**Source: Tencent Cloud, Ericsson**

**Title: [FS\_5GSTAR] Adding shared AR conversational experience usecase, architecture and call flow**

**Agenda Item: 10.9**

**Document for: Discussion and Agreement**

1. **Introduction**

The current AR conversational service is designed around participant devices directly exchanging media streams. This contribution provides a more extended use-case where the communication occurs through a network/cloud entity which builds a common composed scene description that puts all participants in a virtual conference scene and streams the common scene as well as the associated media to all participants and therefore participants share the same virtual conference room.

The contribution proposes adding the use-case as well as the corresponding architecture, call flow, and required media formats.

1. **Changes and additions**

## 6.6 Shared AR conversational experience

## 6.6.1 Introduction

Shared AR Conversational experience is an end-to-end conversational service that includes communication between two or more parties through a network/cloud entity that creates a shared experience. The following building blocks that may be used to realize AR conversational services are identified:

a) Call setup and control: this building block covers the

- signalling to set up a call or a conference.

- fetching of the entry point for the AR experience. The protocol shall support upgrading and downgrading to/from an AR experience. It shall also support adding and removing media. This also includes the device type (Type-1, Type-2, or Type-3) as well as non-AR experience, e.g., tablet.

b) Formats: The media and metadata types and formats for AR calls should be identified. The format for the entry point, namely the scene description, and any extensions to support AR telephony need to be defined. Also, the format for media capturing, e.g., point clouds, colour attributes, etc. need to be identified. For AR telephony media types, the required QoS characteristics need to be defined, as well as the format properties and codecs.

c) Delivery: the transport protocols for the AR media need to be identified. AR telephony and conferencing applications require a low latency exchange of real-time media. A protocol stack, e.g. based on RTP, will be required.

d) 5G system integration: offering the appropriate support by the 5G system to AR telephony and conferencing applications includes:

- signalling for QoS allocation,

- discovery and setup of edge resources to process media for AR telephony,

- usage of MBS,

- data collection and reporting.

The building blocks may have different instantiations and/or options. For example, the delivery may be mapped to a WebRTC protocol stack or to an MTSI protocol stack. Furthermore, a single session may combine several delivery methods to accommodate the different media types supported by an AR conversational service.

In addition, AR telephony and conferencing applications may support asymmetrical and symmetrical experiences. In an asymmetrical case, one party is sending AR immersive media and the backchannel from other participants may be audio only, 2D video, etc. In a symmetrical case, all involved parties are sending and receiving AR immersive media.

### 6.6.2 Relevant use cases

The use cases relevant to this scenario may be further categorized including:

- UC#8: 360-degree conference meeting

- UC#9: XR meeting

- UC#10: Convention / Poster Session

- UC#12: AR avatar multi-party calls

- UC#13: Front-facing camera video multi-party calls

- UC#19: AR conferencing

- UC#22: shared AR conferencing experience

### 6.6.3 Basic architecture

To describe the functional architecture for shared AR conversational experience use-case such as clause Annex A.7 and identify the content delivery protocols and performance indicators, an end-to-end architecture is addressed. The end-to-end architecture for AR conferencing (one direction) is shown in Figure 6.6.3-1.

**63**

Camera(s) are capturing the participant(s) in an AR conferencing scenario. The camera(s) for each participant are connected to a UE (e.g. laptop or mobile phone or AR glasses) via a data network (wired/wireless). Live camera feeds, sensors, and audio signals are provided to a UE which processes, encodes, and transmits immersive media content to the 5G system for distribution. In multi-party AR conversational services, the immersive media processing function on the cloud/network receives the uplink streams from various devices and composes a scene description defining the arrangement of individual participants in a single virtual conference room. The scene description as well as the encoded media streams are delivered to each receiving participant. A receiving participant’s 5G STAR UE receives, decodes, and processes the 3D video and audio streams, and renders them using the received scene description and the information received from its AR Runtime, creating an AR scene of the virtual conference room with all other participants.

Also note that if the format conversion is desired, the immersive media processing function on the cloud may optionally use media services such as pre-processing of the captured 3D video, format conversion, and any other processing before compression of immersive media content including 3D representation, such as in form of meshes or point clouds, of participants in an AR conferencing scenario.

Editor’s Notes:

1. The exact functionality of the immersive media processing function, whether it includes a partial scene rendering such as Unity or it is pure scene description needs to be clarified. Clarification of 6.5 regarding the scene processing on the device is also needed.
2. For the composite scene generation, the immersive media processing function takes the input from the participants’ physical constraints, so that the generated scene is consistent with every participants environment and can be rendered at each device consistently.

Finally, while the sender’s functionality shown in Figure 6.6.3-1 is identical in the STAR and EDGAR devices, if an EDGAR device is used to receive the AR conversational services, it uses the split-rendering function on Cloud/Edge. In this case, the 5G STAR UE (receiver) of Figure 6.6.3-1, is replaced with the 5G EDGAR UE (receiver) and the cloud/edge functionality shown in Figure 6.6.3-2.



Figure 6.6.3-2: Replacement of the STAR UE device with EDGAR and cloud/edge prerendering.

Editor’s Notes the call setup and control in the case of EDGAR is still FFS.

### 6.6.4 Generic Call flow

Figure 6.6.5-1 illustrates the call flow for an immersive AR conversational for a receiving EDGAR UE. Only one sender is shown in this diagram without showing its detailed call flow.



Figure 6.2.4.1-1: Shared AR conversational experience call flow for a receiving EDGAR UE

Procedures:

1. Session Establishment:
   1. The AR/MR Application requests to start a session through EDGE.
   2. The EDGE negotiates with the Scene Composite Generator (SCG) and the sender UE to establish the session.
   3. The EDGE acknowledges the session establishment to the UE.
2. Media pipeline configuration:
   1. MAF configures its pipelines.
   2. EDGE configures its pipelines.
3. The AR/MR Application requests the start of the session.

Loops 4, 5, 6, and 7 are run in parallel:

1. AR uplink loop:
   1. The AR Runtime sends the AR data to the AR/MR Application.
   2. The AR/MR Application processes the data and sends it to the MAF.
   3. The MAF streams up the AR data to the EDGE.
2. Shared experience loop:
   1. Parallel to 9, the sender UE streams its media streams up to Media Delivery (MD).
   2. The sender UE streams its AR data up to the Scene Graph Compositor (SGC).
   3. Using the AR data from various participants, the SCG creates the composted scene.
   4. The composted scene is delivered to the EDGE.
   5. The media streams are delivered to the EDGE.
3. Media uplink loop:
   1. The AR Runtime captures the media components and processes them.
   2. The AR Runtime sends the media data to the MAF.
   3. The MAF encodes the media.
   4. The MAF streams up the media streams to the EDGE.
4. Media downlink loop:
   1. The EDGE parses the scene description and media components, partially renders the scene, and creates a simple scene description as well as the media component.
   2. The simplified scene is delivered to the Media Client and Scene Manager.
   3. Media stream loop:
      1. The pre-rendered media components are streamed to the MAF.
      2. The MAF decodes the media streams.
      3. The Scene Manager parses the basic scene description and composes the scene.
      4. The AR manager after correcting the pose, renders the immersive scene including the registration of AR content into the real world.

### 6.6.5 Various instantiations

Similar to clause 6.5.3, the shared AR conversational experience can be instantiated in various 5G systems:

a) The MTSI architecture (TS 26.114 [15]) supports audio and 2D video conversational services.

b) Extending the 5GMS architecture (TS 26.501 [26]) to support AR conversational services by combining live uplink and live downlink.

c) An architecture based on something different than MTSI / IMS or 5GMS, for example WebRTC.

For the comparison between different instantiations, please refer to Table 6.5.3-1.

6.6.6 Content formats and codecs

Based on the use cases, the following formats, codecs and packaging formats are of relevance for Media Streaming of AR:

- Scene Graph/Description

- 2D Video Formats

- 3D Formats such as static and dynamic point clouds or meshes

- Animated 3D meshes

- 2D Video Formats with depth

- 2D, stereo, and spatial audio formats

- Several video decoding instances

- Decoding tools for such formats

- Encoding tools for 2D formats

- Low-latency downlink and uplink real-time streaming of the above media

- Uplink streaming of pose information

- Uplink streaming of media

6.6.7 Standardization areas

TBD.

# A.7 Use Case 22: Shared AR Conferencing Experience

|  |
| --- |
| **Use Case Description: Shared AR Conferencing experience** |
| This clause describes an AR conferencing use case that allows participants to participate in a shared virtual conference room experience. For each participant, the other participants' video objects are registered on the participant’s AR scene, creating the sense that the conference room is held in the physical location of the participant. The arrangement of participants (i.e. their location relative to each other) in the virtual room are the same which creates a consistent sense of conference room layout for all participants.  Bob, Alice, and Tom are participating in a virtual meeting. Bob is having the call at his home's kitchen and sees Alice and Tom with their 3D volumetric representation on his AR glasses. In Bob’s view, Alice and Tom are sitting at Bob’s kitchen table with Alice on left and Tom on the right-hand side. Alice is in her office conference room. She sees Bob and Tom their 3D volumetric representation on her AR glasses. For Alice, Tom is sitting on the left and Bob on the right-hand side of her conference table. Finally, Tom is having the call at the airport lounge. For him, Bob is sitting on a chair on the right and Alice is sitting on the couch on the left side. While the real worlds of Bob, Alice, and Tom are different, in all scenes the participants are seated in the same arrangement relative to each other. Therefore, when Alice turns to Tom, Tom and Bob see the consistent views of Alice as if they are in the same physical room. |
| **Categorization** |
| **Type: AR**  **Degrees of Freedom: 3DoF+ or 6DoF**  **Delivery: Conversational**  **Device: AR glasses** |
| **Preconditions** |
| - The participants are wearing AR glasses that allow the 3D volumetric representation of other participants. |
| **Requirements and QoS/QoE Considerations** |
| The network shall support the delivery of 3D volumetric streams for real-time conversational services:  - Support of creating a composed scene in the network  - Support of different volumetric user representation formats.  - Bitrates and latencies that are sufficient to stream volumetric user representations under conversational real-time constraints. |
| **Feasibility** |
| The bandwidth and latency requirements for AR conferencing using 3D volumetric representations present a challenge to mobile networks. The complexity of the 3D volumetric representations is challenging for the endpoints and introduces additional delay for processing and rendering functions. Intermediate edge or cloud components are needed.  In the following are some indicative values of a potential solution and transmission format for different types of user representation:  - A point cloud stream has raw bandwidth requirement of up to 2 Gbps. The transmission bandwidth is expected to be lower after encoding and optimization.  - Preliminary data from MPEG V-PCC codec evaluation indicates compression ratios ”in the range of 100:1 to 300:1”[40]. For dynamic sequences of 1M points per frame this could result into an encoding bitrate of “8 Mbps with good perceptual quality” [40]. For conversational services, we expect lower compression ratios.  - 2D/RGB+Depth: >2.7Mbps (1 camera @ 30fps with total resolution of 1080x960 [37]), >5.4Mbps (2 Camera @ 30fps with total resolution of 1080x  - 3D Mesh: ~30 Mbps @ 20-25 FPS (with a voxel grid resolution of 64x128x64 and 12-15k vertices) [39].  - Preliminary data from 3D GPCC show that bitrates in the range of 5-50 Mbps @ 30 fps with varying octree depth and varying JPEG QP are expected [39]. |
| **Potential Standardization Status and Needs** |
| The following aspects may require standardization work:  - Standardized formats for 3D volumetric representation of participants on AR glasses.  - Cloud APIs for processing and rendering of 3D volumetric streams.  - Conversational methods for call initiation.  - Spatial audio formats and associated metadata.  - Metadata for Spatial characteristics of the AR environment (e.g. positioning of users). |

# 5 Core Use Cases

This clause documents the core use cases and scenarios for AR/MR devices, which serve to extract requirements, functional structure, related media format, and protocols for the 5G systems. Parts of the use cases are derived from XR use cases in TR26.928 [2] based on the relevance to AR/MR device type. In addition, the other use cases and scenarios are collected in Annex A of this document.

Table 5-1 provides a list of all the collected use cases.

Table 5-1: List of use cases for AR/MR services

|  |  |  |
| --- | --- | --- |
| No | Use Case | Reference |
| 1 | 3D Image Messaging | Annex A.2 in [2] |
| 2 | AR Sharing | Annex A.3 in [2] |
| 3 | Real-time 3D Communication | Annex A.8 in [2] |
| 4 | AR guided assistant at remote location (industrial services) | Annex A.9 in [2] |
| 5 | Police Critical Mission with AR | Annex A.10 in [2] |
| 6 | Online shopping from a catalogue – downloading | Annex A.11 in [2] |
| 7 | Real-time communication with the shop assistant | Annex A.12 in [2] |
| 8 | 360-degree conference meeting | Annex A.13 in [2] |
| 9 | XR Meeting | Annex A.16 in [2] |
| 10 | Convention / Poster Session | Annex A.17 in [2] |
| 11 | AR animated avatar calls | Annex A.18 in [2] |
| 12 | AR avatar multi-party calls | Annex A.19 in [2] |
| 13 | Front-facing camera video multi-party calls | Annex A.20 in [2] |
| 14 | AR Streaming with Localization Registry | Annex A.21 in [2] |
| 15 | 5G Shared Spatial Data | Annex A.24 in [2] |
| 16 | AR remote cooperation | Annex A.1 |
| 17 | AR remote advertising | Annex A.2 |
| 18 | Streaming of volumetric video for glass-type MR devices | Annex A.3 |
| 19 | AR Conferencing | Annex A.4 |
| 20 | AR IoT | Annex A.5 |
| 21 | AR gaming | Annex A.6 |
| 22 | Shared AR Conferencing experinece | Annex A.7 |

The use cases may be grouped into several categories based on the similar requirements for media flow and device functional structure.