments and advances in VR and network technology

Mauricio Aracena, VRIF President

Virtual Reality Industry Forum / Ericsson AB, mauricio.aracena@ericsson.com

Thierry Fautier, VP Video Strategy, Harmonic

Virtual Reality Industry Forum / Harmonic Inc., Thierry.Fautier@harmonicinc.com

Ozgur Oyman, VRIF Guidelines WG Chair

Virtual Reality Industry Forum / Intel Corporation, ozgur.oyman@intel.com

Written for presentation at the

SMPTE 2020 Annual Technical Conference & Exhibition

Abstract. *Live VR sports events are gaining traction among viewers since immersive services add a closer experience than watching on a TV providing even better point of views than being at the event. The content acquisition of the Live VR feed is captured using additional cameras dedicated for this purpose (in addition to the traditional cameras). In the venue, there might be several locations where the VR content is captured. The director may choose which Live VR feeds are distributed to home delivery, either by producing one single VR feed, or by giving the choice to the end user to select which one to watch, or a by offering a combination of a produced feed and individual cameras.*

In this paper, we describe and analyze the live VR workflows from an end-to-end perspective such as production, contribution, distribution and consumption aspects and describe how this can be complementary to the traditional broadcast services with a 4K like experience. We also describe how the recent advances in VR technology can improve VR experience in an optimal bandwidth using view port dependent technologies (for instance using MPEG OMAF format) and how VR production techniques (such as volumetric) provide advance tools for capturing footage from many angles for producing live free-point of view consumption and volumetric replays. This paper also presents Live VR service deployments using those technologies and it will describe the technical challenges from an end-to-end perspective on how to achieve high quality (VR360 8K equivalent) considering aspects such as bit rate, latency, CDN configurations and device capabilities. Finally, we also address how any high speed broadband infrastructure like 5G , Fiber of DOCSIS 3.1 infrastructure and new 5G devices with their built in 8K capabilities can simplify the Live workflow, and how it can accelerate the 8K VR deployments.

Keywords. live VR, VR360, Volumetric, VR workflows, 8K, 5G, Omidirectional Media Format (OMAF).

The authors are solely responsible for the content of this technical presentation. The technical presentation does not necessarily reflect the official position of the Society of Motion Picture and Television Engineers (SMPTE), and its printing and distribution does not constitute an endorsement of views which may be expressed. This technical presentation is subject to a formal peer-review process by the SMPTE Board of Editors, upon completion of the conference. Citation of this work should state that it is a SMPTE meeting paper. EXAMPLE: Author's Last Name, Initials. 2020. Title of Presentation, Meeting name and location.: SMPTE. For information about securing permission to reprint or reproduce a technical presentation, please contact SMPTE at jwelch@smpte.org or 914-761-1100 (445 Hamilton Ave., White Plains, NY 10601).

Live VR Workflows

Live VR sports events are gaining traction among viewers since immersive services add a closer experience than watching on a TV providing even better point of views than being at the event.

Use case description

The content acquisition of the Live VR feed is captured using additional cameras dedicated for this purpose (in addition to the traditional cameras). In the venue, there might be several locations where the VR content is captured. The director may choose which Live VR feeds are distributed to home delivery, either by producing one single VR feed, or by giving the choice to the end user to select which one to watch (in the secondary screen), or a by offering a combination of a produced feed and individual cameras.

The broadcaster has support for hybrid delivery of live TV content, that is, the main Live TV event is broadcasted using “traditional means” (such as terrestrial, satellite, cable, etc.) while the VR content is delivered via unicast-streaming service in parallel with the broadcast content[1]. As in traditional Live TV, the content may be exchanged between other affiliates and even distributed to other regions including international destinations. This exchange occurs using the full 360° video[2] content and is normally at very high quality, at least 8K (and possibly even higher). Finally, a local broadcaster (or service provider) delivers the Live VR content to the home with additional bandwidth optimization, such as viewport dependent (VPD)[3] delivery. In this case, the VR headset is required to transmit information (uplink) to the delivery server or CDN. It is important to mention that in this use case VR content broadcasting (i.e. without a bidirectional connection) is not considered.

[1] This means that both feeds are not necessarily in sync.

[2] This means the full (omnidirectional) camera capture, e.g 360° video or lower such as 180° video

[3] VPD: Viewport dependent: only a portion of the omnidirectional 360 video corresponding to the viewport is delivered in high quality to reduce bandwidth and complexity of decoding at the receiver (e.g. instead of decoding 8K for full 360, VPD can achieve the same quality with lower decoding capabilities in the device).

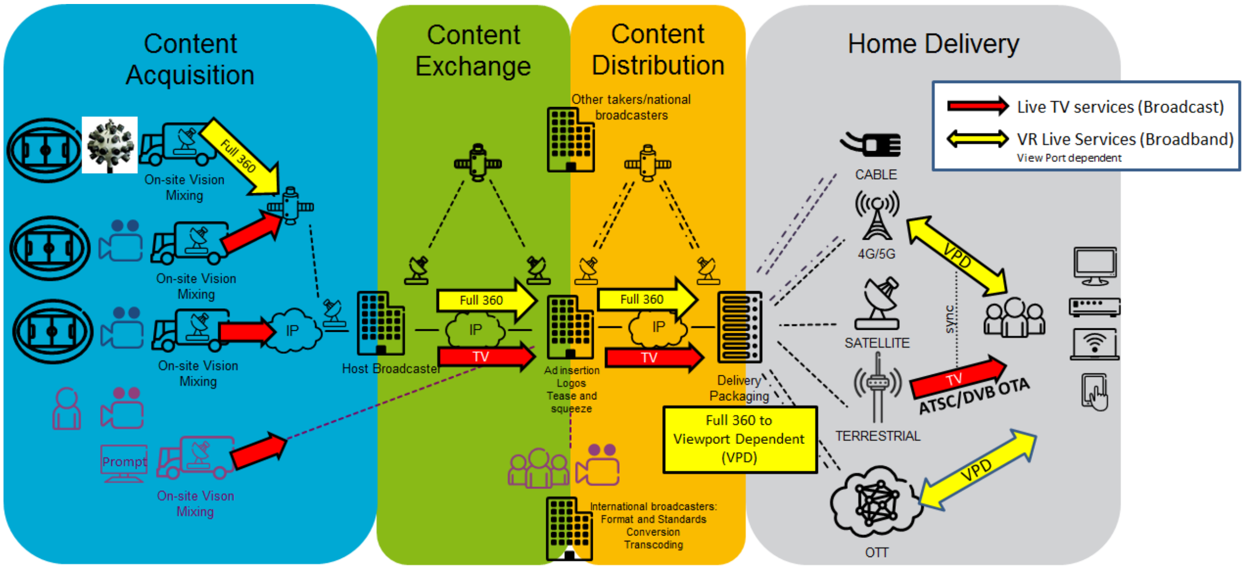

Figure 1: Live workflow for Television augmented with VR

Consumption of VR content
In order to facilitate the adoption of VR services to a wider audience, this use case also presumes the distribution of VR content to 2D displays, such as tablets, smartphones, and possibly set top boxes or even smart TVs. In the case of tablets and smartphones, the viewport is generally controlled by touching the screen or moving the device, and an STB or TV by a remote control. Other scenarios may utilize a connected smartphone or tablet to function as the swiping input mechanism for the STB or smart TV.

Reference architecture and workflows

The following diagram depicts the parallel content workflows where the live event is captured using traditional cameras (in the upper part of the diagram) and omnidirectional cameras (in the lower part of the diagram)

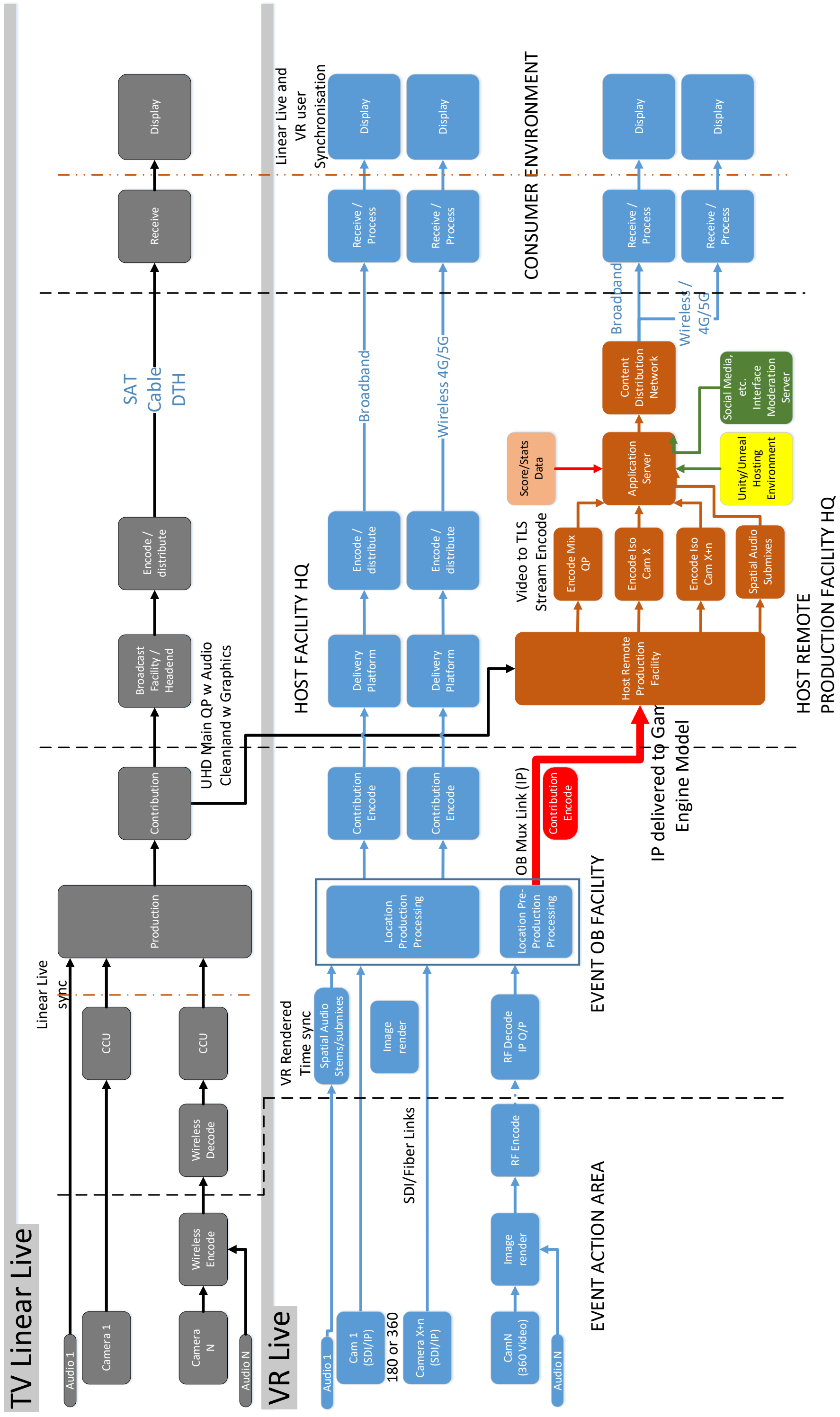

Figure 2: Overview Linear TV versus VR Live w/wo Host Remote Production Facility

It can be observed from **Figure 2** that similar workflows are utilized for traditional Live TV and Live VR. This should not be unexpected as the same operations of acquisition, production, processing and delivery apply for all entertainment experiences although the nature of the functions may differ.

Current broadcasting production workflows can be used for VR production by making use of an equirectangular projection in a vision mixer, for instance a 4K broadcast infrastructure will work without problems if the full 360 VR video is captured in 4K. If VR is captured in 8K video or even higher, then VR and traditional TV Linear live workflows will start to differ. MPEG's Omnidirectional Media Format (OMAF) format [1] can be used to distribute such 360 VR video content [2]. It is important to note that in order to take full advantage of the OMAF format with view port dependent (VPD) profile, it would require to capture a 360 video with at least 8K resolution.

Figure 3 describes the workflows in detail for Event Action Area and Event OB Facility.

Event Action Area
(1) TV Linear

- General Audio from microphones to OB facility for mixing
- Multiple Cameras using SMPTE Fiber to OB facility for processing to SDI (or IP 2110)
- Remote wireless cameras encoded with MPEG2 / 4 for wireless transmission to main OB facility for decoding to SDI (or IP 2110). Local audio embedded.

(2) VR Live

- General Audio from microphones to OB facility for mixing
  Note: if conventional Broadcaster is present – feeds may be taken here and delayed
- Spatial Audio Sub-mixer may be derived for Per-Camera view Audio feeds
- Multiple VR Cameras outputs multiplexed together onto SMPTE Fiber or on IP to OB facility for rendering to flat rectilinear image. Note: on most current "Live" camera systems the render is on-board
- Multiple VR Cameras outputs rendered locally together to flat rectilinear image. Encoded with MPEG2 / 4 for wireless transmission to main OB facility for decoding to SDI (or IP 2110). Local audio embedded.

Event OB Facility
(3) TV Linear

- Local audio input for production mix
- Camera signals decoded and processed in CCU. SDI / IP output to vision mixer
- Wireless Camera signals received, decoded and processed in CCU. SDI / IP output to vision mixer. Local audio extracted for production mix.

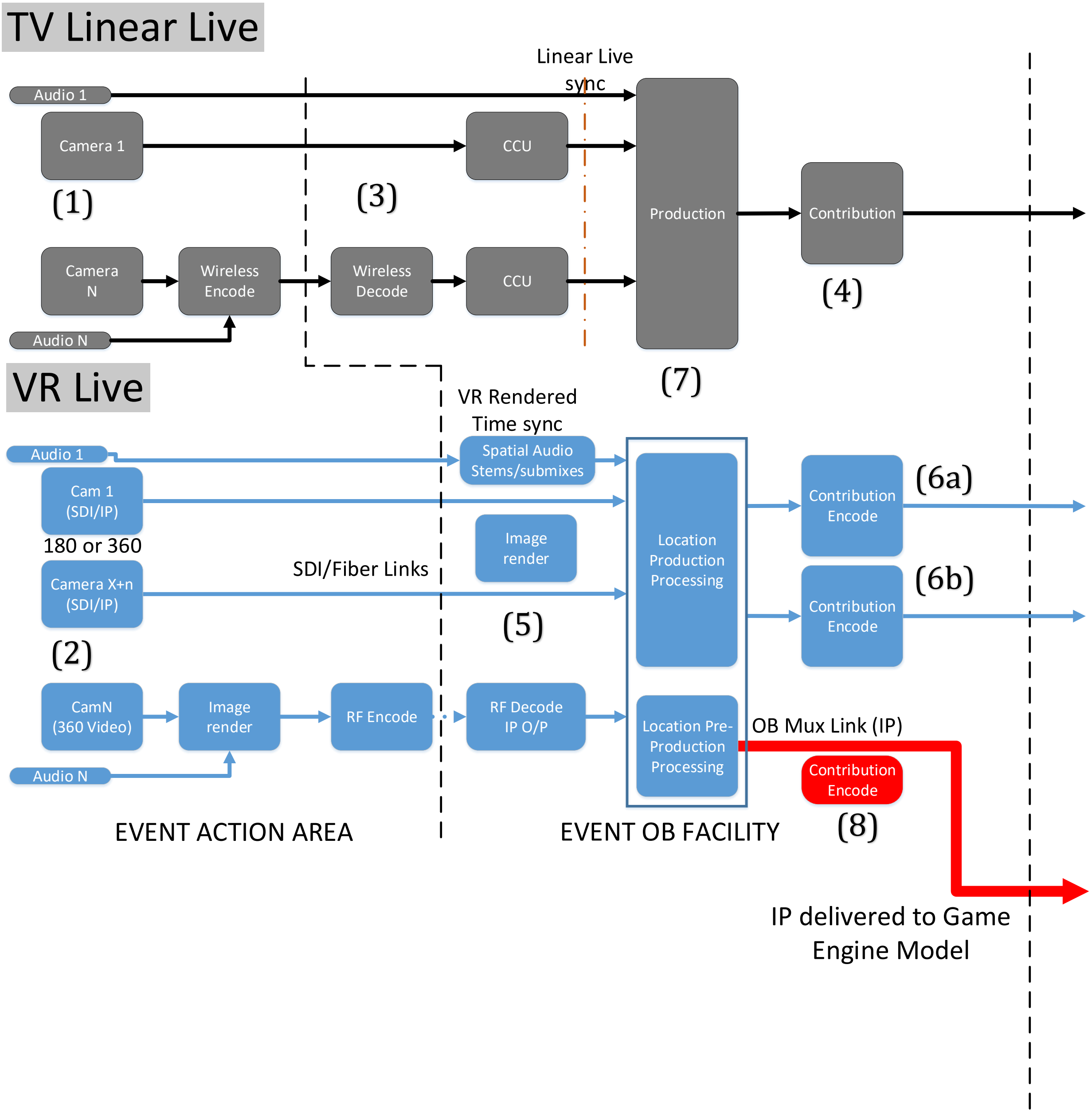

Figure 3: Event Action Area and Event OB Facility

(4) TV Linear

- Production facilities mix and cut sources, add graphics etc. to create consumer program
- Production output is encoded as contribution feed (HEVC ASI / IP) for delivery to host broadcast facility.

(5) VR Live

- Local audio input for production mix
- Individual VR Camera signals demultiplexed, processed in “CCU” to correct and rendered to SDI / IP output to vision mixer
- Wireless Camera signals received, decoded and processed in CCU. Rendering to flat rectilinear image.
- SDI / IP output to vision mixer. Local audio extracted for production mix.

(6a) VR Live

- Production output is encoded as contribution feed (HEVC ASI / IP) for delivery to host broadcast facility
- For IP based delivery to consumer
- Discrete Spatialized Audio Mixes (per camera view) may be added

(6b) VR Live to Mobile

- Production output is encoded as contribution feed (HEVC ASI / IP) for delivery to host broadcast facility
- For Mobile delivery to consumer
- May have differing visual needs (graphics size, cuts etc.)
- Could be identical to main production O/P

(7) For Linear / VR simulcast, time synchronization / encoder latency / PTS must be maintained to allow user to switch between linear and VR application with minimal interruption to program timing

(8) It would be feasible to implement remote production for VR Live

- Camera feeds from a small Head-end Location Unit would be fed ASI/IP to a remote Gallery for Processing
- OB Parking space is generally at a premium
- Hardware would solely comprise Fibre Camera interface and ASI interface to link to Broadcast Centre

TV Linear and VR Live

- All sources are co timed to production mixer to ensure all sources are synchronized and can be cut together

Figure 4 describes the workflows in details for Host Facility HQ, Host Remote Production Facility HQ and Consumer Environment.

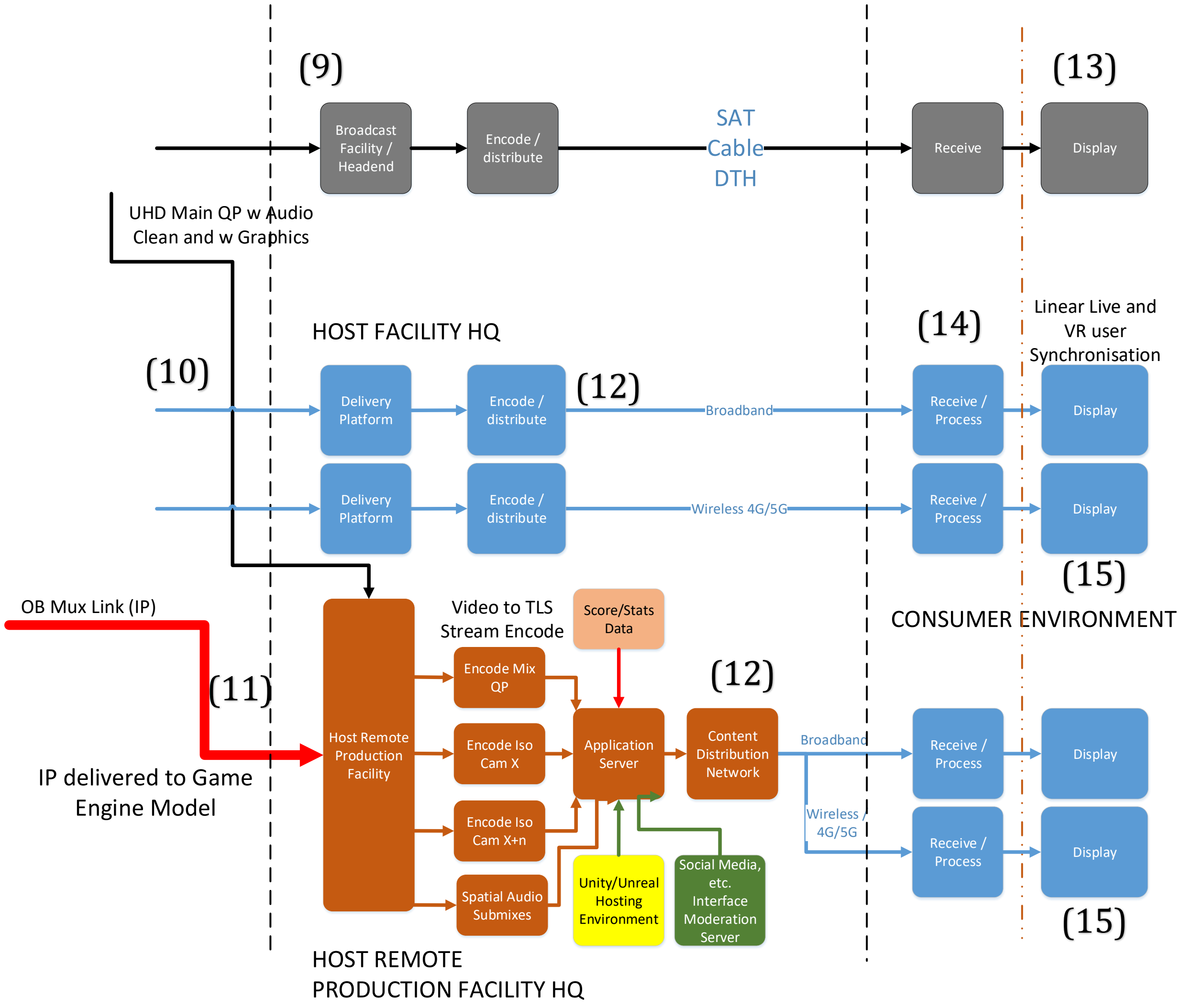

Figure 4: Host Facility HQ / Remote Production Facility HQ and Consumer Environment

Host Facility HQ

(9) TV Linear

- Live contribution feed decoded to SDI / IP 2110 for host broadcast processing
- Local audio voiceover and graphics added as well as breaks and links

(10) VR Live

- Live contribution feed decoded to SDI / IP 2110 for host delivery platform processing
- Programme toped and tailed for consumer package

Host Remote Production Facility HQ

(11) VR Live Remote Production

- Live contribution feeds (IP) for platform processing

- Hosted in Game Engine environment
- Encode Camera feeds for delivery
- Spatial Audio submixes (per Camera View)
- Picture feed optimization for best Game Engine performance

Host Facility HQ & Host Remote Production Facility HQ

(12) VR Live and VR Live Remote Production

- Content is prepared for delivery
- It may use Viewport Independent Delivery (VPI)[4]
    - May or may not use CDN, possible to broadcast
- It may use Viewport Dependent Delivery (VPD)[5]
    - CDN is required for best performance

Consumer Environment

(13) TV Linear

- Content received, demultiplexed, decoded and rendered to display format.
- Local navigation and programme selection
- Local consumer display device - TV

(14) For Linear / VR simulcast, time synchronization must be maintained to allow user to switch between linear and VR application with minimal interruption to programme timing

(15) VR Live

- Content received, processed in receive device and output to display format.
- Local navigation and content selection
- Integrated receive / consumer display device HMD / Tablet

Potential of SMPTE 2110 in VR workflows

Early workflows for VR360 content made use of customized applications of devices that were repurposed from regular 2D production. This included the use of SDI (Serial Digital Interface) for transporting uncompressed video from camera rigs to processing devices and between processing devices. The Nokia OZO camera, with its 8 wide angle lenses and 8 microphones, made use of a single link 1.5G HD-SDI interface to deliver proprietary compressed sensor data to a powerful computer running the OZO Live suite which performed the stitching, rectification and other content processing and formatting functions. Along with several GPUs, the computer

[4] VPI: Viewport Independent: the entire omnidirectional video is delivered. In order to reduce the bandwidth the video may be downscaled to a lower resolution.

[5] VPD: Viewport dependent: only a portion of the omnidirectional 360 video corresponding to the viewport is delivered in high quality to reduce bandwidth and complexity of decoding at the receiver (e.g. instead of decoding 8K for full 360, VPD can achieve the same quality with lower decoding capabilities in the device).

required dedicated PCIe SDI boards. While SDI provides a suitable interface for “close” connections, inherently fixed nature lacks flexibility and dynamic configurability.
The general ambition for SMPTE ST 2110 is as a replacement of coaxial SDI interfaces in the production suite, including the venue or studio and any outside broadcast facilities.
As we look towards the role that ST 2110 can play in the VR domain, the follow usage scenarios are feasible [2].

Omnidirectional 360/180 camera with inbuilt stitching

A single omnidirectional camera rig with either full or partial spherical coverage looks to use ST 2110 for delivering the panoramic scene via IP to additional processing systems. The camera rig includes built in stitching software which spatially composites each sensors bit stream and outputs in “raw” format. The camera rig also includes multiple microphones to capture the “sound field” which is processed by an inbuilt processor and output as a bit stream that may or may not be temporally aligned with the stitched video output.

Omnidirectional 360/180 camera without inbuilt stitching

A single omnidirectional camera rig with either full or partial spherical coverage looks to use ST 2110 for delivering the captured scene via IP to additional processing systems. The camera rig outputs each sensors bit stream as a separate “track” in an uncompressed format. Any microphone sources are also output as separate tracks. In order to reconstruct an omnidirectional visual scene and its corresponding sound field, orientation information (often referred to as intrinsic/extrinsic metadata) should be provided for each track.

Multiple Omnidirectional 360/180 cameras with/without inbuilt stitching

At a live event, multiple omnidirectional camera rigs are deployed at strategic locations. These cameras may or may not include inbuilt processing of the video and audio data (according to the “Omnidirectional 360 camera with inbuilt stitching” and “Omnidirectional 360 camera without inbuilt stitching” scenarios presented above). All video and audio bit streams are sent over an IP contribution link to a cloud based processing system which creates the necessary composite view. In order to allow the positions of each camera rig to be shown within a single omnidirectional view, accurate positioning information (either global or relative) is required.

Personal Point-of-View

At a sporting or artistic performance, several fixed wide angle IP connected cameras are positioned at various locations (for example, in the goal mouth, in front of the band) and the content is uplinked to a cloud based processing center where a virtual PTZ-style view can be created for each viewer based on their pose along with the ability to hop between camera positions.

Follow my favorite

This scenario uses fixed wide angle cameras such as that described in “Personal Point-of-View” but the cloud based processing system performs object/actor/player recognition across all the IP uplinked camera feeds and creates a view tailored to the desires of the viewer.

Volumetric Acquisition

Volumetric capture systems use large numbers of fixed, calibrated inward facing cameras to record still or motion objects in real time. The feed from each camera needs to be uploaded to a local or cloud service such that processing, including the final production of a fixed or dynamic 3D model (using point clouds or texture-and-mesh) can be developed.

Recent Advances on Live VR Services in MPEG

OMAF 2nd Edition

The 1st edition of MPEG’s Omnidirectional Media Format (OMAF) specification in ISO/IEC 23090-2 [1] finalized in October 2017 defines a media format that enables omnidirectional media applications, focusing on 360° video, images, and audio, as well as associated timed text and has been the basis of the profiles addressed in VRIF Guidelines 1.0 and 2.0 [2]. These OMAF-based profiles, composed of a viewport-independent and a viewport-dependent profile, are equally applicable for both on-demand and live 360 VR content distribution.

MPEG will release the 2nd edition of their OMAF specification in October 2020 with additional features, such as multiple viewpoints, overlays and new tiling profiles [1]. The architecture for OMAF 2nd edition is depicted in Figure 5. Toward enabling richer live VR360 experiences, VRIF is working to develop recommendations around OMAF 2nd edition as part of its Guidelines 3.0 activity.

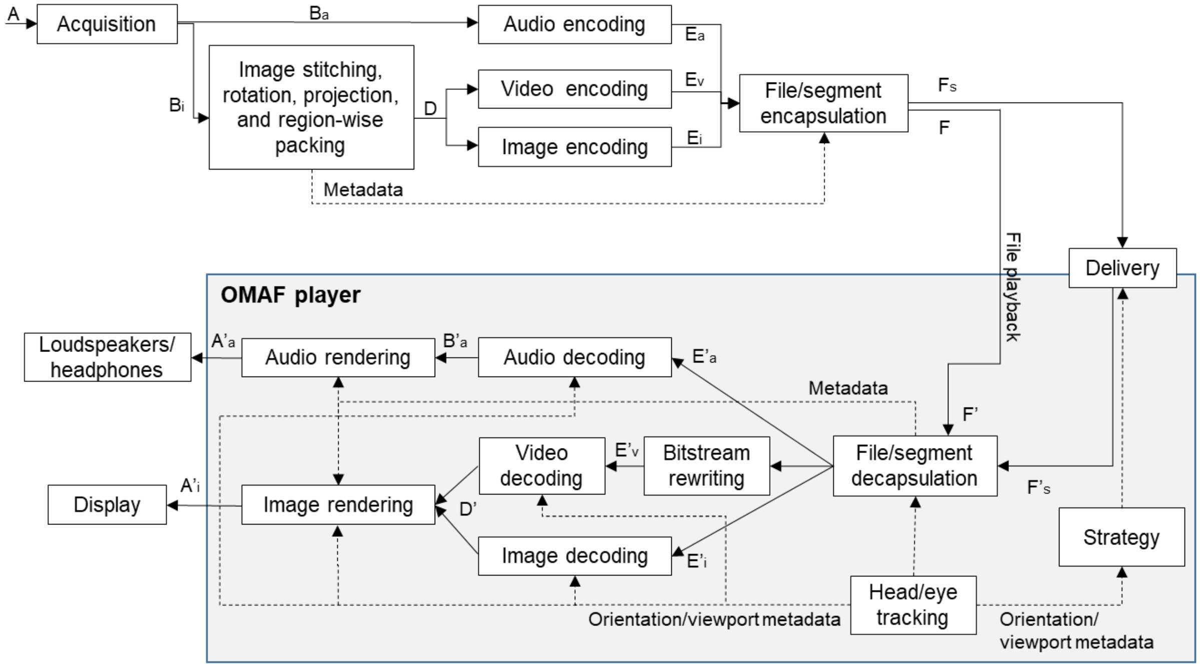

Figure 5: OMAF 2nd edition architecture

In the sub-sections below, we provide a summary of the OMAF 2nd edition features.

Multiple Viewpoints

Multiple viewpoints can be thought as a set of 360º cameras which, for example, may be scattered around a basketball field (see Figure 6). The OMAF 2nd edition specification enables a streaming format with multiple viewpoints to allow, for example, switching from one viewpoint to another, as done by multi-camera directors for traditional video productions. This allows watching an event or object of interest from a different location and facilitates leveraging the well-established cinematic rules for multi-camera directors that make use of different shot types, such as wide-angles, mid-shots, close-ups, etc.

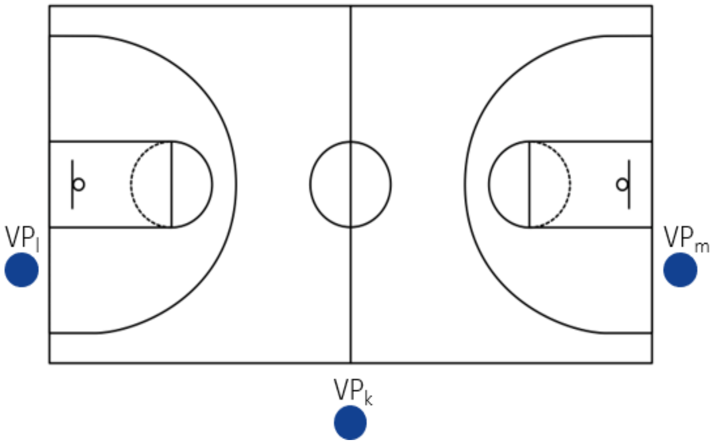

Figure 6: Example usage of multiple viewpoints in a basketball event.

Figure 6 illustrates an example of OMAF 2nd edition content with multiple viewpoints (VP_k, VP_l and VP_m). This allows the user to experience the action from different perspectives and facilitate interactive content consumption and creative storytelling [1].

OMAF 2nd edition [1] provides means to signal multiple viewpoints including

- Cartesian coordinate (X, Y, Z) position of the viewpoint,
- GPS position of the viewpoint,
- Geomagnetic position information for the viewpoint, i.e., orientation
- Yaw, pitch, and roll rotation angles of X, Y, and Z axes, respectively, of the global coordinate system of the viewpoint relative to common reference coordinate system,
- Viewpoint group information,
- Viewpoint switching information,
- Viewpoint looping information.

Viewpoint data can be static or dynamic. In case of dynamic viewpoints, OMAF defines a new timed metadata track to signal the changing viewpoint data.

Overlays

Overlays are a way to enhance the information content of 360 video. They allow superimposing another piece of content (e.g., a picture, another video with news, advertisements, text or other) to be rendered on top of the main (background) omnidirectional video. Overlays also allow the creation of interactivity points or areas.

Example usage scenarios for overlays include:

- Annotations of 360 video content, e.g., player statistics in sports games, etc.
- Advertisements
- Selectable hotspots for user interactions, e.g. switching viewpoints, turning overlay on/off
- Closeups or 2D camera views to complement 360 background video
- Recommended viewport for the content, e.g., director’s cut

- Logo and trademark display

OMAF 2nd edition defines four different overlay types, based on their spatial position (see Figure 7, [3], [4], [5]:

- Overlays may be positioned on the users' viewing screen and always present on the users' viewport (viewport-relative or viewport-locked overlay);
- Overlays could be positioned at a depth from the user viewing position (sphere-relative 2D overlay);
- Overlays may be positioned over the background video without any gap between the two (sphere-relative omnidirectional overlay);
- 3D mesh at a given location within the unit sphere (sphere-relative 3D mesh overlay).

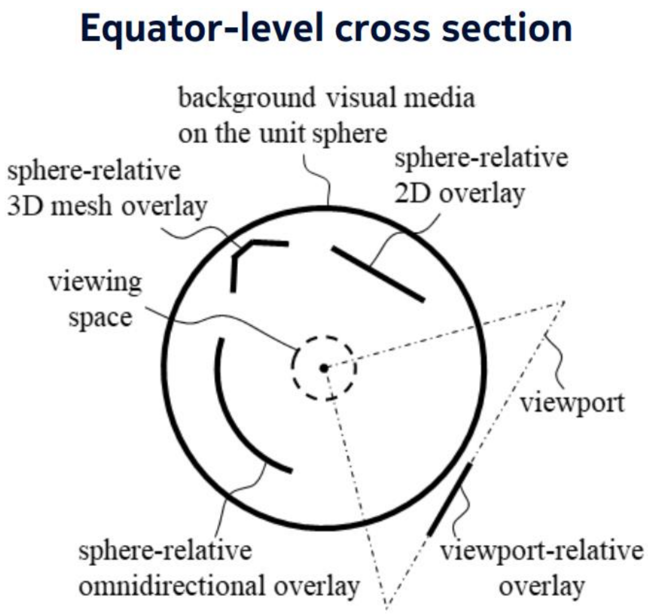
Equator-level cross section
background visual media on the unit sphere
sphere-relative 3D mesh overlay
sphere-relative 2D overlay
viewing space
viewport
sphere-relative omnidirectional overlay
viewport-relative overlay

Figure 7: Different types of overlays defined in OMAF 2nd edition [4].

As controls for user interaction with the overlays, OMAF 2nd edition defines signalling of the following flags:

- change_position_flag, when set to 1, specifies that users are allowed to move the overlay window to any location on the viewing sphere or the viewport.
- change_depth_flag, when set to 1, specifies that the depth of overlay can be chosen by user interaction. When both change_position_flag and change_depth_flag are set to 1 then the X,Y,Z position of the overlay can be freely choosen by user interaction.

- switch_on_off_flag, when set to 1, specifies that the user is allowed to switch ON/OFF the overlay.
- change_opacity_flag, when set to 1, specifies that the user is allowed to change the opacity of the overlay.
- resize_flag, when set to 1, specifies that the user is allowed to resize the overlay window. The field-of-view of the resized overlay window shall be same as that of original overlay window.
- rotation_flag, when set to 1, specifies that the user is allowed to rotate the overlay window to different directions. The field-of-view of the rotated overlay window shall be same as that of original overlay window.
- change_position_flag, change_depth_flag, switch_on_off_flag, change_opacity_flag, resize_flag, or rotation_flag when set to 0, specifies that the user is disallowed to perform the respective operation on the overlay.

Advanced Tiling Profile

Viewport-dependent delivery of OMAF content can be achieved by either of the following:

- Multiple versions of the content, where different versions are optimized for different viewing orientations, e.g., via region-wise packing which enables encoding different portions of the 360 video content at different quality levels.
- Splitting the content into tile sequences, where a tile sequence is defined as a rectangular subset of the original video content, and encoded tile sequences have the capability to be merged with other encoded tile sequences in coded domain without decoding mismatch by rewriting only header data.

Each tile sequence can be encoded at multiple resolutions and/or at different quality levels. Viewport-dependent delivery can be achieved by receiving different sets of available encoded tile sequences.

In OMAF 1st edition, tile-based viewport-dependent delivery relied upon the use of extractor tracks. Each extractor track is targeted at a particular range of viewing orientations and providing bitstream rewriting instructions to merge specific tile sequences into a single video bitstream. The content author must prepare multiple extractor tracks, and then the strategy module on the client side selects between available extractor tracks, and the bitstream rewriting module follows the instructions of the selected extractor track to create a video bitstream.

In OMAF 2nd edition, a new viewport-dependent delivery solution, a.k.a. Advanced Tiling Profile was adopted, that leaves complete freedom to the party that prepares the content to make encoding decisions and provides freedom to the player vendor to implement strategies for adapting to conditions such as available decoding resources and dynamically changing bandwidth. The strategy module selects any tile sequences that can be merged to a single HEVC-compliant bitstream. The tile selection strategy can be tailored according to the current viewport. Tile selection and bitstream rewriting entirely player's responsibility. As such, this profile leverages the bitstream rewriting capabilities of the client

Volumetric Content and Point Clouds

Volumetric video has been recently gaining significant traction in delivering live VR experiences. Volumetric video contains spatial data and enables viewers to walk around and interact with people and objects, and hence it is far more immersive than 360 video footage because it captures the movements of real people in three dimensions. Users can view these movements from any angle by using positional tracking. Point clouds are a volumetric representation for describing 3D objects or scenes. A point cloud comprises a set of unordered data points in a 3D space, each of which is specified by its spatial (x, y, z) position possibly along with other associated attributes, e.g., RGB color, surface normal, and reflectance. This is essentially the 3D equivalent of well-known pixels for representing 2D videos. These data points collectively describe the 3D geometry and texture of the scene or object. Such a volumetric representation leads to immersive forms of interaction and presentation with 6 degrees of freedom.

As part of its “Coded Representation of Immersive Media” (MPEG-I) project, MPEG is currently developing the “Visual Volumetric Video-based Coding (V3C) and Video-based Point Cloud Compression (V-PCC)” standard in ISO/IEC 23090-5 to compress point clouds using any existing or future 2D video codec, including legacy video codecs, e.g., HEVC, AV1, etc. [6]. An associated carriage format for V3C data is also developed by MPEG in specification ISO/IEC 23090-10 [7]. These standards are expected to be finalized and published before the end of this year. As part of its Guidelines 3.0 development, VRIF is currently considering these standards toward defining industry profiles for volumetric content compression and distribution.

Industry deployment information on Live VR services

TiledMedia

TiledMedia’s 360 Live distribution system demonstrated at IBC 2019 and depicted in Figure 8 gives consumers an 8K experience while transmitting and decoding much less information (4K decoding using View Port Dependent tecnique); the platform sends only what a user actually sees. To make this possible, the image is cut up into around 100 high-quality tiles and only the actual tiles in view are sent, decoded and displayed. The real time nature of the experience, including the need to instantly respond to a user’s head motion, makes this approach particularly challenging. When the user turns their head, the system retrieves new high resolution tiles, decodes and displays them – all within a tenth of a second. This happens so fast that users will hardly notice the low resolution background layer that is always present to prevent black areas from appearing in their field of view. The technical specification of the TiledMedia system is depicted in *Figure 9* [2].

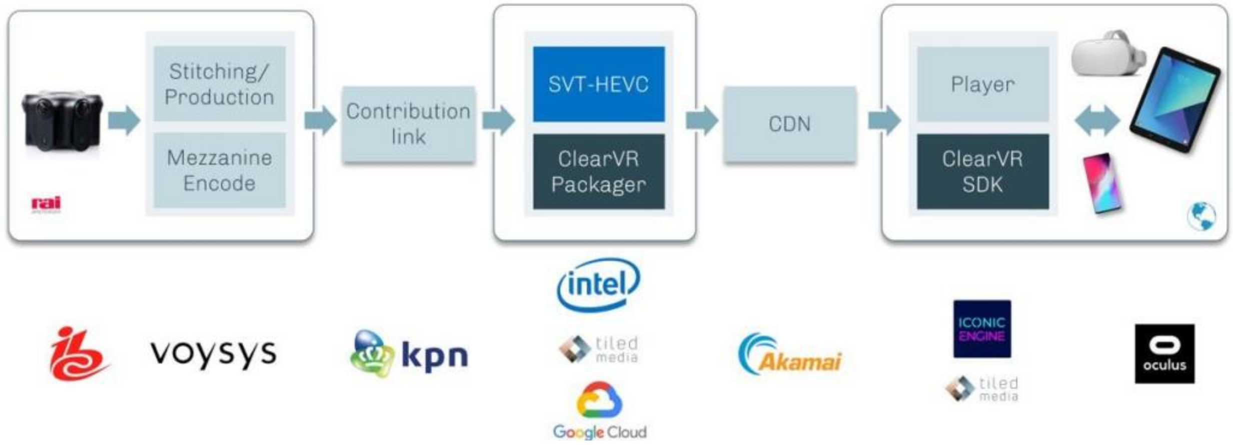

Figure 8: TiledMedia 360 live distribution system [2].

Number of Cameras	*Two, user-switched*	**Distribution Format**	*MP4-based ClearVR packaging*
Event Duration	*5 Days; 8+ hours per day*	**Distribution Bitrate**	*12 – 15 Mbit/s*
Capture Resolution	*8 092 x 4 046*	**Streaming Protocol**	*Standard http/2 with multipart byterange requests*
Contribution Bitrate	*150 Mbit/s per camera (HEVC)*	**User device decoder**	*HEVC Main Level 5.1*
Distribution Resolution	*8 192 x 4 096*	**Glass-to-glass latency**	*~30 seconds*
Distribution Encoder	*SVT- HEVC*	**Supported Devices**	*Oculus headsets, iOS and Android tablets and phones*
Cloud Processing	*Well over 1000 Intel cores in Google Cloud Platform for the two streams (dynamically managed)*		

Figure 9: Technical specs for TiledMedia 360 live distribution system [2]

Intel Sports

Intel's immersive media platform provides partners an end-to-end solution for processing immersive media content and delivering immersive media experiences to fans worldwide. During the recording of a sporting event, the capture systems, Intel® True View and Intel® True VR work independently or side-by-side at the stadium. Intel® True VR is stereoscopic camera solution (with multiple cameras placed in specific locations in the stadium) and outputs panoramic video from the stadium and sends it through the immersive media processing pipeline. Intel® True View is comprised of a camera array that is built into the perimeter of the stadium and outputs volumetric video through the same immersive media processing pipeline as Intel® True VR, as depicted in Figure 10. Once the content from either capture system is processed, the distribution pipeline enables delivering a catered stream for each supported device. The Intel Sports immersive media platform supports distributing both live and non-live content. The immersive media platform also supports 2D screen viewing experiences like mobile phones, traditional broadcast television, and web players as well as VR/AR/MR Head Mounted Devices (HMDs) and other immersive media streaming platforms.

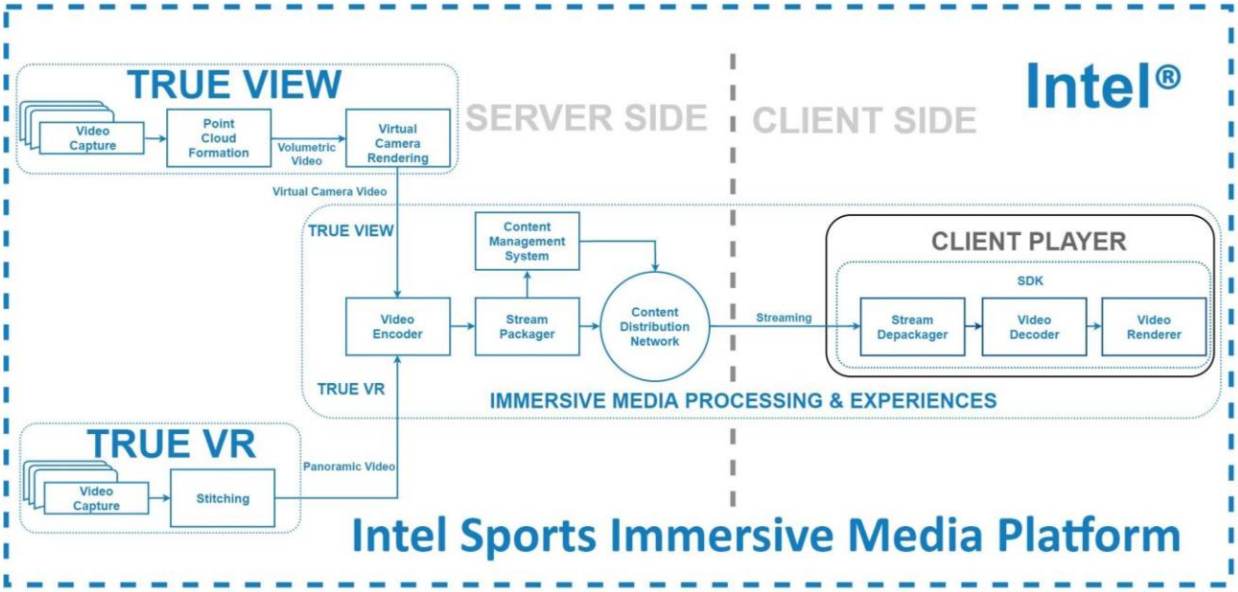

Figure 10: Intel Sports Immersive Media Platform [2].

8K Viewport Independent VR Content Distribution

Concept

VR has evolved from initially being 4K viewport independent to 4K viewport dependent to double the resolution per eye in the viewport and providing a better quality of experience. With the advent of a new 8K-capable devices — including Qualcomm XR2[6] and next-generation 5G smartphones released in 2020 — it is now possible to send and decode the full 8K sphere in the device and provide the same resolution in the viewport as the best 4K viewport-dependent solutions. The scheme has two main advantages. First, it uses DASH [8] existing infrastructure, and second it is quite straightforward to implement on the encoder, streaming and player sides. The proposed viewport-independent approach fits quite well on new 5G networks, where network capacity is estimated to reach 10 times the amount of 4G, enabling more bits to be transmitted to 8K-capable 5G devices, without any lag compared unlike a viewport-dependent system. Of course, more storage and bandwidth are required on the network, but we expect only a 50% increase versus 4K viewport dependent, when using state-of-the-art, cloud-based content-aware encoding (CAE) technology featuring HEVC compression.

[6] https://www.qualcomm.com/products/snapdragon-xr2-5g-platform

Table 1 shows the different capabilities of each technology.

Features	**4K viewport independent**	**4K viewport dependent**	**8K viewport independent**
Standard based	OMAF 1.0	OMAF v1.0/OMAF v2.0	Extension of OMAF 2.0
Quality	No lag in any network conditions Limited (HD) perceived resolution	Lag in stressed NW conditions Limited ABR support	No lag in any NW conditions Reduced resolution on 4G
Decoder performance	4K decoder required	4K+ decoder required	8K decoder required
Decoder base	All head-mounted displays (HMDs)	All high-performance 4K HMDs/phones	Gear VR/S10 Skyworth v901 XR2 S10/S20 phones
Protocol	DASH	DASH with tiling	DASH
Latency	5-7s (1)	20-30s	5-7s (1)
DRM integration	Easy	Complex	Easy
Multi-client integration	Easy	Complex	Easy
Bitrate top quality	10-15 Mbps	20 Mbps	35 Mbps
CDN independent	Yes	No (2)	Yes
Broadcast support	Yes	No	Yes

Table 1: Technology comparison. (1) When DASH CMAF low latency is used (2) Requires CDN optimization for best performance

VRIF Adoption

VR-Industry Forum (VRIF) has developed a new distribution profile for viewport-independent delivery of 8K video content in its Guidelines version 2.2 [9] that will match the same resolution per eye as achieved by the 4K viewport-dependent OMAF profile today. The 8k viewport-independent profile inherits all properties of the OMAF viewport-independent profile, but requires support for HEVC Main 10 Level 6.1 decoding capability in order to process 8k content.VRIF believes that this is a pragmatic means to increase the adoption of high-quality live VR experiences by leveraging 8K processing capabilities of new 5G mobile devices, including phones and HMDs, expected to be available by end of 2020.

Support in MPEG OMAF 2.0

The current version of OMAF 1.0/2.0 limits the viewport-independent and viewport-dependent profiles to HEVC Main Level 5.1 (4Kp60). VRIF is proposing that MPEG extend the HEVC profile to at least 8Kp60 in a future version of OMAF 2.0, with support for Level 6.1. This proposal has been submitted to MPEG and is expected to be included as part of MPEG OMAF 2.0 specification.

DASH

DASH is now becoming universal for streaming services, and we believe that in order to offer a compelling VR service it has to be based on DASH. The DASH foundation offers many benefits and is a key feature to offer state-of-the-art services similar to what was demonstrated at the 2019 French Open, when 8K broadcast content was streamed using DASH to address mobile devices and connected TVs [10].

In the past, the viewport dependent profiles have defined a specific streaming format based on tiling using either OMAF 1.0 or 2.0 (as discussed in "OMAF 2nd Edition" section above) but people realized this prevented them from using already-deployed DASH-based OTT solutions and their associated benefits – this is due to the required extensions to the legacy DASH deployments in order to support the tile-based profiles of OMAF. The legacy DASH framework enables low latency, seamless DRM integration with CENC, and support for digital ad insertion and content replacement.

Encoder

Using CAE, content providers can deliver VR at the lowest bitrate and at the highest quality. The fundamentals of CAE, as implemented by Harmonic with its AI-based EyeQ technology, has been described in detail [11]. The Ultra HD Forum is recommending using CAE for 4K encoding and, of course, the same technique can also be used for 8K. Some early results of 8K encoding using CAE for streaming applications have been reported [12].

Table 2 describes the encoding profiles being used. As CAE is used, an encoding cap is defined as well as the average encoded rate, which depends entirely on the complexity of the content.

Profiles	**Cap (Mbps)**	**Resolution**	**Frame rate (fps)**	**HEVC encoded bitrate (Mbps)**
8K	42 Mbps	7680x4320	30 fps	30-35 Mbps
4K high quality (HQ)	25 Mbps	3840x2160	30 fps	15-18 Mbps
4K low quality (LQ)	15 Mbps	3840x2160	30 fps	10-12 Mbps
1080p HQ	8 Mbps	1920x1080	30 fps	3.5-5 Mbps
1080p LQ	5 Mbps	1920x1080	30 fps	2.5-3.5 Mbps
720p	3 Mbps	1280x720	30 fps	1.5-2.5 Mbps

Table 2: 8K encoding profiles

Decoder Capabilities

The following devices have been tested using 8Kp30 content monoscopic, encoded with DASH ABR. Table 3presents the progress on interoperability.

Device	Video format	Streaming format	Player	Mode
Galaxy S10	8Kp30	DASH	Viaccess-Orca	Magic window
Gear VR/Galaxy S10	8Kp30	DASH	Viaccess-Orca	HMD
Skyworth v901	8Kp30	DASH	Native player	HMD
Qualcomm XR2	8Kp60	DASH	Viaccess-Orca	HMD
S20	8Kp60	DASH	Viaccess-Orca	Magic window

Table 3: Interoperability table

We expect to test more HMDs in 2020 based on the XR2 platform.

Backward Compatibility

Backward compatibility with 4K-only devices, such as the Oculus Quest, is ensured by using 4K profiles and below. Moreover, when delivered through a limited capacity network like 4K, even if the device is 8K capable, it might have to fall back on 4K and below profiles.

Figure 11 describes the match between content and devices in the different scenarios.

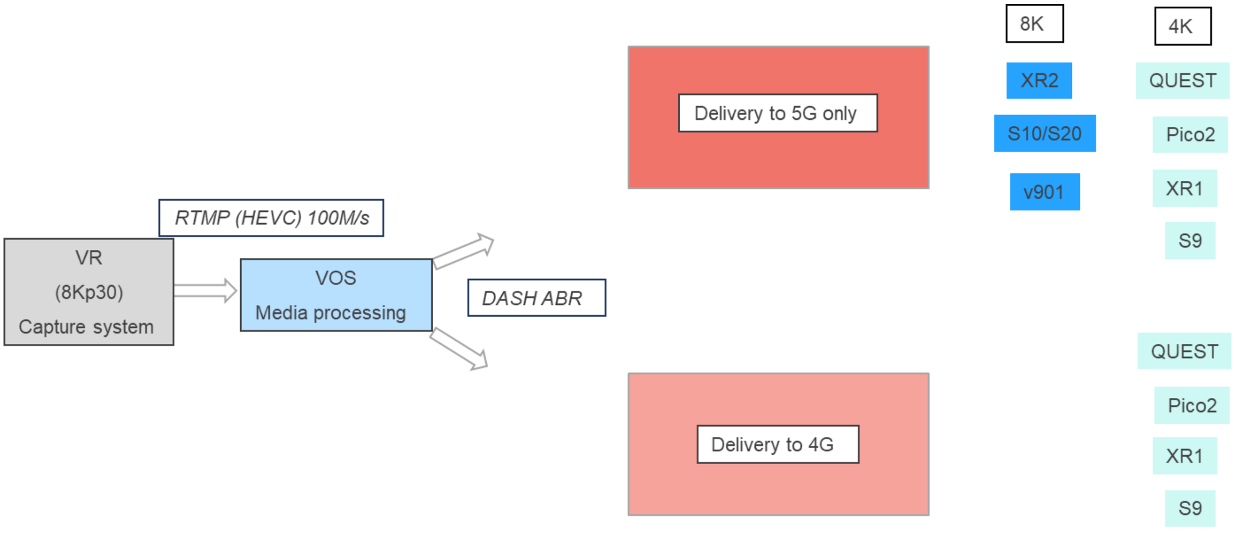

Figure 11: Different scenarios for 8K VR ABR delivery

Bitrates

Table 4: Formula 3 8K encoding profilesshows the VR IF test clips of Formula 3 encoded with an 8K profile ladder, highlighting the different bitrates.

Profiles	**Resolution**	**Frame rate (fps)**	**Average video bitrate (Mbps)**	**Max video bitrate measured over chunk duration (Mbps)**
8K Cap 42 Mbps	7680x4320	25 fps	28.6-32 Mbps	42.6 Mbps
4K Cap 25 Mbps	3840x2160	25 fps	16.6-18.3 Mbps	26 Mbps
4K Cap 15 Mbps	3840x2160	25 fps	10.2-12.1 Mbps	15.5 Mbps
1080p Cap 8 Mbps	1920x1080	25 fps	3.6-5 Mbps	6.3 Mbps
1080p Cap 5 Mbps	1920x1081	25 fps	2.2-3.6 Mbps	4.9 Mbps
720p Cap 3 Mbps	1280x720	25 fps	1.3-2.4 Mbps	2.9 Mbps

Table 4: Formula 3 8K encoding profiles

The 8K profile, compared with a high-quality 4K profile (with a cap of 25 Mbps), is only 74% higher. As CAE compression is used, we also indicate the top bitrate measured during a chunk duration, and the 8K bitrate tops out at 42.6 Mbps versus 26 Mbps for 4K, a 62% increase.

It is important to note that these data are provided for live encoding with a prototype encoder, meaning that they are not optimized yet. Over time it's expected that 8K will not be more than 50% higher than 4K viewport dependent.

Low Latency

One important VR use case is to be able to consume VR content on a mobile phone in a stadium. There are two major applications for this use case: live and replay. For live, the user needs to be able to follow the live action with his phone. The delay between the live action and what appears on the phone must be minimal. Using state-of-the-art DASH CMAF low-latency chunk transfer with local edge processing in the stadium, experiments [13] have shown a five- to seven-second delay is possible. Lower delay can be achieved by reducing encoding delay, but this will impact the bitrate, which might not be desirable for a mobile operator who wants to serve tens of thousands of users simultaneously using a 5G network. For the replay, the user wants to go back in time to replay after an important event has occurred. Thus, in this instance the delay is also critical.

Broadcast Mode

3GPP is defining the use of broadcast technology for 5G. The initial name for 4G was eMBMS, now called FeMBMS[7] for 5G applications. Today, unicast is not the right technology to scale 100,000 simultaneous sessions during big events like the Olympics or the FIFA World Cup where users want to be able to stream different games that are broadcast live over TV networks via 5G for home viewing. In that case, a viewport-independent scheme is the right technology to use, as one feed is sent to all users. The EBU has provided requirements for this mobile

[7] See 5G field measurements for broadcast: https://lab.irt.de/5g-today-field-measurements/

broadcast technique and also explains the different standardization activities taking place inside 3GPP [14].

Conclusions

This paper described Live VR workflows from an end-to-end perspective such as production, contribution, distribution and consumption aspects and describe how this can be complementary to the traditional broadcast services with a 4K like experience. We also summarized the recent advances in MPEG standards to improve and enrich live VR experiences with 360 degree video and volumetric content, and presented a few example Live VR service deployments in the industry. On the technology side, we showed that 8K viewport independent, using DASH technology for streaming and FeMBMS for broadcast, is an attractive approach when used over 5G and also using modern devices that have an 8K decoding capability. Backward compatibility with legacy 4K devices is achieved by the virtue of ABR. The DASH environment offers continuity with existing OTT platforms that are already using DASH and is crucial to accelerate 5G deployments as well as address other devices at home. The proposed viewport-independent scheme can, of course, also be used at home via high-speed broadband networks like fiber, DOCSIS 3.1 or 5G fixed wireless where HMD and magic window modes can both be used.

Acknowledgements

The following individuals are acknowledged for their contributions to VRIF on the above content during the course of VRIF Guidelines development: Xavier Ducluox (Harmonic), Vincent Lepec (Viaccess Orca), Chris Johns (Sky), Rob Koenen (TiledMedia), Angus Yeung (Intel), Kim Loza (Intel), Miska Hannuksela (Nokia) and Igor Curcio (Nokia).

References

[1] ISO/IEC 23090-2, "Information technology — Coded representation of immersive media — Part 2: Omnidirectional media format (OMAF)".

[2] Virtual Reality Industry Forum (VRIF), "VRIF Guidelines," available at: https://www.vr-if.org/guidelines.

[3] I. Curcio, K. Kammachi-Sreedhar and S. Mate, "Multi-Viewpoint and Overlays in the MPEG OMAF Standard," *ITU Journal: ICT Discoveries, Vol. 3(1),* 18 May 2020.

[4] M. M. Hannuksela and Y.-K. Wang, "An overview of Omnidirectional MediA Format (OMAF)," *submitted to Proceedings of the IEEE.*

[5] K. Kammachi-Sreedhar, I. Curcio, A. Hourunranta and M. Lepistö, "Immersive Media Experience with MPEG OMAF Multi-Viewpoints and Overlays,," *ACM Multimedia Systems Conference,* 8-11 June 2020, Istanbul, Turkey.

[6] ISO/IEC 23090-5, "Visual Volumetric Video-based Coding (V3C) and Video-based Point Cloud Compression (V-PCC)".

[7] ISO/IEC 23090-10, "Carriage of Visual Volumetric Video-Based Coding Data".

[8] "DASH Industry Forum Guidelines," https://dashif.org/guidelines/.

[9] Virtual Reality Industry Forum (VRIF), "VRIF Guidelines 2.2," June 2020, https://www.vr-if.org/wp-content/uploads/VRIF_Guidelines-2.2.pdf.

[10] T. Fautier. and E. Mazieres, "8K Experiment at Roland Garros 2019," *France TV Labs,* May 2019, https://www.francetelevisions.fr/lab/projets/8K-Experiment-at-Roland-Garros-2019.

[11] Harmonic, "EyeQ: Achieving Superior Viewing Experience," March 2016, https://info.harmonicinc.com/technical-guide/achieving-superior-viewing-experience/.

[12] T. Fautier, Harmonic, "8K Is Making Progress Bit by Bit," November 2019 https://www.harmonicinc.com/insights/blog/8k-making-progress.

[13] Harmonic, "DASH CMAF LLC to Play Pivotal Role in Enabling Low Latency Video Streaming," December 2018, https://info.harmonicinc.com/white-paper/dash-cmaf-role-in-enabling-low-latency-video-streaming/.

[14] EBU, "5G for the Distribution of Audiovisual Media Content and Services," *Tech Report 054,* June, 2020. https://tech.ebu.ch/publications/tr054.