IMMERSIVE MEDIA OVER 5G - WHAT STANDARDS ARE NEEDED?

T. Stockhammer[1], I. Bouazizi[3], F. Gabin[2], G. Teniou[4]

[1] Qualcomm Incorporated, United States
[2]Ericsson, France
[3]Samsung, United States
[4]ORANGE, France

ABSTRACT

More than three-fourths of the world's mobile data traffic will be video by 2021. Mobile video will increase 9-fold between 2016 and 2021, accounting for 78 percent of total mobile data traffic by the end of the forecast period. Hence, Media distribution will remain to be the most relevant traffic on cellular networks. However, different to Managed services approaches in earlier Releases in 3GPP, new enablers are necessary to support both Over-The-Top (OTT) and managed media services over 5G networks. This presentation summaries the findings in the Third Generation Partnership Project (3GPP) in the context of 5G Media Distribution and addresses new opportunities and services including media production, immersive media, APIs for Augmented Reality (AR) and Virtual Reality (VR), support of browser and mobile terminals architectures, capability discovery and additional use cases. Based on the new 5G service architecture, this document provides an overview of how advanced media services can be mapped to the 5G architectures and how service and content providers leverage 5G enablers for new media services. The MPEG-I Immersive Media architecture, including rendering-based service layers as well as networked-based media processing, are expected to be a building block in the 5G media architecture. Some special use cases are highlighted including 6DoF services including network-media processing, mapping of AR services to 5G media distribution, as well as new enablers for managed media distribution services.

1 INTRODUCTION

Immersive Media will play a major role in the coming years. According to a forecast from IDC[i], the market for AR/VR will have a size of 215 Billion USD by 2021. By 2021, it is also expected to see massive adoption of 5G-based services, so just by the timelines, these two technologies will naturally go hand in hand. At the same time, the demand and

challenges for immersive services are expected to be addressed by the 5G system: lower latency, higher throughputs and ubiquitous access. In this paper we will address some efforts to harmonize immersive 5G and immersive media through open standardization approaches.

Figure 1 provides an evolutional approach from Virtual Reality (VR) towards integrated VR and Augmented Reality (AR). Nowadays AR and VR are more or less separate applications. It is expected that entire scenes, such as entertainment events, accessible with an AR device, will become so realistic and interactive that they'll be nearly indistinguishable from reality. In this context, VR becomes one mode of consumption of a larger AR service.

Figure 1 Future Evolution of VR and AR experience

Another relevant aspect about immersive media is the heavy use of many different technology components in a highly-integrated mobile computing platform. These components need to be used jointly and fulfil massive computational and real-time requirements, both individually but also through networking interfaces. Similar to VR experiences, this includes media decoding, graphics rendering as well as processing sensor input in real-time. Further to VR experiences, AR with 6DoF is expected to require significantly more media traffic. Aspects around the communication interfaces include:

- High throughput in both directions (uplink and downlink), and equivalently optimized compression technologies
- Low latency in the media communication to address the service requirements
- Consistency and universally high throughput

In this context, standards will be relevant to enable multi-vendor interoperable AR/VR services, enabling different players in the market to develop applications and services. Standards should provide the core component enablers. At least three organizations will provide significant enablers for immersive media services:

- Khronos, with OpenXR, is creating an open standard for VR/AR applications and devices, in particular on interfaces between applications and rendering engines. The present document provides an overview of the status of this effort.
- MPEG has initiated the new project on “Coded Representation of Immersive Media”, also referred to as MPEG-I, with multiple parts. One of them addresses initial 3DoF experiences in the Omnidirectional Media Format (OMAF) [6], but additional parts for new immersive media enablers are on the way.
- 3GPP has been working since Release-15 on 5G and the first specifications are approved to enable New Radio (NR) with capabilities for higher throughput and lower latency. Among those, 3GPP has also started initiatives for adding and combining immersive media to 5G. This document reviews the first approaches and the considered next steps in the following Releases.

2 IMMERSIVE MEDIA – AUDIOVISUAL QUALITY AND NATURAL INTERACTION

Immersive media creates the ability that users feel a sense of presence and immersion in an authored environment. Whereas in traditional media, steps towards passive consumption of more immersive media is relevant, for example by the inclusion of Ultra High Definition (UHD) and High Dynamic Range (HDR) for more realistic presentation of content in terms of contrast and colours (something addressed in 3GPP by the TV Video profiles defined in 3GPP TS 26.116 [2]), full immersion is achieved by the addition of the three pillars as documented in Figure 2. This includes:

- **Visual Quality** including high fidelity, spherical coordinates and stereoscopic rendering and depth information.
- **Sound Quality** including high-resolution and 3D audio, positioned such that the directivity of the sound sources can be rendered
- **Intuitive interactions** with the content using natural user interfaces, precise tracking of the motion and imperceptible latency to avoid lag and motion sickness

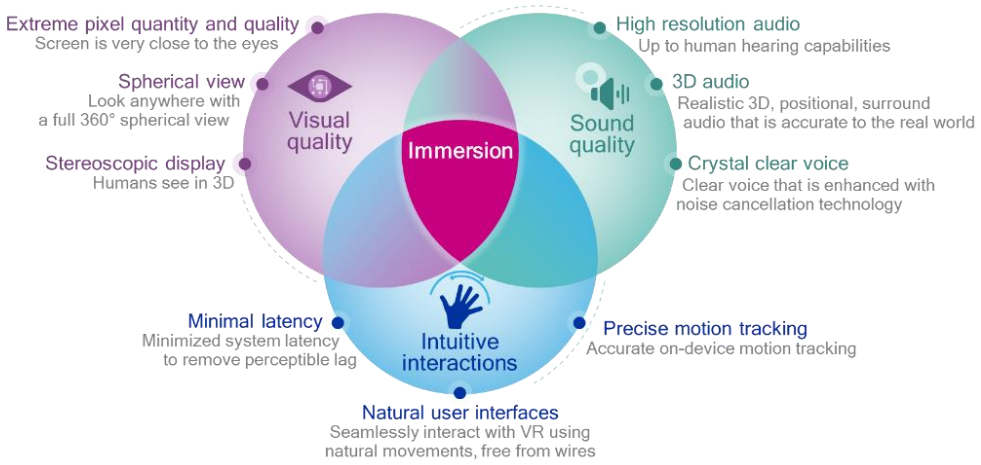

Figure 2 The three pillars towards immersion

User interfaces used to “interact” with traditional 2D content are well-known today like the consumption in apps and browsers. However, in this case the interaction is typically limited to content selection, random access, seeking and component selection. This selection/interaction impacts not only the media playback, but also the media decoding and delivery and hence formats and communication aspects are involved. However, the consumption once the media is accessed is predominantly passive.

The architecture of the first phase of immersive media in 3GPP and MPEG is based on Omnidirectional Media Application Format OMAF [6] as shown in Figure 3 including the key specification area for OMAF. The key specifications are on formats such that a device and applications can be build that permit rendering the formats.

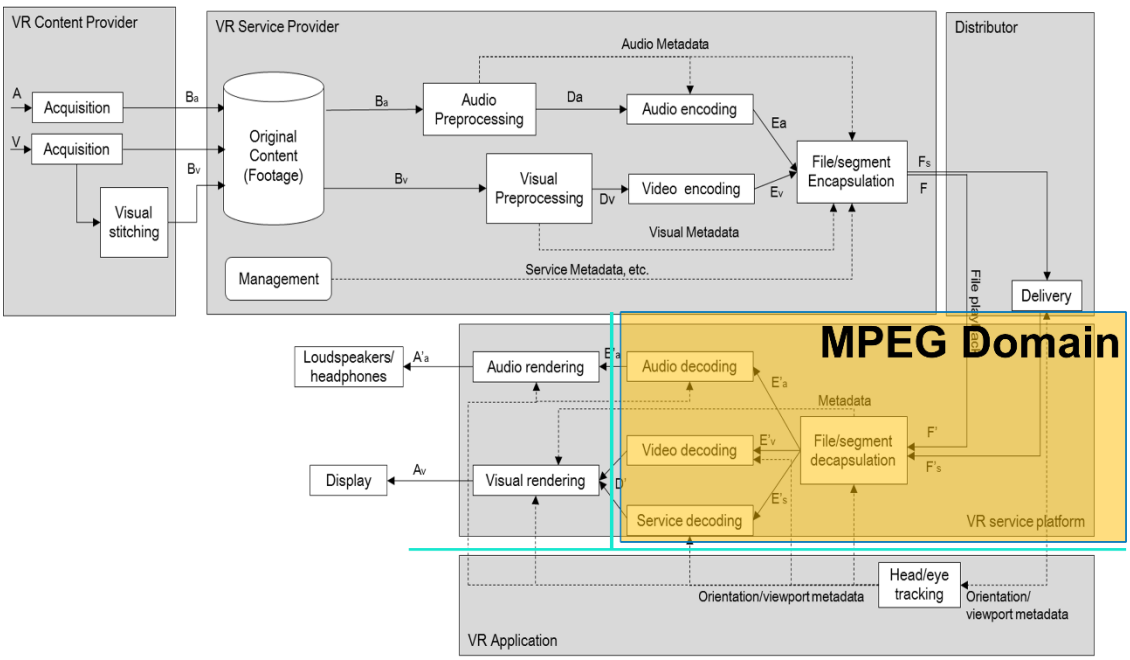

Figure 3 OMAF Architecture and the key MPEG domain specifications

3GPP adopted the OMAF architecture and addresses detailed specification for VR360 streaming in TS 26.118 [5]. This will be introduced in more details in section 3.

However, in moving towards more degrees of freedom as well as more interaction with the content, the additional sensors, contextual information and graphics engines will play a significantly more important role. Sensor such as light sensors, multiple cameras, multiple microphones, gyroscopes, etc. will be used to determine the user position, environmental factors and lightning conditions. In addition, graphics engines will be used to render traditionally coded 2D content such as the texture for objects that are controlled by certain geometries. Physically-Based Rendering (PBR), inherited from computer graphics shader engines, takes this approach to the extreme. With PBR, realistic light propagation, reflection/refraction patterns are mimicked with a high fidelity. In the natural video environment, it is defined as the Light Field representation, i.e. a model capable of representing the amount of light of every point in space in every directions.

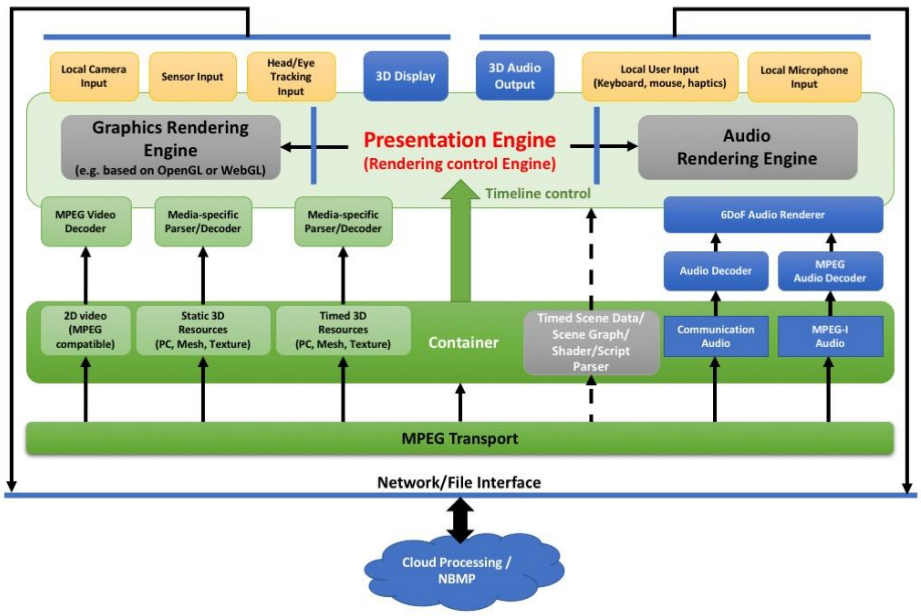

Figure 4 Graphics Centric Rendering

Most widely used graphics and game engines today rely on an OpenGL core (this may be different on some Operating Systems). They act as wrappers around OpenGL and offer more advanced functionality in a more abstract and simple way to use for the developers.

As depicted on Figure 4, when rendering VR/AR or 6DoF content, the rendering engine usually sets up a scene first. The scene maybe read from a scene graph/scene description document or it may be inferred from the content (e.g. a scene with a single sphere geometry for 360 video). The client may be given the option to choose for a full 6DoF scene rendering, it may opt for a simplified rendering or it may delegate parts of the scene rendering to the network. In the latter case, the network converts a full 6DoF scene into a simplified 6DoF scene, a 3DoF+ or 3DoF scene, or even into a 2D video.

The media resources of a content may be of a wide range of formats and types. They can either be 2D or 3D, natural or synthetic, compressed or uncompressed, provided by the content provider or captured locally (e.g. in the case of AR).

Key issues for interoperability expected to be solved include:

- **Definition of the spatial environment**, i.e. the space in which the presentation is valid and can be consumed. Typically, for example what is referred to as “windowed 6DoF” is the limited amount of possible movements within the 3D space.
- **Presentation timeline management**. The different resources may have an internal timeline for their presentation. The scene graph may have animations

and other scripts that incur an internal media timeline. In addition, scene graphs should also be updateable in a 6DoF presentation, where updates are timed or event driven. Finally, the container format may also specify the media timeline for the presentation of the embedded media.

- **Positioning and rendering of the media sources in the 6DoF scene appropriately**. Each individual media source may for itself have descriptive metadata of its geometry or it may be described by the scene graphs. In any case, such objects need to be properly integrated in the 6DoF scene.
- **Interacting with the scene based on sensor and/or user input**. The rendered viewport can be dependent on simple aspects such as viewing position or may include complex sensor input or captured signals such as geolocation, gyroscope, temperature, camera out, eye tracking, etc.

A more detailed discussion on 6DoF and AR standardization considerations is provided in section 4.

3 VR360 FOR STREAMING APPLICATIONS

Based on findings in the 3GPP TR 26.918 [4], 3GPP identified the necessity to define a consistent set of media enablers and interoperability points for VR360 Streaming applications in the first Release of 5G. The developed specification in TS 26.118 [5] is built upon the work in MPEG OMAF [6], but provides additional considerations for detailed interoperability aspects, especially in the client. The specification includes existing hardware capabilities but also the currently emerging ones that will be deployed together with the first set of 5G radio technologies. The interoperability aspects of the specification are considered in Figure 5.

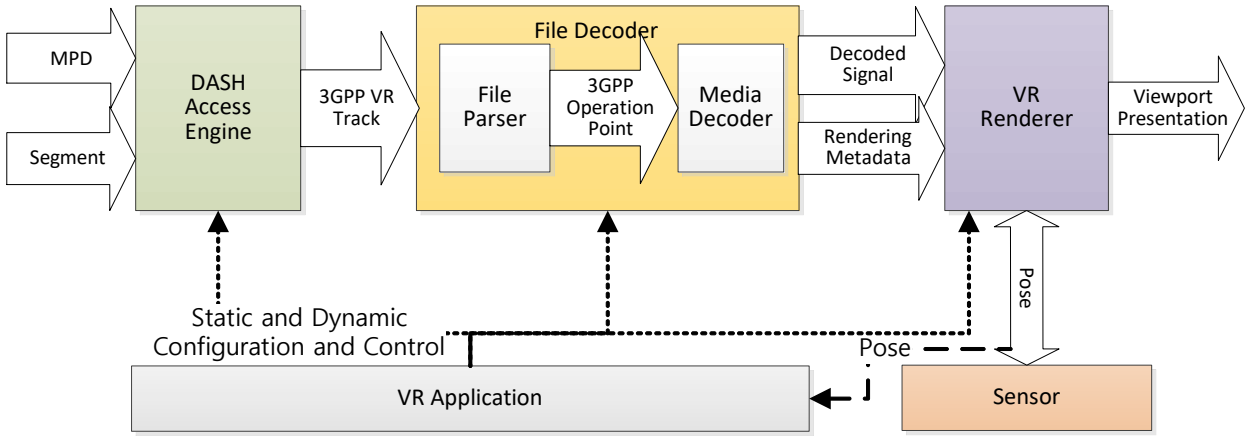

Figure 5 Interoperability aspects of TS26.118

Primarily, the specification focuses on using media decoders that are already available on common mobile devices, together with GPU and audio rendering engines, such that a 3DoF experience is achieved. For video, the operation points are shown in Table 1. Two of the operation points address legacy receivers for which only mono (2D) and full equirectangular projection (ERP) is used. One additional operation point is defined that

permits the use of Cube-Map projection (CMP), region-wise packing (RWP) for viewport dependent coding and rendering as well as the support for stereoscopic contents organized in Top and Bottom (TaB) frame packing.

Table 1 Video Operation Points in TS26.118

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

For distribution, it is expected that for most operation points, regular DASH architectures can be used. Only in the case of “H.265/HEVC flexible”, will the viewport-dependent streaming with tiling extension be considered. This is work in progress and the final decisions are expected to be provided in time for the presentation to IBC.

Audio for 360VR can be produced using three different formats. These are broadly known as channels-, objects- and scene-based audio formats. Audio for 360VR can use any one of these formats or a hybrid of these (where all three formats are used to represent the spherical sound-field).

This specification expects that an audio encoding system is capable of producing suitable audio bitstreams that represent a well-defined audio signal in the 3DoF reference system. 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. The signal representation is shown in Figure 6.

In TS 26.118, all audio 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.

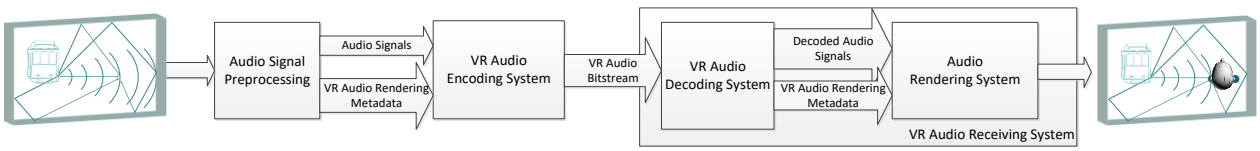

Figure 6 Audio Signal Representation

The audio specification is planned to be completed by September 2018, and details on the selected media profiles and their characteristics will be provided at the presentation to IBC.

4 6DOF AND AR – RENDERING AND NETWORK CENTRIC APPROACHES

The basic architecture developed in MPEG for more degrees of freedom is shown in Figure 4. This architecture foresees input through different modalities at the client to account for applications such as 3DoF/6DoF and even AR/MR. Input from local cameras and microphones is usually required for AR/MR applications.

Rendering is performed by a Graphics Rendering Engine for visual content and a 2D/3D Audio Rendering Engine for audio content. The Graphics Rendering Engine is usually based on some graphics libraries such as OpenGL or WebGL or even some higher-level engines such as Unity. Decoded media resources are composited together by the rendering engine to produce the Presentation.

The architecture supports MPEG defineD and non-MPEG media defined resources. These can be timed and non-timed. The container parser extracts information about resources and media timeline as well as any embedded or referenced media resources and make them available at the presentation engine.

The Architecture supports consuming the content in different forms, e.g. a simplified 2D version may be rendered on basic implementation clients, a limited 3DoF, 3DoF+, or 6DoF version may also be consumed by clients with higher capabilities. This pre-rendering/simplification may be done locally or in the network in a pre-rendering (baking) step. This is necessary for clients that are limited in their processing capabilities or network resources. Alternatively, the full-fledged 6DoF presentation may be consumed by clients capable of consuming it and having the required network resources and processing power.

The rendering engine may recompose 3D content from 2D-contents. An example is point clouds that are encoded using MPEG-encoded Point Cloud Compression.

To describe the scene of a presentation, a scene graph may be used. The scene graph may be provided in alternative formats to offer the renderer the choice of picking a supported scene graph format. Alternatively, a basic rendering operation may be described in the container format to support simple clients that do not support any of the included scene graph formats. Other scene description files such as scripts or shaders may also be included in the container.

Beyond local processing, rendering in the cloud and combining this with local presentation is one of the key issues to distribute workload. These efforts are currently being studied in 3GPP and a study item is under preparation at the presentation time of this paper.

6 CONCLUSIONS

After many years spent on improving the 2D content quality (more pixels, more colours, more contrasts…), immersive media represents the next big challenge for offering new quality of experiences in the TV and audio-visual environment.

Many significant impacts have been identified in this paper: Complex scene representation formats that will require a huge amount of data to be compressed before being distributed

on access networks with particularly high bandwidth capabilities and low latencies for interactive cases. Furthermore, all this data will have to be processed by highly efficient devices in terms of computing capability and rendering fidelity.

One option considered in this paper is to offload the end-user device processing need, by distributing the computation/processing load on to the network - this is particularly suitable for 5G networks with their high bandwidth and low latency capabilities.

In order to be able to provide such a device-complexity dynamic adaptation in the 5G network with mass scale interoperability, standardized interfaces and enablers need to be developed by 3GPP, taking into account the support of the latest defined immersive formats in MPEG and the recently defined client-APIs by Khronos/OpenXR. The 3GPP effort is likely to reach the Release 16 timeframe by end of 2019 (5G phase 2).

7 REFERENCES

[1] Nick Whiting, “Standardizing All the Realities: A Look at OpenXR”, GDC 2018, accessible here and here: https://www.youtube.com/watch?v=U-CpA5d9MjI

[2] 3GPP TS 26.116: “Television (TV) over 3GPP services; Video profiles”.

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

[4] 3GPP TR 26.918: “Virtual Reality (VR) media services over 3GPP”.

[5] 3GPP TS 26.118: “3GPP Virtual reality profiles for streaming applications”.

[6] ISO/IEC 23090-2: Information technology -- Coded representation of immersive media -- Part 2: Omnidirectional media format

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues for the collaborative and innovative work spirit in order to drive immersive media work in 3GPP and MPEG.

---

[i] https://www.idc.com/getdoc.jsp?containerId=prUS42959717