VR VIDEO ECOSYSTEM FOR LIVE DISTRIBUTION

T. Fautier

President of the Ultra HD Forum and Vice President, Video Strategy, at Harmonic, USA

ABSTRACT

New VR (Virtual Reality) HMDs (Head Mounted Displays) being introduced in 2016 are creating increased demand for VR video content. A growing amount of content — including documentaries, movies and live events — have already been covered in VR video. For the market to truly take off, some standardization is required with regards to the rules of content writing, the content acquisition and stitching methods, and the approach for mapping content for encoding and delivery. In addition, the industry needs to define a unified mechanism to address all of the different ecosystems, ranging from the various VR devices to mobile devices, STBs and connected TVs, to avoid the fragmentation that resulted with 3D and over-the-top (OTT) video delivery. This paper will present reference architectures that can be deployed with existing technology to pave the way for future evolutions of VR.

INTRODUCTION

VR represents an entirely new way for consumers to experience video. No longer is the TV viewer or game player a passive participant in the action; VR video simulates the experience of entering the video content itself, with the ability to see a full 360 degrees in any direction. The entertainment and educational possibilities afforded by the technology are “virtually” unlimited, and stand to change the way that video is produced, prepared and consumed for generations to come.

VR video can be thought of as a panoramic representation of content either captured on camera or generated via computer graphics, and then viewed on a 2D or 3D HMD. The workflow to create and deliver content includes production, encoding and transmission of audio, video and graphic elements. The displays worn by VR video consumers are a key part of the VR ecosystem and come in a variety of form factors. They may be either tethered (e.g., Oculus Rift, HTC Vive, Sony PlayStation VR) or untethered (i.e., connected to a device wirelessly, such as Gear VR and LG VR) to a VR player or PC. They can also be fully self-contained 2D devices, such as Google Cardboard, in which the user views content from a smartphone.

This paper will focus on VR video content preparation (i.e., acquisition, processing, encoding, transmission). Audio, graphics and devices are an entire other subject.

VR VIDEO DEFINITION

VR, sometimes referred to as immersive multimedia, is a computer-simulated environment that can mimic physical presence in places in the real world or imagined worlds. Virtual reality can recreate sensory experiences, virtual taste, sight, smell, sound, and touch, which include virtual VR video is a panoramic (180 or 360 degrees) video environment that is captured on a single or stitched multi-camera system and sent to a wireless HMD for an immersive experience or to a 2D device such as a PC, mobile device or TV set. Content can be consumed locally, streamed or broadcast.

VR VIDEO CONTENT CREATION

VR video is a complete ecosystem that is still under construction. This section will provide a high level overview of VR video content creation. VR video content creation is composed of different steps. Some of them can be skipped depending on the solution, as described in Figure 1. Content is captured via a camera rig, then stitched, then mapped to a variety of different geometries, then encoded in HEVC and transmitted using either broadcast or unicast mechanism.

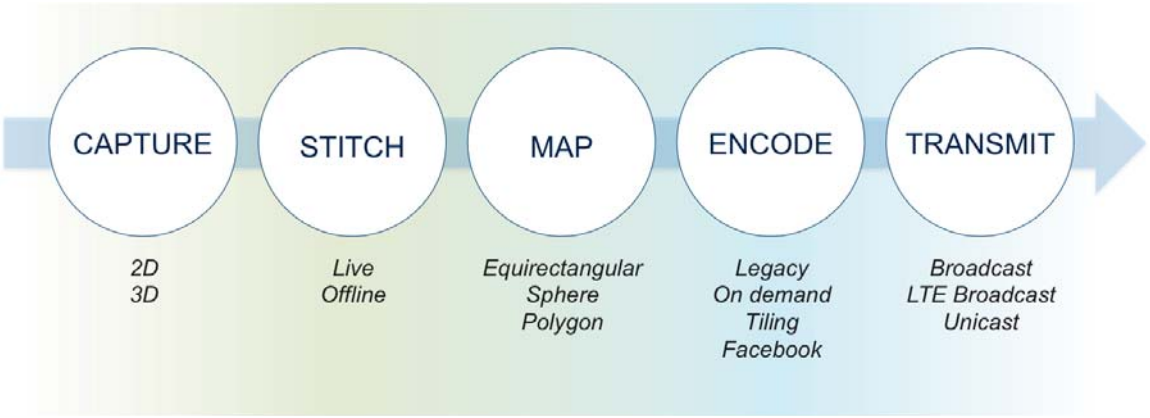

Figure 1 - VR video content creation

1. Capture

The capture system can be segmented in three categories: camera orientation, panoramic view and 3D relief.

Camera orientation

Capture is made of different types of synchronized cameras with multiple lenses. For natural video, there are different camera systems:

- **Concave view:** Multi-lens cameras, shot from one camera into multiple directions like what's being done currently by most VR camera companies.
- **Convex view:** Multiple camera system spread across the arena, all filming the same subject.

Most of the deployments are using a concave system, but the convex system is also used today with still pictures. We expect this to become more deployed for natural video in the future.

Prosumer cameras capture at p30, while pro cameras capture at p60.

Panorama view

VR video today is declined either in 180 degrees or 360 degrees. 180 degrees is more used for sports events in a fixed setup, such as a stadium, whereas the 360-view is more used for documentary, news and open space sports.

The main drawback of 360-degree systems is the ratio of FOV (Field Of View)/captured video, which is roughly 1/5, meaning reduced resolution on the display. With a 180-degree system, the ratio is closer to 1/3, offering a higher resolution than with a 360-degree system, but only a 180 view.

3D relief

When video is captured, it can be either 2D, which can be played on any devices, or 3D, which can only be played on a HMD. 3D could be reproduced on a 3D TV, but for the time being 3D on TV has not sufficiently convinced the market, so it's only on HMD.

The video is today encoded via top/bottom approach, meaning not optimized in terms of compression, as opposed to the multi-view MPEG mode that will encode each view together.

2. Stitching

Today's stitching systems can stitch, in real time, four to six HD streams into one video stream up to 2160p60 8-bit resolution. The function can either be inside the camera with a limited performance or outside of the camera like what's been done with Nokia OZO or Video Stitch Vahana VR with a maximum level of performance. It is important to note that the stitching should be performed in real time even though the content will have to be post produced later, in order to streamline the production workflow.

3. Mapping

The classical approach is to map the video into an equirectangular projection. This consists of sending the full VR video to the decoder, which picks the Region of Interest (ROI).

Facebook is proposing a new scheme [1] where video is mapped on a polygon structure, taking into account the line of sight to remove details that are not in line of sight.

4. Encoding

Encoding can be done in several ways, using different techniques described in this section:

- Equirectangular
- Equirectangular tiled
- On-demand transcoding
- Polygon mapping

This section will analyse the different encoding techniques and the way this might impact the QoE (Quality of Experience) over the short and long terms.

For all presented encoding schemes, adaptive streaming can be added. In this case, different resolutions will have to be encoded.

Equirectangular

Equirectangular encoding is encoding of 2160p60 equirectangular mapped [2] content using DVB Ultra HD (UHD)-1 specification [3]. This is considered “state of the art” today.

The current MPEG HEVC compression algorithm is not taking advantage of the fact that the poles could be encoded with fewer bits, although the encoder could make those decisions.

The main advantage of this technique is the encoding is straightforward; the encoder gets UHD-1 (3840x2160x60) content at the input and provides HEVC Main encoded at the output. The device will extract the ROI based on the head’s position. The downside of this technique is the displayed video is coming from roughly 1/12 or 8 percent of the transmitted video, as depicted on Figure 2, while the bitrate transmitted to the device is the bitrate of the full picture (UHD-1 resolution). This explains the fuzziness aspect of the video when watched on an HMD.

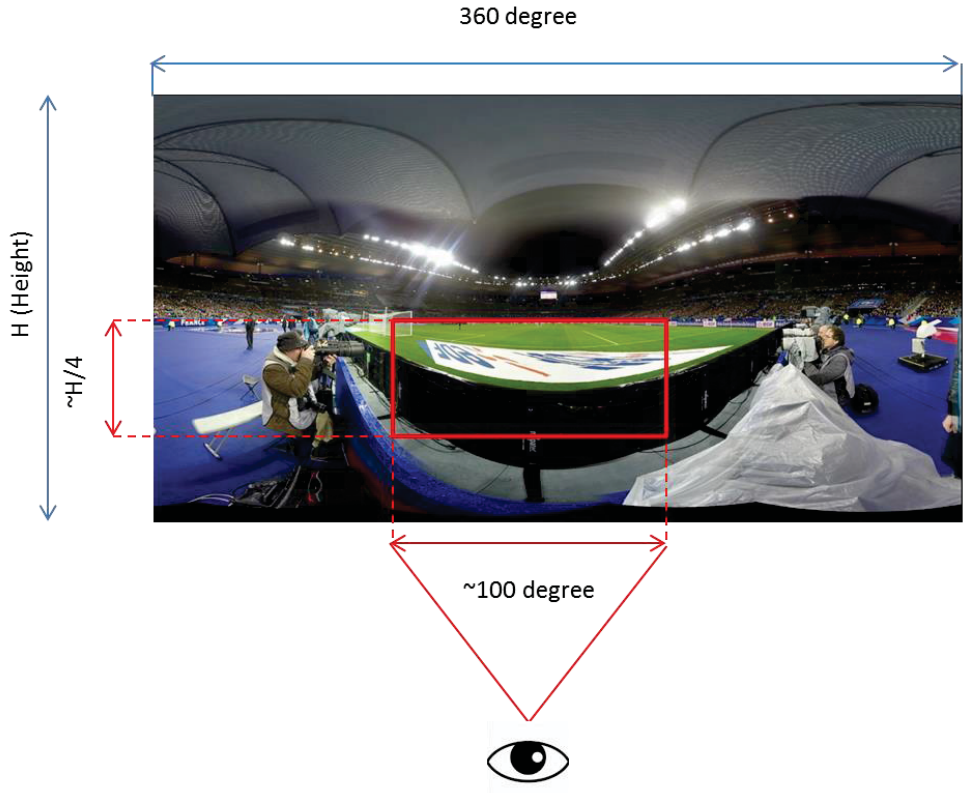

Figure 2 - Effect on resolution for an equirectangular system

We describe in figure 3 the system diagram of an equirectangular solution.

Figure 3 – Legacy Equirectangular System

In this scheme, the entire video is sent to the client that will extract window the ROI inside the full resolution and upscale it to the resolution of the display. In the HMD case, there is a clear fuzziness feeling, while on 2D devices like mobile or tablet, the experience is still acceptable.

This system can accommodate both a broadcast, multicast and unicast transmission, which makes it the most scalable, although what is deployed today is mostly Unicast over Internet.

The latency is minimum as there is no sever/client communication, deployed systems are below 20 ms latency.

Equirectangular tiles

Tiling systems have been proposed by BBC, TNO and HHI amongst others [4]. In this scheme, video is captured at 4x UHD or more, and the transmitted video will be UHD, meaning a full UHD experience vs. the HD experience achieved from the equirectangular ROI approach. Figure 4 represents an example with four UHD quadrants, each made of four tiles encoded independently.

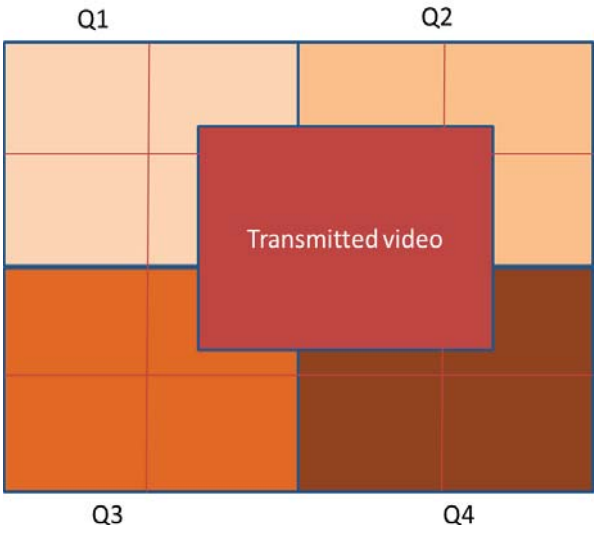

Figure 4 - Equirectangular unicast encoding of tiled content

In the example shown in Figure 4, only nine tiles out of 16 will be transmitted, which would represent half of the total bitrate if all the video was transmitted. With smaller tiles, the

transmitted video bitrate will asymptotically decrease to one-fourth of the total encoded resolution, to the detriment of the coding efficiency.

The tiling system is placed after the encoder in order to provide a minimum delay. Here also the E/E latency will be critical and the networking aspects will be key to avoid motion sickness. This will require 4x the encoding power vs. the equirectangular approach.

Figure 5 describes the architecture of an equirectangular unicast of tiled content system.

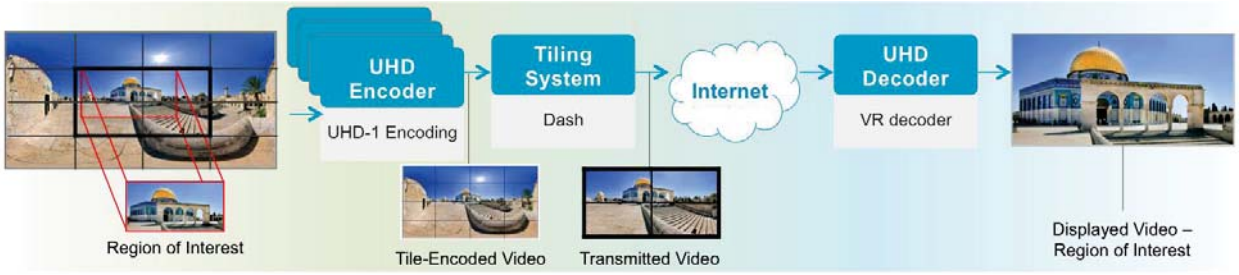

Figure 5 - Equirectangular Unicast of Tiled Content System

In this scheme, content is encoded at 4x the UHD resolution; the client requests a set of tiles to the server that receives the coordinates of the ROI. The content already encoded is parsed by the tiler that will send the appropriate tiles to the player, as described in Figure 4.

This system requires a unicast transmission, as each client has obviously a different ROI. As the system is unicast, it will require a strong CDN solution to scale for mass events. The system is very sensitive to network delay, and it might only be applicable on a wired network where ping time is below 50 ms and on future 5G networks that will have ultra-low delay in a similar range.

The latency added will be the titling processing time plus network transit. Total delay is made of Tiling processing (~20ms) + Network delay (~50ms) + HMD delay (20ms) should be below 100ms, it might be too long for a seamless experience.

This technology looks promising as it will enable a true UHD native experience since the player receives full UHD resolution.

On-demand transcoding

[5] Has introduced a system where the video sent to each user is the exact position of the VR ROI window. Under this approach, only the ROI is sent, thus saving on transmitted bandwidth. Details are described in Figure 5.

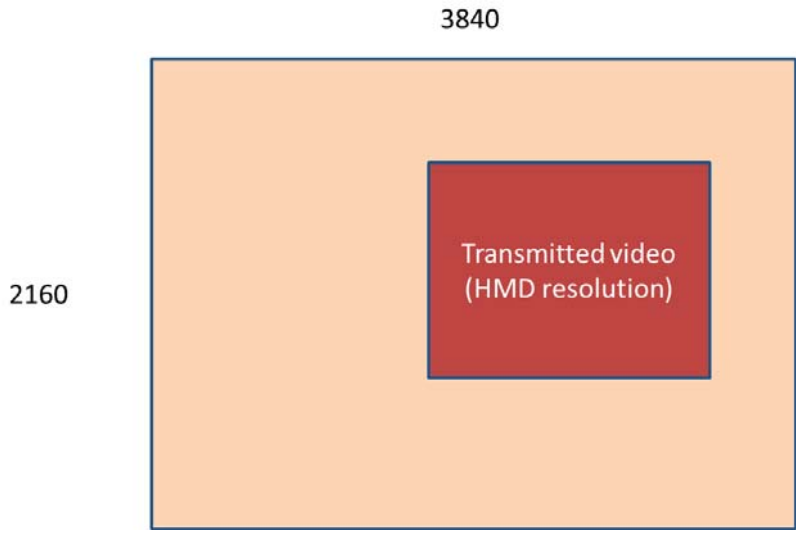

Figure 6 - On-demand transcoding windowing

Each video is transcoded per user, and massive transcoding techniques known as Transcoding on the Fly (TOTF), developed for cloud DVR applications, have to be applied. See Figure 7 for details.

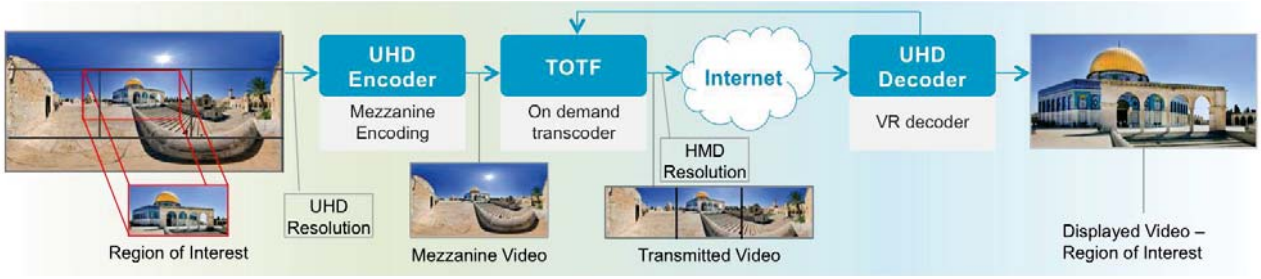

Figure 7 - On-demand Transcoding Workflow

The UHD mezzanine encoder encodes at the full UHD resolution (3840x2160). The TOTF system extracts the ROI window and will encode to the native resolution of the VR player. Table 1 illustrates the potential savings using this technique.

Number of pixels	Input	ROI	Transmitted to Gear VR
Horizontal	3840	1280	1280
Vertical	2160	540	540
% vs input	NA	8%	8%

Table 1 - Savings using an on-demand transcoding scheme

With this technique, we can divide the total number of transmitted pixels by 12. This does not improve the resolution of the video watched on the VR devices, but reduces

dramatically the bandwidth on the network, to the detriment of adding transcoding resources required by TOTF. Resolution can be increased, to the detriment of the transcoding scalability.

Total delay is made of TOTF retrieval from Cache (~20ms) + Network delay (~50ms) + HMD delay (20ms) should be below 100ms, it might be too long for a seamless experience.

From an economical point of view, such a system requires one transcoder per user, which will only work with a small amount of sessions. Once the number of sessions increases, a caching mechanism will have to be put in place to communalize the same request coming from different users.

Polygon mapping

Facebook has announced in January 2016 its next-generation video encoding techniques for 360 video and VR [1]. This consists of mapping the stitched video into different geometries. According to Facebook, this approach offers certain benefits compared with a classical ROI system:

- Moving from equirectangular layouts to a cube format in 360 video reduces the file size by 25 percent against the original. This exploits the fact that video out of the stitching process maps better on a sphere than on a rectangle.
- Encoding 360 video with a pyramid geometry reduces file size by 80 percent against the original. This is explained by the fact that pixels not in the line of sight will be more compressed than pixels in the line of sight.

For such a solution, we recommend the use of TOTF in order to increase the transcoding scalability. Figure 8 details the implementation.

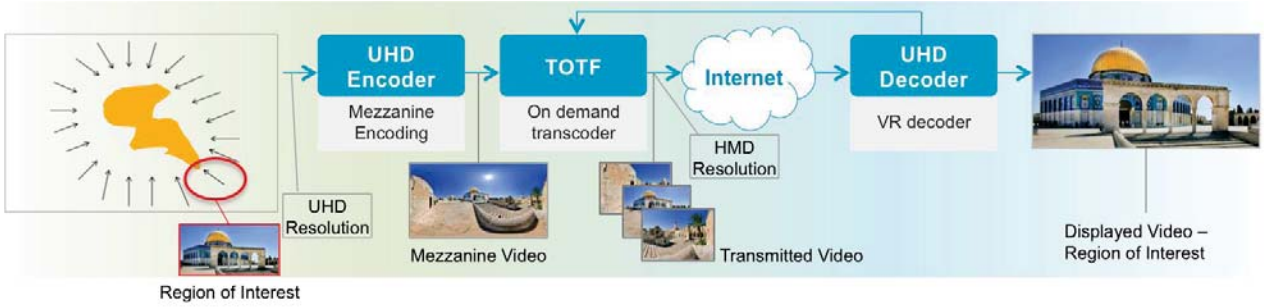

Figure 8 - Polygon Mapping Scheme

The transition between the different views will have to be seamless, which could be a challenge when the network starts to be congested or has a long ping time (typically larger than 100ms).

This system requires a unicast transmission and is very sensitive to network delay. As the system is unicast, it will require a strong CDN solution to scale for mass events. It might only be applicable on a wired network where ping time is below 50 ms and on future 5G networks that will have ultra-low delay in a similar range.

Facebook recommends 30 views x 5 profiles (at different bitrates) for ABR transmission. Overall this will be 150 views to encode, which obviously does not fit a real-time delivery business model, but could be applicable when a very high number of users can justify the encoding cost.

This technology looks promising as it will enable a 5x lower bitrate than the equirectangular broadcast approach. As quality will depend on the network response and transition between different views, it is too early to make an assessment before trials are performed with loaded networks.

Total delay is made of TOTF retrieval from Cache (~20ms) + Network delay (~50ms) + HMD delay (20ms) should be below 100ms, it might be too long for a seamless experience.

Encoding summary

Table 2 summarizes the different technologies available to encode VR video.

	Equirectangular		**Equirectangular tiles**		**On-demand transcoding**		**Polygon mapping**	
	H res	V res	H res	V res	H res	V res	H res	V res
Source resolution (1)	3840	2160	15360	2160	3840	2160	3840	2160
ROI resolution (2)	1280	540	2560	1440	1280	540	1280	540
Transmitted resolution	3840	2160	2560	1440	1280	540	3840	2160
Display (GearVR)	2560	1440	2560	1440	2560	1440	2560	1440
Zoom factor (3)	5.3		1.0		5.3		5.3	
Transmitted bitrate	10-15 Mbps		6-10 Mbps		<2 Mbps		2,5-4 Mbps	
Target delay (4)	Base		Base + 80ms		Base + 80ms		Base + 80ms	
Performance	Low resolution High bitrate		High resolution Medium bitrate		Low resolution Low bitrate		Low resolution Low bitrate	

Table 2 - Summary of encoding schemes for VR video

(1) Source resolution: Resolution after stitching

(2) ROI resolution: Resolution of what is watched by the end-user using a Gear VR (2560 x 1440).

(3) Zoom factor: Ratio between the displayed and the ROI pixels.

(4) Target values, not measured

The equirectangular system is what is currently deployed in devices and is the less performing in terms of quality and bandwidth.

The equirectangular tiled system has been researched over the past years and is expected to be demonstrated in the 2016-2017 timeframe, as all elements of the food chain are technologically available today. It offers the highest resolution.

The on-demand transcoding scheme helps to reduce the bandwidth, with a guaranteed QoE. This solution if the most bandwidth efficient. It has the possibility to increase resolution with a decrease in scalability.

The polygon mapping system looks promising, although the compression artefacts will be more pronounced than with other systems, and is also depending on network latency, with potential bandwidth savings.

Except equirectangular system, all other systems' QoE will depend on the system (encoding + distribution) latency to the device.

Future developments for mapping

A next-generation system might be a hybrid between the tiling, on-demand transcoding and polygon mapping. Either the resolution on the device is going to increase or the bitrate is going to decrease in the future, depending on the technology path chosen, which should overall improve the quality of experience of VR video.

MPEG is going to investigate VR encoding technologies for its Next Generation Video Codec initiative for 2020. Harmonic has already contributed in MPEG to push VR requirements in the Next Generation Video Codec initiative [6].

5. Transmission

VR video can be delivered either in broadcast or unicast mode. Table 3 provides a summary of all the different delivery mechanisms to transport VR video.

Transmission	Equirectangular	Equirectangular tiled	On-demand transcoding	Polygon mapping
Broadcast	TS	NA	NA	NA
LTE broadcast	DASH ISO BMFF	NA	NA	NA
Unicast	DASH ISO BMFF	DASH ISO BMFF	DASH ISO BMFF	Proprietary

Table 3 - Transmission system for VR video

The next section will look at how the different techniques described above can be mapped into the different transport protocols.

Broadcast

For TS delivery, the content will be decoded by TS capable devices, mainly STBs that will display on a TV. So far, DVB has not standardized the delivery of VR video over TS, but technically, it is possible today to carry VR video over TS, using DVB UHD-1 specification [2].

There is no plan to have a direct tuner inside a mobile VR device, so the most probable scenario would be gateway conversion from TS to IP (DASH) inside the home, such as defined by SAT>IP [7] .

The broadcast mechanism can only support the equirectangular scheme and will always have the lowest quality vs. other unicast schemes. Meanwhile, as the bandwidth available can increase, in 2018-2019 there may be a 2160p120 transmission that will be part of the HFR (High Frame rate) of DVB UHD-1 Phase2.

LTE broadcast

For LTE broadcast [8], the mobile device will host a VR application that will have to display on the phone or connected to a HMD. This is a single bitrate distribution as currently deployed in LTE broadcast.

The LTE broadcast mechanism can only support the equirectangular ROI and therefore will always have the lowest quality vs. other unicast schemes. Moreover, as the bandwidth available will always be limited, we do not foresee more than 1080p120 transmission for pure bandwidth constraints, therefore quality might suffer until 5G is deployed.

This transmission mechanism still offers the direct connection to the device and massive scalability in sports arenas, so this scheme could find an application.

Unicast

Unicast delivery over the Internet requires an adaptive streaming mechanism, while distribution via a QoE network (up to the device) can be done with a single bitrate scheme. Unicast can be delivered in UHD within a single bitrate using a live UHD off-the-shelf encoder with DASH packaging. This scheme was demonstrated by Harmonic and its partner Viaccess-Orca with TF-1 for the France vs. Russia soccer game [9].

As most of the VR devices will be connected to the Internet via Wi-Fi with an OTT service, ABR encoding is recommended.

The danger of ABR on HMD devices, where the resolution is critical, is if the resolution drops to HD, the perceived quality might degrade dramatically. More experiments will have to be conducted on HMD devices as well as 2D devices to see what is acceptable and adjust the profiles accordingly per devices.

Table 4 provides a recommendation for ABR Video profiles in a equirectangular configuration.

Profile	Resolution	Frame rate	TS bitrate
1	3840x2160	30	15 Mbps
2	3840x2160	30	10 Mbps
3	1920x1080	30	5 Mbps
4	1280x720	30	2 Mbps

Table 4 - Adaptive bitrate profiles

For unicast delivery, DASH ISO BMFF is the recommended streaming format. All HMD devices will support DASH ISO BMFF except iOS devices that are HLS. Note that adaptive streaming will impact the quality of experience when the bitrate drops too dramatically.

Unicast delivery can be applied to all VR video encoding schemes. Due to the massive amount of data involved for live events, ABR multicast will be a must in order to scale the service at peak usage.

In terms of complexity, we expect an ABR encoder to be 3x the complexity of a broadcast encoder. For an ABR tiling system, there would be a 4x3=12 x increase vs. equirectangular ROI system, while for the polygon mapping, it is 30x3=90x vs. an

Equirectangular ROI system. For the TOTF solution, the adequate profile will be transcoded on demand, so no additional complexity should be added.

Transmission Summary
Table 5 provides a summary of the different transport modes vs. devices.

Techniques applied	**Equirectangular**			**Equirectangular tiled**	**On-demand transcoding**	**Polygon mapping**
Transmission	Broadcast		Unicast	Unicast	Unicast	Unicast
Transport	TS	TS > IP (via Home GW)	DASH ISO BMFF	DASH ISO BMFF	DASH ISO BMFF	DASH ISO BMFF
Device	STB/TV	Any IP device	Any IP device	Any IP device	Any IP device	Any IP device

Table 5 - Summary of transmission options for VR video

CONCLUSION

This paper looked at the different technical options to transmit VR video using broadcast and unicast delivery mechanisms together with different mapping and encoding schemes. There is a significant difference in the user experience, network engineering, scalability and cost between the various options considered, and testing is expected over the next coming years to compare all those technologies, at scale. The production of VR video can be done in the more classical broadcast way or in a more personalized approach that better fits the HMD and mobile experience. Latency will be an important parameter of the user experience in a Mobile environment and this should be a key attribute for 5G services.

To conclude, VR video is still in its infancy. The different techniques presented in this paper will have to be evaluated at scale to determine which features bring real benefits. Based on those experiments, the industry will be in a better position to define the next generation VR delivery system for live, on demand across broadcast and unicast networks.

REFERENCES

1. Kuzyakov, E. and Pio, D. 2016. Next-generation video encoding techniques for 360 video and VR. https://code.facebook.com/posts/1126354007399553/next-generation-video-encoding-techniques-for-360-video-and-vr/
2. Equirectangular projection definition : http://www.kolor.com/wiki-en/action/view/Understanding_Projecting_Modes
3. ETSI TS 101 154 V2.2.1, (2015-06) specification

4. D'Acunto, L., Gregory-Clarke, R., Niamut, O. A., Thomas, E., Thomas, G. A. and van Brandenburg, A. 2014. Immersive Live Event Experiences – Interactive UHDTV on Mobile Devices. BBC Research & Development White Paper WHP 284.

5. Kolstad, K. 2016. Vantrix Showcases Disruptive Virtual Reality in 4K/HEVC at NAB. http://blog.vantrix.com/vantrix-showcases-disruptive-virtual-reality-in-4k/hevc-at-nab

6. Fautier, Thierry and Fogg, Chad. Commercial requirements for next-generation video. MPEG San Diego Jan'16 contribution M37665, by Harmonic.

7. Sat > IP : http://www.satip.info/technology-0

8. ETSI TS 126 346 V12.3.0, (2014-10) specification

9. http://www.viaccess-orca.com/what-s-new/pr/839-tf1-viaccess-orca-create-history-with-world-s-first-live-streaming-of-sports-event-in-360-and-in-ultra-hd-to-virtual-reality-headsets.html

ACKNOWLEDGEMENTS

I would like to thank those who have helped me write this paper: Kevin Le Jannic, Business manager at Viaccess-Orca, Raoul Monnier, Technology Innovation manager at Harmonic, Avni Rambhia, Industry Principal at Frost & Sullivan, Aytac Biber, Product Manager at Qualcomm as well as the Viacess-Orca team (Pascal Perrot, Alain Nochimowski), the Digital Immersion team (Julien Levy & al) and the Video Stitch team (Nicolas Burley, Claire Van de Voorde, Stephane Valente) who have helped Harmonic in the VR journey.