STATE-OF-THE-ART OF EXTENDED REALITY IN 5G NETWORKS

I.D.D. Curcio[1], S.N.B. Gunkel[2], T. Stockhammer[3]

[1]Nokia Technologies, Finland
[2]KPN/TNO, The Netherlands
[3]Qualcomm Incorporated, United States

ABSTRACT

Virtual Reality (VR) and Augmented Reality (AR) both retain importance in the media industry. However, the overall market adoption of VR technologies is still slower than expected. At the same time, industry efforts are shifting towards AR. This is also foreseen as a much larger market in the future, particularly for AR on mobile devices which are more widespread compared to VR devices. On the one hand, the AR explosion is due to recent developments in mobile technologies and on the other hand this is based on the promises of 5G networks. 5G, as a wireless communication infrastructure, will be able to satisfy the resource hungry demands (e.g., in terms of required bandwidth and low delay) of new Extended Reality (XR) services and applications. This paper will cover use cases, architectural, protocols and codec aspects for XR systems over 5G networks and their state-of-the-art in 3GPP and MPEG standardization organizations.

1 INTRODUCTION

5G promises new capabilities compared to existing network architectures, including higher bandwidth, lower latencies and new functions such as slices, virtualization and edge computing [1]. Immersive media and extended realities quite often are viewed as one of the key experiences that are enabled by 5G, taking into account the requirements in terms of bit rates, latency and ubiquitous availability of such services. Immersive media and 5G is considered a combination that will enable new services and opportunities within the mobile ecosystem, both in the entertainment and gaming environments, but also for new verticals such as industrial services, public safety and automotive.

To structure the work on 5G and XR, 3GPP has launched a feasibility study to identify use cases, technologies, and possible gaps that need specifications for interoperable services. In the context of this paper, Extended Reality refers to all real-and-virtual combined environments and human-machine interactions generated by computer technology and wearables. It includes representative forms such as Augmented Reality, Mixed Reality (MR) and Virtual Reality and the areas within the continuum among them. The levels of virtuality range from partially sensory inputs to fully immersive VR. The ultimate objective for immersive media is the experience of *Presence* providing the feeling of being physically and spatially located in the virtual environment. The sense of presence provides significant minimum performance requirements for different technologies, such as tracking, latency, persistency, resolution and optics [2]. Such experiences may be consumed on smart phones, Head-Mounted Displays (HMDs), AR glasses, heads-up displays or new emerging form factors devices.

In terms of integrating XR into networks, different aspects need to be considered. Among others are the (cumulative) applications data download size (e.g., Fortnite [26] has several GBs in download) or streaming immersive scenes at up to 100 Mbps. In advanced systems, the XR pose is not only processed in the device, but it is sent to the network in order to adapt to the current viewport or to a predicted viewport. In order to maintain the immersive experience, the pose needs to be processed in a matter of few milliseconds (typically 20ms). If the processing is delegated to the network, it either requires very low processing and communication latencies, or a smart separation of network-based pre-rendering and local rendering (also referred to as split rendering).

This document provides updates compared to what was presented 12 months ago [16], and provides a summary of the status of the study in 3GPP on XR over 5G as well as related aspects.

2 EXTENDED REALITY USE CASES FOR 5G

Consumers could envision a rich set of use cases covering XR experiences. One major effort has been undertaken by 3GPP in the definition of XR use cases in the context of 5G [3]. These use cases cover also a wide deployment temporal horizon that goes from immediate feasibility (e.g., over 4G networks) to 1+ years in the future (e.g., over 5G networks). In this section, we will survey a few of the representative use cases considered by 3GPP for XR over 5G.

Streaming

The typical media streaming experience is enhanced with the capability of 6 Degrees of Freedom (6DoF) within a scene. 6DoF motion and interaction are allowed in two possible ways: by changing the viewer's angle within the scene, and by head movements with an HMD. Additionally, the viewer's emotional reactions (e.g., facial expressions, eye movements, heartbeats, biometric data, etc.) could be collected by means of body sensors during a watching session and a personalized storyline could be created based on the type of emotions (for example while watching a movie). The stream display may occur over AR glasses on a chosen augmented wall within the home, after the spatial room configuration has been analyzed. Synchronized playback and interaction with multiple co-located viewers are possible.

This use case relies on volumetric video and 6DoF capture systems. New standardized methods for scene composition and description, social interaction, as well as new formats for storage and cloud access, content delivery and optimized streaming protocols and formats for biometric, emotion, and spatial metadata would need to be defined.

Gaming

Multi-player VR games will allow remote people to play and interact on the same game space. Here interaction and video quality (e.g., at least 60 fps and 8K resolution) are very important features. Users can change their in-game position by using controllers and body movements. It is possible for a game spectator to take two possible views: player's view and spectator's view. Interaction between a spectator and the players is also envisioned.

For this use case, decoding, rendering and sensor standard APIs are essential for success over multiple platforms. Also, low-delay streaming protocols are a must.

Real-Time 3D Communication

Video chats are captured using 3D models of people's heads, which can be rotated by the receiving party. Multi-party VR conferences support the blended representation of the participants into a single

360-degree video with a pre-recorded office background. Some of the conference participants may also be overlaid on an AR display. Shared presence using depth cameras is one of the features of this use case. In an instance, a virtual meeting space could be created, and the participants' avatars could move and interact with other avatars using 6DOF. Remote participants use an HMD and audio is binaural or spatially rendered.

The current 3GPP standards need to be extended to support dynamic 3D objects, metadata for spatial audio and 6DoF framework, multi-device media synchronization, network-based media processing functions for background removal, HMD replacement, etc.

Industrial Services

One of the use cases considered by 3GPP in this area covers an AR guided assistant at a remote location for augmented instructions/collaboration. It requires AR glasses. A remote assistant is guiding a local person to perform maintenance on an industrial machine. The remote assistant can see in real-time, through the local person's AR glasses, the local environment and the machine to be repaired. Part of the repairing instructions are sent as overlays to the AR glasses.

Messaging

The current MMS service could be extended to support 3D image messaging. This is realized by using devices equipped with a depth camera in order to capture a 3D image.

Standards will be extended to support formats for 3D images (e.g., meshes, point clouds and/or depth-layered images).

Other use cases

The list of XR use cases that could benefit from 5G also include others in the areas of training (possibly the largest market), automotive (engineering, design, marketing/sales), location-based entertainment, digital models, 3D holographic shows, passenger entertainment in self-driving cars, health care, data visualization among the others [4].

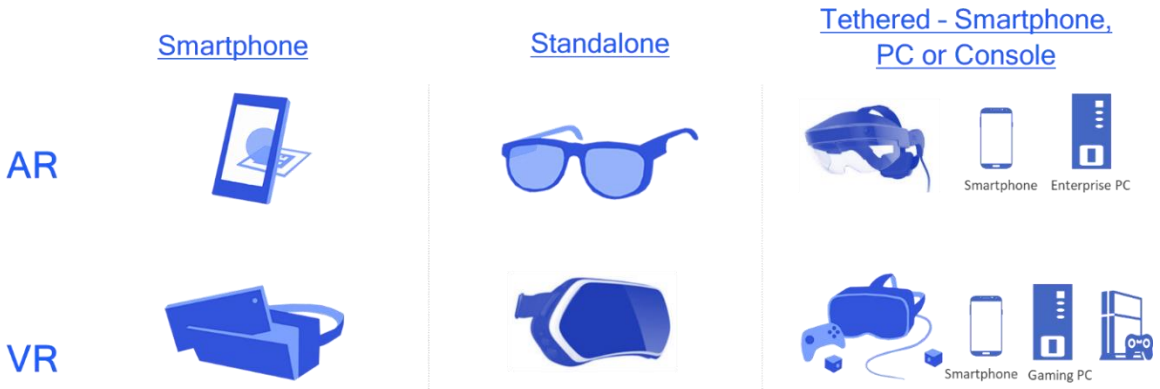

Figure 1 XR Form Factors

3 TECHNOLOGIES FOR EXTENDED REALITY OVER 5G

XR covers a wide range of technologies and end devices (see Figure 1). In this section we give an overview of some of those technologies relating to how to display 3D visual media with the help of volumetric video data and a view on a generic architecture for XR.

One essential part of XR applications is to rely on 3D immersive audio and volumetric visual media (sometimes referred to as holograms) to be displayed in the virtual, augmented or mixed into the real environment. The key aspect of such data formats is a full 3D representation of scenes and objects to allow experiences that are 6DOF. Currently the most relevant formats to store and transmit such content are:

- **PLY File Format [21]**, describes a 3D object as a list of vertices, faces and other elements, along with associated attributes. A single PLY file describes exactly one 3D object. The 3D object may be generated synthetically or captured from a real scene. Attributes of the 3D object that can be stored with the object include color, surface normal, texture coordinates and transparency.
- **OBJ File Format [22]**, is a textual file format that represents a set of commands to be run by the renderer. Commands are represented in separate lines, terminated by a newline. The first character of each line specifies the type of command.
- **Universal Scene Description (USD) [23]**, has been developed and made open source by Pixar with the goal of providing an interoperable scene graph format for the VFX industry. USD offers multi-file assembly of assets to enable splitting and composition of scenes. The parallel processing of the scene has been a key aspect of USD.
- **glTF 2.0 [24]**, is a new standard that was developed by Khronos to enable Physically Based Rendering. glTF 2.0 offers a compact and low-level representation of a scene graph. glTF 2.0 offers a flat hierarchy of the scene graph representation to simplify the processing. glTF 2.0 scene graphs are represented in JSON to ease the integration in web environments. The glTF 2.0 specification is designed to eliminate redundancy in the representation and to offer efficient indexing of the different objects in the scene graph.
- **Open Scene Graph (OSG) [25]**, is designed as a set of libraries that provide core functionalities of scene graphs but also allow for extensions for more customized scene graph processing.

The above formats offer advantages and disadvantages, and currently no common file format is adopted within the industry. Thus, more alignment and standardization effort is needed in order to evolve these formats for many new XR application and use cases. In order to capture scenes and objects in 3D (as volumetric mesh or point cloud data), several volumetric capture solutions exist:

- Volucap [5] captures an area with a diameter of 3m and 6m height, with the help of 32 high resolution cameras. The capture results into 2 TB of data per minute that is further processed and finally converted into mesh 3D data.
- 4DViews [6] captures an area with a diameter of 3m and 2.4m height with a rate of 60fps at a processing rate of 10 hours per minute of video.
- Microsoft Mixed Reality Capture [7]: multiple deployments of this capture studio technology exist. For instance, the latest deployment in Culver City (at Metastage) allows a capture area of 3m diameter with the help of 106 cameras (53 RGB ones, and 53 infrared). The final processed capture outcome is a in a mp4 movie format containing mesh information and texture images.
- Intel Studios [8]: this studio is designed to capture a large area of over 900m2 with the help of 96 high-resolution cameras. The capture data is transferred over fiber-optic cables at a data rate of 6 TB/min to allow a server farm of 90 processing units to create the volumetric video data.
- 8i [9] offers different size of capture at 2.5x2.2m (Volumetric Studio Stage) and 1.5x2m (Portable Volumetric Stage) equipped with 30 capture cameras each. A local processing farm is creating the final volumetric video stream as proprietary point cloud mp4 stream that can be streamed as live or on-demand content.
- Jaunt XR-Cast [10] uses 6 small depth cameras (Intel real-sense) to capture a small area with a low hardware and performance footprint but lower resolution than professional capture studios.

One of the promising formats to represent volumetric media are point clouds due to its high spatial resolution. MPEG has proven that a video coding-based approach for encoding dynamic 3D point cloud data outperforms the state-of-the-art methods significantly, both in objective and subjective quality [17]. One advantage of Video-based Point Cloud Compression (V-PCC) is its reliance on commercially available 2D video coding standards and technologies [11]. Thus, V-PCC can be easily implemented on current hardware, exploiting existing 2D hardware video decoders and mobile platforms [18]. Geometry-based Point Cloud Compression (G-PCC), also actively pursued by MPEG, stores the Point Cloud in 3D (rather than 2D) in an octree geometry structure. In this sense, G-PCC is a more complex and processing intensive encoding format [12].

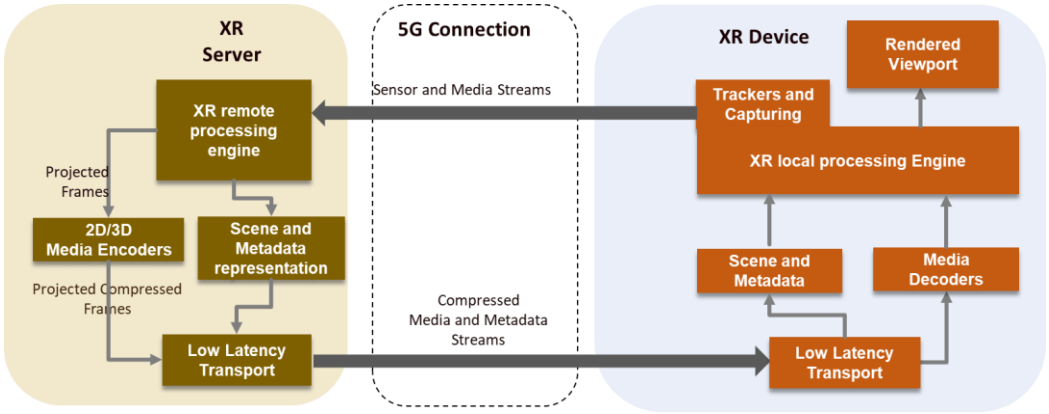

Figure 2 Generalized XR Distributed Computing Architecture

For the distribution of static content and assets to be displayed in XR applications it is possible to rely on distribution channels that are already in place. When it comes to live content and complex rendering in the cloud (or on the network edge) however, there is need for a new architecture in order to facilitate the full end-to-end pipeline. Figure 2 shows a preliminary Generalized XR Distributed Computing Architecture that is currently in discussion in 3GPP [3]. A highlight of this architecture is the 3D scene rendering, which could be either the direct viewport of the user or be converted into a simpler format to be processed by the device. The device does the final rendering based on local correction of the actual pose. In this way the graphics workload can be split into a rendering workload on powerful XR server and simpler rendering on the device, which is essential for mobile and standalone XR devices in terms of power consumption and limited processing. This is still an initial design of such an architecture and thus needs careful decompositions and analysis.

4 STANDARDS FOR EXTENDED REALITY OVER 5G

Technologies contributing to immersive media experiences quite significantly build on existing functionalities in the area of media distribution and compression as well as in existing computer graphics functionalities and APIs. It is expected that work in organizations other than 3GPP will support the standardized deployment of immersive media services. In the following, a few relevant activities in the MPEG, 3GPP, Khronos and W3C are summarized.

In October 2016, MPEG initiated a new project on “Coded Representation of Immersive Media”, referred to as MPEG-I. The project was motivated by the lack of common standards that do not enable interoperable services and devices providing immersive, navigable experiences. The MPEG-I project is expected to enable the evolution of interoperable immersive media services. Enabled by

the Parts of this Standard, end users are expected to be able to access interoperable content and services and acquire devices that allow them to consume these.

Core technologies as well as additional enablers are implemented in *Parts* of the MPEG-I standard (ISO/IEC 23091-X with X for the part). The most relevant activities [19] and the resulting parts are summarized in the following:

- Part 1 – Immersive Media Architectures: this overview summarizes use cases and architectures that motivate the development of specific components that contribute and support the distribution of immersive media services. While initially there was a focus on streaming distribution, new work is ongoing to decompose architectures for social VR and split rendering and extract the media-centric functions and requirements. Also, the development of architectures and requirements for the integration of networked and compressed media into 6DOF scenes is underway.
- Part 2 – Omnidirectional MediA Format [15]: OMAF addresses a first set of enablers for 3DoF experience based on existing MPEG technologies. It is the first published standard in MPEG (published in early 2019) that specifically addresses immersive media by combining and reusing existing MPEG compression (HEVC, MPEG-H audio) as well as storage and file formats. Currently, the development of the second edition is underway addressing extensions including limited 6DoF with multiple viewpoints, overlays and improvements to the efficiency of viewport-adaptive streaming.
- Part 3 – Versatile Video Coding (VVC): this is a future video compression standard to be finalized around 2020 by the Joint Video Exploration Team (JVET), a united video expert team of the MPEG consortium and the ITU. VVC predominantly addresses improved compression efficiency for high-resolution video (which directly benefits the efficient storage and distribution of video based immersive signals), but is also expected to provide better support for the integration of immersive media signals into the decoding. In particular the HEVC tiling concept is expected to be enhanced for more flexibility to enable decoding of multiple objects in one scene.
- Part 4 – Immersive Audio Coding: to support rich, immersive and highly interactive audio, MPEG addresses immersive audio in a new part in order to enable fully integrated audio-visual experiences. This specific project is expected to use the existing MPEG-H 3D Audio Low Complexity Profile (ISO/IEC 23008-3) as a compression engine and will define a rich audio rendering engine with supporting metadata for 6DoF and AR experiences. Integration of low-latency speech codecs is expected to be supported and enables the combination of 3GPP speech codecs (such as EVS and IVAS), for example to support social VR use cases.
- Part 5 and 9 – Point Cloud Coding: Point clouds are becoming popular to present immersive volumetric video due to the relative ease of capture and render when compared to other volumetric video representation. Several applications include 6DoF immersive video, VR/AR, immersive real-time communication, autonomous driving, cultural heritage and a mix of individual point cloud objects with background 2D/360 video. MPEG addresses two ways to compress Point Clouds, Part 5 (V-PCC), and Part 9 (G-PCC).
- Part 8 – Network-Based Media Processing (NBMP): it defines a framework that allows content and service providers to describe, deploy, and control media processing for their content in the network/cloud. The NBMP framework provides an abstraction layer on top of existing cloud platforms and is designed to integrate with 5G Core and edge computing. A particular aspect on NBMP is the integration of immersive media.

Other explorations such as 6DoF video representation and compression are underway. Additionally, traditional MPEG technologies such the MP4 file format and DASH streaming are enhanced to support immersive media storage and delivery.

Within 3GPP [20], the following are the standardization efforts for immersive services:

- The Immersive Voice and Audio Service (IVAS) Codec to be completed by 2021 is developed to address use cases including, but not limited to, conversational voice, multi-stream teleconferencing, VR conversational and user generated live and non-live content streaming. The approach proposed is to build upon the EVS codec with the goal of developing a single codec with attractive features and performance (e.g., excellent audio quality, low delay, spatial audio coding support, appropriate range of bit rates, high-quality error resiliency, practical implementation complexity). In the scope of 3GPP, the main audio rendering instrument is envisaged to be headphones, but configurations with, for example, tablet speaker playback may also be of relevance.
- QoE Metrics for VR with the objective to define device capability and latency metrics for optimizing the quality of experience.
- Immersive Teleconferencing and Telepresence for Remote Terminals (ITT4RT) with the objective to introduce immersive media support for conversational services.
- 5G Media Streaming architecture (5GMSA) with the objective to provide a modular architecture for streaming services including edge computing.

Finally, an important initiative for the harmonization of device functions and APIs is progressed by the *Khronos group*. OpenXR (Cross-Platform, Portable, Virtual Reality) defines APIs for VR and AR applications. The OpenXR 0.90 provisional specification was released in March 2019 [13]. It defines two levels of API interfaces that a VR platform's runtime can use to access the OpenXR ecosystem:

- Apps and engines use standardized interfaces to interrogate and drive devices. Devices can self-integrate to a standardized driver interface.
- Standardized hardware/software interfaces to reduce fragmentation, while leaving implementation details open to encourage industry innovation.

In order to support immersive experiences in browsers, the WebXR Device API Specification [14] developed in W3C provides interfaces to VR and AR hardware to allow developers to build compelling and comfortable VR/AR experiences on the web.

5 CONCLUSIONS

Extended Realities provide disruptive user experiences. However, only with capabilities provided by 5G, can a truly unbounded experience be expected. Essential features are higher data rates in both directions, ultra-low latency and reliability, greater density of connections, as well as massive compute resources close to the user. 3GPP is in the process of identifying and specifying the key functionalities for broad support of such XR services and experiences. These include radio capabilities, core network functionalities as well as media processing functions. Simple access to complex and powerful 5G and XR technologies will provide completely new business opportunities – both for third party service providers and MNOs.

6 REFERENCES

[1] ETSI White Paper, "MEC in 5G networks", https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf

[2] https://xinreality.com/wiki/Presence

[3] 3GPP TR 26.928: "Extended Reality in 5G".

[4] 3GPP/VRIF/AIS 2nd Workshop on VR Ecosystems and Standards, "Immersive Media meets 5G", 15-16 April 2019, Culver City, U.S.A.

[5] http://www.volucap.de

[6] https://www.4dviews.com
[7] https://www.microsoft.com/en-us/mixed-reality/capture-studios
[8] https://newsroom.intel.com/tag/intel-studios/
[9] https://8i.com/
[10] https://www.jauntxr.com/xr-cast/
[11] E.S. Jang at al., “Video-Based Point-Cloud-Compression Standard in MPEG: From Evidence Collection to Committee Draft”, IEEE Signal Processing Magazine, May 2019, pp. 118-123.
[12] https://mpeg.chiariglione.org/standards/mpeg-i/geometry-based-point-cloud-compression/g-pcc-codec-description-v2
[13] https://www.khronos.org/openxr
[14] https://immersive-web.github.io/webxr/
[15] ISO/IEC 23090-2: Information technology -- Coded representation of immersive media -- Part 2: Omnidirectional media format
[16] T. Stockhammer et al., "Immersive media over 5G - What standards are needed?", Proceedings IBC 2018. Amsterdam, The Netherlands, September 2018.
[17] S. Schwarz et al., “Emerging MPEG Standards for Point Cloud Compression”, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 10 December 2018, DOI: 10.1109/JETCAS.2018.2885981, https://ieeexplore.ieee.org/document/8571288
[18] M. Pesonen, S. Schwarz, “On implementing V-PCC standard”, Document m46074, MPEG #125 meeting, Marrakesh, Morocco, January 2019.
[19] R. Koenen, "MPEG-I - “I” is for Immersive", 3GPP/VRIF/AIS 2nd Workshop on VR Ecosystems and Standards, “Immersive Media meets 5G”, 15-16 April 2019, Culver City, U.S.A., https://www.vr-if.org/wp-content/uploads/1-2-2-Koenen-I-is-for-Immersive-1.pdf
[20] G. Teniou, "3GPP achievements on VR & ongoing developments on XR over 5G", 3GPP/VRIF/AIS 2nd Workshop on VR Ecosystems and Standards, “Immersive Media meets 5G”, 15-16 April 2019, Culver City, U.S.A., https://www.vr-if.org/wp-content/uploads/3GPP-achievements-on-VR-ongoing-developments-on-XR-over-5G.pdf
[21] http://paulbourke.net/dataformats/ply/
[22] http://paulbourke.net/dataformats/mtl/
[23] https://graphics.pixar.com/usd/
[24] https://www.khronos.org/gltf/
[25] http://www.openscenegraph.org/
[26] https://en.wikipedia.org/wiki/Fortnite

ACKNOWLEDGEMENTS

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