**3GPP TSG SA WG4#117e S4-220245**

**E-meeting, 14th – 23rd February 2022**

|  |
| --- |
| *CR-Form-v12.0* |
| **PSEUDO CHANGE REQUEST** |
|  |
|  | **26**.**804** | **CR** | Pseudo | **rev** | **4** | **Current version:** | **1.0.0** |  |
|  |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME |  | Radio Access Network |  | Core Network |  |

|  |
| --- |
|  |
| ***Title:***  | [FS\_5GMS\_EXT] TV-grade mass distribution of unicast Live Services |
|  |  |
| ***Source to WG:*** | Qualcomm Incorporated |
| ***Source to TSG:*** |  |
|  |  |
| ***Work item code:*** | FS\_5GMS\_EXT |  | ***Date:*** | 16/12/2021 |
|  |  |  |  |  |
| ***Category:*** | **B** |  | ***Release:*** |   |
|  | *Use one of the following categories:****F*** *(correction)****A*** *(mirror corresponding to a change in an earlier release)****B*** *(addition of feature),* ***C*** *(functional modification of feature)****D*** *(editorial modification)*Detailed explanations of the above categories canbe found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | *Use one of the following releases:Rel-8 (Release 8)Rel-9 (Release 9)Rel-10 (Release 10)Rel-11 (Release 11)Rel-12 (Release 12)**Rel-13 (Release 13)Rel-14 (Release 14)Rel-15 (Release 15)Rel-16 (Release 16)* |
|  |  |
| ***Reason for change:*** |  |
|  |  |
| ***Summary of change:*** |  |
|  |  |
| ***Consequences if not approved:*** |  |
|  |  |
| ***Clauses affected:*** |  |
|  |  |
|  | **Y** | **N** |  |  |
| ***Other specs*** |  |  |  Other core specifications  | TS/TR ... CR ...  |
| ***affected:*** |  |  |  Test specifications | TS/TR ... CR ...  |
| ***(show related CRs)*** |  |  |  O&M Specifications | TS/TR ... CR ...  |
|  |  |
| ***Other comments:*** |  |
|  |  |
|  |  |

**===== CHANGE =====**

[T] 3GPP TR 26.925: "Typical traffic characteristics of media services on 3GPP networks".

[V] 3GPP TR 26.917: " Multimedia Broadcast Multicast Services (MBMS) and Packet-switched Streaming Service (PSS) enhancements to support television services".

[X] "DASH-IF WEBRTC-based Streaming", https://dashif.org/news/webrtc/

**===== CHANGE =====**

## 5.11 TV-grade mass distribution of unicast Live Services

### 5.11.1 Description

#### 5.11.1.1 General

Live TV services of different scale (professional, user-generated, session-based, etc.) are increasingly distributed over broadband and mobile networks. Live TV services are characterized by:

- scalability (in terms of concurrent users),

- consistent quality,

- high bandwidth requirements,

- target latency constraints, and

- TV Experiences.

#### 5.11.1.2 Scalability

Consistent support of the distribution of such services to a different scale of users and in a concurrent fashion is a prime concern. 5G Media Streaming is expected to support such service distribution and end-to-end optimizations. Improvements and optimizations on the architectural level and stage 3 are expected to be studied.

#### 5.11.1.3 Consistent quality

TV Services are expected to provide a consistent quality over time. Consistent quality can be defined as the service staying within a set of Operation Point boundaries to the largest extent. Specific aspects are:

* Meeting the latency requirements, namely staying at the target, not exceeding the maximum, and not falling below the minimum. Measuring deviation from the target, as well creating events when exceeding the boundaries, is relevant. For more details see clause 5.11.15.
* Meeting the target playback rate, i.e. how often the playback rate is changed from 1.0, and if there are cases when the playback rate is outside of the range. For more details see clause 5.11.15.
* Staying within the boundaries of the target bit rates and meeting the target. This measurement includes those cases where the bit rate falls to 0, i.e. the service is stalled. All of this can be measured by the 5GMS Client.

Table 13.2.6‑2 in TS 26.512 [16] defines a set of Operation Point parameters, repeated below, to which the Media Player is configured and that it exposes to the 5GMS-Aware Application as status information in notification events at reference point M7 when the current Operation Point changes. This configuration setting may be included in the media presentation manifest, or it may alternatively be provided through private application means. Generally, consistent quality refers to meeting the Operation Point parameters as shown in table 5.11.1.3-1 below . The Media Player and the network are expected to collaborate to meet the quality requirements.

Table 5.11.1.3-1: Operation Point Information (see TS 26.512, Table 13.2.6-2)

|  |  |  |
| --- | --- | --- |
| OperationPoint | Operation Point Parameters | The currently configured operation point parameters according to which the DASH client is operating. |
|  | Mode | Enum | The following operation modes are defined:live: The DASH client operates to maintain configured target latencies using playback rate adjustments and possibly resync.Vod: The DASH client operates without latency requirements and rebuffering may result in additional latencies |
|  | maxBufferTime | Integer | maximum buffer time in milliseconds for the service. |
|  | switchBufferTime | Integer | buffer time threshold below which the DASH clients attempts to switch Representations. |
|  | Latency |  | Defines the latency parameters used by the DASH client when operating in live mode. |
|  |  | Target | Integer | The target latency for the service in milliseconds. |
|  |  | Max | Integer | The maximum latency for the service in milliseconds. |
|  |  | Min | Integer | The maximum latency for the service in milliseconds. |
|  | PlaybackRate | MediaTypeaudio, video, all | Defines the playback rate parameters used by the DASH client for catchup mode and deceleration to avoid buffer underruns and maintaining target latencies. |
|  |  | Max | Real | The maximum playback rate for the purposes of automatically adjusting playback latency and buffer occupancy during normal playback, where 1.0 is normal playback speed. |
|  |  | Min | Real | The minimum playback rate for the purposes of automatically adjusting playback latency and buffer occupancy during normal playback, where 1.0 is normal playback speed. |
|  | Bandwidth |  | Defines the operating bandwidth parameters used by the DASH client used for a specific media type or aggregated. The values are on IP level. |
|  |  | Target | Integer | The target bandwidth for the service in bit/s that the client is configured to consume. |
|  |  | Max | Integer | The maximum bandwidth for the service in bit/s that the client is configured to consume. |
|  |  | Min | Integer | The minimum bandwidth for the service in bit/s that the client is configured to consume. |
|  | PlayerSpecificParameters |  | Player specific parameters may be provided, for example about the used algorithm, etc. |

#### 5.11.1.4 High bandwidth requirements

TV Services typically have certain bit rate requirements that need to be met. Clause 5.1 of TR 26.925 [T] provides an overview. These figures are valid for both HDR/non-HDR video:

|  |
| --- |
| - 720p HD: 2 - 5 MbpsNOTE 1: Today typically 3 Mbps for HEVC [15] and 5Mbps for AVC [14], but bitrate reductions expected with better encoding and coding tools. See Clause 6.1.2.- Full HD: 3 - 12 MbpsNOTE 2: Today typically 5-7 Mbps for HEVC [15] and 10-12 Mbps for AVC [14], but bitrate reductions expected with better encoding and coding tools. See Clause 6.1.2.- 4k UHD: 5- 25MbpsNOTE 3: today typically 8-16 Mbps for HEVC [15] and 15-25 Mbps for AVC [14], but bitrate reductions expected with better encoding and coding tools. See Clause 6.1.2.- 8k UHD: 20 - 80 MbpsNOTE 4: Today typically up to 80 Mbps for HEVC [15], but bitrate reductions expected with better encoding and coding tools. See Clause 6.1.2.NOTE 5: On 8k UHD: currently not specified in 3GPP TS 26.116. Bitrate figures are based on limited available deployment data. |

These figures may be used as indication of bit rate requirements for services.

#### 5.11.1.5 Target latency constraints

Based on a report developed jointly between DVB and DASH-IF on Low-Latency DASH [9], this clause defines details on how to support consistent latency in DASH for linear TV services. In [9], several definitions had been introduced, repeated here for consistency.

*- End-to-End Latency (EEL)*: The latency for an action that is captured by the camera until its visibility on the remote screen.

*- Encoder-Display Latency (EDL)*: The latency of the linear playout output (which typically serves as input to distribution encoder(s)) to the screen.

*- Packager-Display Latency*: The latency after the output of the distribution encoder to the screen.

*- CDN latency*: The delay caused by the CDN delivery from CDN input to CDN output.

*- Live Edge Start-up Delay (LSD)*: The time between a user action (service access or service join) and the time until the first media sample of the service is perceived by the user when joining at the live edge. Typically also the channel change time.

*- Seek Start-up Delay (SSD)*: The time between a user action (service access or service join) and the time until the first media sample of the service is perceived by the user when seeking to a time shift buffer.

Those two categories, latency and delay are subject to be controllable by the service provider for a consistent service offering. In the remainder, primarily the Encoder-Display Latency (EDL) and the Live Edge Start-up Delay are considered, but for some use cases also the End-to-End Latency (EEL) may be relevant. Figure 5.11.1‑1 provides a schematic overview of the different latencies.



Figure 5.11.1‑1: Different latencies and delays relevant for low-latency distribution

The Low Latency DASH scenario is a variant of the Live Services recommended approach focused on ensuring that the Encoder-Display Latency of the DASH Media Presentation is comparable to the latency when distributing over terrestrial, cable or satellite broadcast. Latency in broadcast is not a unique universal value, as it is influenced by many factors such as the duration of the broadcast encoding pipeline, the latency of the transport channel which can slightly differ per type (satellite, cable, IPTV or, DTT...), or the artificial delays introduced by local content moderation regulations. However, most of the measurements converge on a 3 - 10 seconds latency between the moment where the source signal is acquired for encoding and the moment when it's played back on the TVs, i.e the EDL. Start-up delay requirements are typically in the range of 1-2 seconds. For details refer to [9].

As a possible solution that is considered backward-compatible to existing clients is provided in the following. Low-latency mode are supported to minimize the architectural impacts on existing workflows. Figure 5.11.1‑2 provides a basic flow of information for operating a low-latency DASH service as defined in DASH-IF’s Low-latency Modes for DASH [10]. The DASH packager gets information on the general description of the service as well as the encoder configuration. The encoder produces CMAF chunks and fragments. The chunks are mapped by the MPD packager onto Segments and provided to the network in incremental fashion using HTTP/1.1 chunked transfer.



Figure 5.11.1-2 Basic operation flow Low-Latency DASH

HTTP chunked transfer coding needs to be supported up from the ingest into the packager up to the CDN edge, whereas the last mile delivery is expected happen using HTTP chunked transfer coding or HTTP in regular mode. If HTTP chunked transfer coding is supported by the DASH player, it basically means that a media segment carrying the latest moment of the program (also known as the "live edge time" as defined in clause 4 of this document) could be consumed on the player while it's still being produced by the encoder and the packager.

In case chunked segments are used, clients may want to access partially available Segments for example for fast random access, see ISO/IEC 23009-1 [11]. However, requesting available byte ranges of a partially available Segment, i.e., Segments still being produced, is not consistently supported in CDNs, but solutions are provided in RFC8673 [X6]. This functionality may also be needed to support common segment handling for low-latency DASH and low-latency HLS.

Key aspects for low-latency live distribution include:

*-* Consistent support for chunked transfer from ingest to client.

*-* Support for partially access of non-complete resources.

*-* End-to-end optimizations to support the latency requirements.

At the time of writing, DASH-IF is investigating even lower-latency distribution [X] using technologies that are beyond regular segment-based streaming using WebRTC.

As part of this discussion, also a requirement on Audience Drift Gap (ADG) is defined, referring to the time difference between the first user to see a frame of media and the last user to see that same frame of media. This KPI is expected to be monitored and may be controlled and the subject of a requirement.

#### 5.1.11.6 TV Experiences

TR 26.917 [V] studied and identified requirements for TV Experiences. Clause 4.3 of TR 26.917, reproduced below, provides a summary from a user-experience perspective:

|  |
| --- |
| - content delivery up to UHD resolution.- mechanisms to restrict the reception of some or all Subscribed TV services to groups of subscribers (e.g. based on the recipients of the services are subscribers of the MNO, roaming subscribers of other MNOs, or not subscribed to any MNO).- combinations of SD, FHD and UHD resolution TV transport services.- support network broadcast geographic area coverage management considering following criteria:- OTT provider request (including the potential coverage information of TV service information)- available network unicast/broadcast capacity of 3GPP network- number of users under broadcast network coverage- The location information of UE- support audio and video formats typically supported by TV Content Providers for SD and HD TV transport services and UHD TV transport services.- support codecs typically supported by TV Content Providers for HD TV services and UHD TV services.- support accessibility functions typically supported by TV Content Providers (e.g. subtitling, closed captioning, audio descriptions, anonymous reception, reporting to support ratings, reporting enforcement, etc.).- regulatory mandates typically supported by TV Content Providers (blackouts, emergency alerts, etc.).- interactivity functions typically supported by TV Content Providers (interactive services, second screen, personalization, etc.).- ad insertion use cases typically supported by TV Content Providers (targeted ad insertion, ad replacement, etc.),- encryption, security and conditional access functions typically supported by TV Content Providers.- concurrent delivery of multiple application components (TV service application signalling, statistical multiplexing, etc.).- random access and channel change times comparable to existing HD TV services. |

#### 5.1.11.7 Summary of service requirements for Mass grade TV distribution

The following assumptions for the study are captured

- Target for a e2e service max delay of 3 sec (Packager-Display latency). This is measured from the input into the DASH Packager to the output of the DASH Player.

- Target for a tune-in to an ongoing DASH Live stream of max 1.5 sec (Live Edge Startup Delay). This is measured between the user action to tune-in (e.g. pressing a button) and the first frame is output from the DASH Player (still fulfilling the target of the Packager-Display Latency).

- There is a requirement on a maximum ADG referring to the time difference between the first user to see a frame of media and the last user to see that same frame of media. It should be possible to remain below 1 sec apart and within the target latency. max(ADG) = 1 sec.

- The DASH MPD can prepared so that the DASH Player can operate at different Latencies, e.g. low latency, normal latency or timeshift (video on demand profile).

### - TV Experiences as summarized in clause 5.11.1.6.5.11.2 Deployment Architectures

#### 5.11.2.1 Distribution of low-latency media streams

A deployment architecture suitable for low-latency CMAF streaming is shown in Figure 5.11.2.1-1.



Figure 5.11.2.1-1 Deployment architecture for low-latency CMAF streaming

In this case:

1. A live stream is ingested into a live encoder.

2. The encoded stream is packaged into CMAF chunks.

3. The packaged CMAF chunks are uploaded to an origin server using chunked transfer encoding input.

4. Segments are then available for retrieval by a CDN on demand and moved through the CDN all the way to the client.

#### 5.11.2.2 Operation Point – Establishment and Monitoring

This clause deals with providing consistent quality in 5G Media Streaming as part of an Operation Point. Figure 5.11.2.2-1 shows how Operation Points and policies can be matched in a basic setup.



Figure 5.11.2.2-1 Operation Point workflow

The content-defined Operation Points are shown in the User Plane setup in Figure 5.11.2.2-1.

1. A set of Policy Templates is agreed in advance between the 5GMS Application Provider and the 5G System operator. These may be provisioned in the 5GMS AF via reference point M1. Policy templates represent long term agreements made between the 5GMS Application Provider and the 5GMS System operator.

2. Service Operation Points, defined in terms of the parameters reproduced in table 5.11.1.3‑1, are long-lived profiles that will be used by media streaming sessions as references.

3. Based on communication with the 5GMS-Aware Application, the UE device characteristics, the prevailing network conditions and so on, the Media Player selects a suitable Operation Point.

4. Based on the parameters of the currently selected Operation Point, the policies in the 5G network are established, by the Media Session Handler instantiating one of the available Policy Templates by invoking the Dynamic Policies API at reference point M5. The streaming session uses at most one of the allowed policy templates at any point in time.

The above workflow is expected to be operational for 5G Media Streaming as defined today in TS 26.512 [15]. What is needed in order to execute the workflow is the following

1) Well-defined Operation Points in the content. The Service Description as defined in ISO/IEC 23009-1 is mentioned in TS 26.512 [15] as a high-level statement:

|  |
| --- |
| The MPD may contain a one or several **ServiceDescription** elements that include operational parameters. The MPD may also include multiple configurations for the media (different codecs, different content protection, different resolutions, etc.), for example for playback under different operating policies. The handling of this information is documented in clause 13.2. |

but not yet in TS 26.247 [40]. More details are needed, including some examples.

2) Media Session Handling APIs allowing a 5GMS-Aware Application to make use of dynamic policies, network assistance and metrics reporting in the 5GMS Client are not yet specified at reference point M6 in TS 26.512 [15]. In particular, clause 12.2 is incomplete for Dynamic Policy Information and Network Assistance information.

3) Media Player metrics are not yet defined on the Media Streaming Handler API at reference point M7 defined in clause of TS 26.512 [15]. In particular, the extent to which the Media Player is (un)able to obey the currently selected Operation Point parametersneeds to be signalled to the Media Session Handler in order to potentially drive the selection of a different Operation Point.

### 5.11.3 Collaboration Scenarios

The following collaboration scenarios may be considered:

1. Live content is provided to the MNO as an (uncompressed or mezzanine-compressed) contribution feed, and the MNO does the encoding and packaging for distribution. The live content is augmented with Service Operation Points that determine content encoding parameters. Policy Templates may either be generated by the MNO matching the provided Operation Points, or they may be provided by the content provider. If the latter, these are expected to match the Service Operation Points. Alternatively, Service Operation Points may be derived by the MNO based on a set of Policy Templates previously agreed with the content provider.
2. Live content is encoded and packaged by the content provider (for example as low-latency CMAF) and published to the MNO. The MNO may produce an MPD for its distribution. However, the content provider augments the media with production and/or encoding timestamps (e.g. producer reference times) in order to permit latency measurements. The content provider augments the published content with service Operation Point metadata mapping CMAF structures to Service Operation Points. Policy Templates may either be generated by the MNO matching the required Operation Points, or the content provider may supply them. If the latter, these are expected to match the Service Operation Points. Alternatively, Service Operation Points may be derived by the MNO based on a set of Policy Pemplates previously agreed with the content provider.
3. The content origin server is deployed outside the MNO network and content is pulled through the 5GMS AS on demand by the clients. The Service Operation Points are defined explicitly in the content presentation manifest. The Policy Templates are either derived from these Service Operation Points or else provided by the MNO to match the Service Operation Points described in the presentation manifest.

### 5.11.4 Mapping to 5G Media Streaming and High-Level Call Flows

#### 5.11.4.1 General: Distribution of “Operation Point Services”

This clause provides an extension to the general call flow in clause 6.2.3 of TS 26.501 [15] in order to address operation point services.



Figure 5.11.4.1-1: High Level Procedure for DASH content for Operation Point Handling

Prerequisites:

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

- The 5GMSd-Aware Application has received the Service Announcement from the 5GMSd Application Provider.

Extended Steps:

1: Policy Templates are defined

12: Media Player informs application about the current set of Operation Points.

13: 5GMSd-Awaer Application selects an Operation Point.

14: Media Player provdes Operation Point parameters to the Media Session Handler.

15: Media Session Hahdnler selects a Dynamic Policy based on the provided Operation Point parameters.

21: Media Player provides Operation Point metrics to the Media Session Handler.

22: Media Session Handler sends Operation Point measurements and events to the 5GMSd AF.

#### 5.11.4.2 Collaboration 1: MNO provides encoding and packaging

In this case, the specific aspects are as follows:

1. A native stream is ingested into the network. This may be an MPEG-2 TS stream, an RTMP stream, etc. The stream may have time codes included that relate media time to real-time.
2. Along with the ingest, a provisioning agreement between the content provider and the MNO exists on one or several Service Operation Points and Policy Templates. It may be that only one of the two is defined, and the other one is derived, or both are defined.
3. The MNO encodes, packages and distributes the content according to the agreed Service Operation Points. The encoding and packaging may take into account the information from the Operation Points but may also take into account the information in the ingested media stream in order to control the latency.
4. The Operation Point metrics collated by the 5GMS AF are used by the MNO to validate that the Service Operation Point is met, or to adjust the Policy Templates accordingly.

#### 5.11.4.3 Collaboration 2: MNO provides DASH distribution

In this case, the specific aspects are as follows:

1. CMAF content is provided externally. The content is announced to the MNO for distribution.
2. The ingest happens in a way such that the latency requirements can be met.
3. A provisioning agreement between the content provider and the MNO exists on one or several Service Operation Points and Policy Templates. It may be that only one of the two is defined, and the other one is derived, or both are defined.
4. The MNO distributes the content according to the agreed Service Operation Points, i.e. meeting bit rate and latency requirements. The MNO generates the media presentation manifests accordingly, for example the MPD and/or the HLS manifest.
5. The Operation Point metrics collated by the 5GMS AF are used by the MNO to validate that the Service Operation Point is met, or to adjust the Policy Templates accordingly.

#### 5.11.4.4 Collaboration 3: MNO acts as CDN

In this case, the specific aspects are as follows:

1. DASH or HLS content is provided externally. The content is announced to the MNO for distribution.
2. The ingest happens in a way such that the latency requirements can be met.
3. A provisioning agreement between the content provider and the MNO exists on one or several Servcie Operation Points and Policy Templates. It may be that only one of the two is defined, and the other one is derived, or both are defined.
4. The MNO distributes the content according to the agreed Service Operation Points, i.e. meeting bit rate and latency requirements.
5. The Operation Point metrics collated by the 5GMS AF are used by the MNO to validate that the Service Operation Point is met, or to adjust the Policy Templates accordingly.

#### 5.11.4.5 Operation Point example

An example for an operation point is provided below:

- Mode = Live.

- maxBufferTime = 4000.

- switchBufferTime = 500.

- Latency: target=3000, maximum=5000, minimum=3000.

- PlaybackRate: max=1.2, min=0.8.

- Bit rate: target = 4000000, max=6000000, min=1000000.

### 5.11.5 Potential open issues

The following aspects are not yet addressed in the current 5G Media Streaming specifications:

* On M1d:

- The establishment of consistent provisioning for Live TV services of different scale, providing the required Service Operation Points.

- The ability to support Collaboration 1, for which the MNO provides encoding and packaging.

* On M2d:

- Ingest protocols that support chunk-based ingest and do not add latency by the use of chunked transfer. (No aggregation of chunks into resources.) Note, support for HTTP Chunked Transfer encoding is a requirement of HTTP 1.1 Clients and Servers. The gap is, that an 5GMS AS implementation is free to wait until the full segment is received.

- Signaling of Operation Points along with the content in Collaboration 2.

* On M4d:

- Consistent support of low-latency streaming formats.

- Time synchronisation between server and client so that performance (particularly latency) can be measured accurately.

- Support for advanced TV Experiences.

* On M5d:

- Metrics to monitor the operation of the service and if it meets operation, in particular the latency.

* On M6d:

- Dynamic policies, network assistance and metrics reporting.

* Other:

- Support TV service requirements.

### 5.11.6 Candidate Solutions

For the above open issues, the following candidate solutions are considered:

* On M1d:

- Policy Template updates to support TV services.

- Provisioning extension to support Collaborations 1, 2 and 3.

* On M2d:

- DASH-IF Ingest Specification: <https://dashif-documents.azurewebsites.net/Ingest/master/DASH-IF-Ingest.html>

- MPD extensions to support signalling of Operation Points using Service Description.

* On M4d:

- DASH-IF Low-Latency Extensions: <https://dash-industry-forum.github.io/docs/CR-Low-Latency-Live-r8.pdf>

- UTC Time Sync that can be used by the 5GMS AS and the 5GMS Client in order to measure latency accurately. A 3GPP-based network time source may be provided.

- New DASH and CMAF functionalities that support enhanced TV services, such as pre-selection.

* On M5d:

- Updates to DASH QoE metrics reporting for monitoring latency and Audience Drift Gap.

* On M6d:

- Extensions to M6 to address the requirements.

* General support:

- DASH-IF Low-Latency Extensions: https://dash-industry-forum.github.io/docs/CR-Low-Latency-Live-r8.pdf