**Agenda item:** 9.8

**Source:** Qualcomm Incorporated

**Title:** [5MBUSA] Low-latency Broadcast

**Document for** Discussion andAgreement

# Introduction

This document addresses

* Define relevant call flows and procedures to support 5GMS over 5MBS
* concrete steps in TR 26.802, clause 5.5.2 on DASH/HLS over MBMS as defined in TS 26.346, clause 5.6 and 5.7, including Low-Latency CMAF as defined in 5GMS.

This document assumes that architectures and session concepts from S4-211005 are agreed. Clause 2 summarizes the main proposals.

# Proposed Architecture and Procedures

Figure 3.1-1 updates Figure 5.1-2 from TS 23.247 to provide a more user service centric view. It also provides an update to Figure 4.4.3-1 from TR 26.802. The red highlights provide the main scope for User Service Specification from a northbound interface.



Figure 3.1-1 User Service Centric 5G MBS system architecture in reference point representation

In Figure 3.1-2, a proposed update to the 5MBS User Service Architecture is proposed that addresses the different interfaces defined in the work item description.



Figure 3.1-2 5G Multicast Broadcast User Service (5MBUS) Architecture

The following definitions are assumed

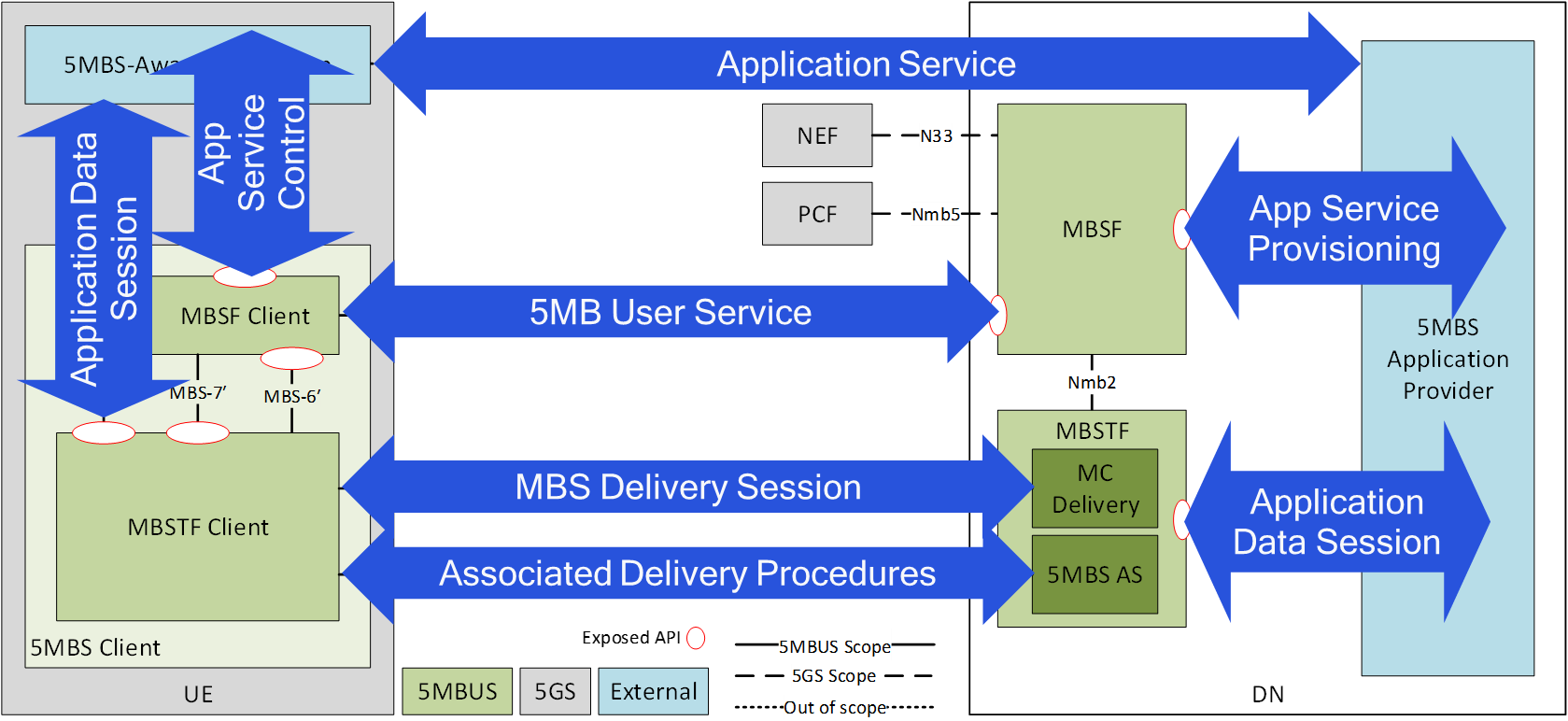
**Application Service**: An end-user service for which parts of the data or all of the data of this service is accessible by joining an 5MB User Service.

**5MB User Service:** The transport-level service using an MBS Delivery Sessions to delivery an Application Service.

**MBS delivery session:** time, protocols and protocol state (i.e. parameters) which define sender and receiver configuration and use an MBS session for the delivery of an application data session.

**Application data session:** time, protocols and protocol state (i.e. parameters) provided by an 5MBS application provider for delivery over 5MBS and provided to the 5MBS aware application.

This concept is now also presented in the below Figure



Initial session and service procedures are provided in Figure 6-1.



Figure 6-1 5MB User Service Workflow

# 5GMS-over-5MBS Architecture and Call Flow

The architecture in Figure X below represents a harmonized architecture for 5G Media Streaming via 5MBS.



Figure X Harmonized architecture for 5G Media Streaming over 5MBS

Two deployment options are provided, as indicated also in Annex A of TS 26.501.

- Consist of multiple (physical) servers, where different servers, or different groups of servers, may be addressed with different FQDNs. The client may be made aware of this via the manifest (i.e. listing multiple base URLs).

NOTE: In this case the servers may be managed by the same or different parties (e.g. MNO and/or 5GMSd Application Provider).

- Be addressed with a single FQDN. For example, the MNO AS is mostly transparent and acts as a proxy/cache.

The second case addresses the scenario for which the service is exclusively being provided through 5 MBS and no unicast for data delivery exists. The first case addresses the scenario, for which parts of the service are also available on unicast, i.e. the hybrid case. This is discussed in more details in S4-210009.

In the architecture, no new functions or interfaces are defined. However, some of the reference points need extensions to fully support the two scenarios.

# Call flow for 5GMS via 5MBS

The initial call flow addresses the delivery of 5GMS media data exclusively via 5MBS broadcast. For this, the call flow in TS 26.501, clause 5.3.2 is extended accordingly.



**Figure X: High Level Procedure for DASH content via 5MBS**

Prerequisites:

- The 5GMSd Application Provider has provisioned the 5G Media Streaming System and has setup content ingest.

- The content ingest is forwarded to the MBSTF using the M1d parameters.

- The 5GMSd Aware Application has received the service announcement from the 5GMS Application Provider.

Steps:

1: The 5GMSd Aware Application triggers the Service Announcement and Service and Content Discovery procedure. The Service and Content Discovery procedure only involves the App and the external Application Server. The Service Announcement includes either the whole Service Access Information (i.e. details for Media Session Handling (M5d) and for Media Streaming access (M4d)) or a reference to the service access information.

2: A media content item is selected.

3: The 5GMSd-Aware Application triggers the 5GMSd Client to start media playback. The Media Player Entry is provided to the 5GMSd Client.

4: When the 5GMS-Aware Application has received only a reference to the Service Access Information (see step 1), the Media Session Handler interacts with the 5GMSd AF to acquire the whole Service Access Information.

5 - 11: The Media Session Handler acts as a 5MBS-aware application and initiates the service acquisition. For details, see TS 26.347. This establishes transport session for MPD and Content.

12: In parallel, the Media Player is invoked to start media access and playback.

13: The Media Player requests the MPD.

14: The Media Player receives the MPD.

15: The Media Player processes the MPD. It determines for example the number of needed transport sessions for media acquisition. The Media Player should be able to use the MPD information to initialize the media pipelines for each media stream. The MPD should also contain information to initialize the DRM client, when DRM is used.

16: The Media Player notifies the Media Session Handler about the MPD. The notification may contain parameters from the MPD.

17: Optional: the Media Player acquires the necessary DRM information, for example a DRM License.

18: The Media Player configures the media playback pipeline.

19: The Media Player requests initialization information. The Media Player repeats this step for each required initialization segment.

20: The Media Player receives the initialization information.

21: The Media Player requests media segments according to the MPD.

22: The Media Player receives media segments and puts the information into the appropriate media rendering pipeline.

23: Previous steps are repeated according to the MPD information.

# End-to-End Latency Analysis of DASH Streaming

An end-to-end latency analysis basically provides a summary as follows:

1. Content of Segment is encoded into GOP structure
   1. Typically GOP latency
2. Segment is generated on application server
   1. Latency is Segment duration
3. Segment is uploaded to 5GMS/5MBS
   1. Latency is dependent on upload speed
4. Segment is delivered as one object over bearer with certain bitrate
   1. Latency until received is roughly segment duration
5. Segment is recovered by FEC and possibly file repair
   1. Additional latency needs to be provisioned
6. Segment is provided to Media Player which buffers one or several segments
   1. Additional latency is 1-3 times segment duration

Total latency is roughly around 3-6 segment durations. Smallest segment durations for efficiency need to be at least 2 seconds. This results in a latency between 6 to 12 seconds.

Other aspects related to low-latency are:

* the timing differences and the need for providing a more accurate time-sync for the UE. Does Low Latency Broadcast require tight time sync like in TSN (using IEEE 1588)? Or is 1sec time sync precision enough?

# Low-Latency DASH with CMAF

Low Latency DASH is introduced in TS 26.247 as part of 5GMS. Instead of dealing with Segments, chunks are used as delivery modes. Chunks typically have durations of for example 320ms. If an end to end system supports the delivery of chunks instead of segments, then the 3-6 segment durations latency can be addressed with 3-6 chunk durations, possibly resulting in latencies of 1-2 seconds. Additional latencies may still be needed, for example if retransmission or FEC is used. However, a latency target of 3-4 seconds is realistic.

Hence, it is important to support Low-Latency DASH with CMAF end to end in 5GMS + 5MBS. This requires

1. New ingest protocol, or at least an understanding how the existing protocols perform and can be optimized.
2. Enhanced content delivery protocol
3. Updated client-side APIs to support for example chunked delivery
4. Additional session reporting to track latency.

Note that a key issue is also available in TR 26.804 on this matter for unicast only.

# Proposal

Based on this contribution it is proposed:

1. Adopt the architecture in clause 3 for 5GMS via 5MBS
2. Adopt the basic call flow in clause 4 for 5GMS via 5MBS
3. Address Low-Latency DASH/CMAF based on the initial discussions in clause 6 and 7.