**Agenda item:** 8.6

**Source:** Nokia

**Title: [FS\_AMD] WT#12: Congestion information for Media Delivery**

**Document for** Discussion andAgreement

# Introduction

In this contribution, we propose to introduce more details of congestion information and discuss the usage of those congestion information in the context of advanced media delivery.

# Exposure of congestion information for Advanced Media Delivery

In current downlink media streaming process, while using congestion-oriented protocols such as L4S, the 5GMS client receives congestion information and send the feedback (ECN echo) to the 5GMS AS. In case of congestion status, 5GMS Client may take actions, e.g., changing the requested representation. But actions from the 5GMS AS side is now missing. For example, the 5GMS AS may remove some high-bitrate representations from the MPD requested by new 5GMS clients if they are located behind a congested network path. Alternatively, a 5GMS AS may provide an adapted results regarding the 5GMS client request for high-resolution segment by temporally serving them lower resolution instead.

It would be beneficial to share or expose congestion information with the 5GMS client and 5GMS server, so that both can detect early signs of congestion, and act accordingly to improve the media delivery process.

# Sharing congestion information within current architectures

As described in KI#12, QoS support for Media Streaming services has been introduced since Rel-16. For example, the dynamic policy feature is introduced to request specific QoS handling, and the network assistance feature is introduced to get aware of the network status. New QoS enhancements and the network information exposure have been introduced in recent releases, which could be useful for Media Streaming services.

# 3.1 Via ECN marking for L4S

Referring to S4al240190 (endorsed at SA4 MBS Telco), “according to clause 6.1.3.22 of TS 23.503 [41], an Application Function may provide an explicit indication that the uplink and/or downlink path of a service data flow supports ECN marking for L4S by means of the *Nnef\_AFsessionWithQoS* service at reference point N33 or the *Npcf\_PolicyAuthorization* service at reference point N5.”

**Congestion information (including percentage of congestion level for exposure, congestion trend (e.g. congestion increasing or not), etc.) can be populated by the same approach in the 5G Media Delivery architecture. Furthermore, the congestion information can be used by 5GMS AS to facilitate the session establishment for new 5GMS Clients, for example, if significant congestion is detected, the 5GMS AS may, for example, update the DASH presentation manifest to remove/add one or more representations.**

**This feature should be beneficial to improve the general procedure of media delivery and would be implemented consistently as part of 5G Media Delivery.**

# 3.2 Via QoS monitoring

Also referring to S4al240190, QoS monitoring comprises of measurements of QoS monitoring parameters and reports of the measurement result for a service data flow (i.e., QoS Flow) and can be enabled based on 3rd party application requests and/or operator policies configured in the 5GC (i.e., PCF).

The AF may request measurements and subscribe to the event for one or more of the following QoS monitoring parameters by means of the Nnef\_AFsessionWithQoS service at reference point N33 or the Npcf\_PolicyAuthorization service at reference point N5, which may trigger QoS monitoring for service data flow(s):

- Uplink packet delay, downlink packet delay and round-trip packet delay for a service data flow (see clause 5.45.2 of TS 23.501 [23]).

- Congestion (see clause 5.45.3 of TS 23.501 [23]).

- Data Rate (see clause 5.45.4 of TS 23.501 [23]).

- Packet Delay Variation (see clause 5.37.7 of TS 23.501 [23]).

- Round-trip packet delay considering the uplink path of one service data flow and the downlink path of another service data flow (see clause 5.37.4 of TS 23.501 [23]).

**According to TS 23.501 clause 5.45.3, NG-RAN is capable to measure and provide congestion information (i.e., a percentage of congestion level for exposure) as requested. Congestion information (including percentage of congestion level for exposure, congestion trend (e.g. congestion increasing or not), etc.) could re-use the same approach, and it would be implemented consistently with current QoS monitoring procedures.**

# Proposed updates to potential call flows

This contribution proposes updates based on high level call flows provided in S4al240190:

\* \* \* 1st Change \* \* \* \*

##### 5.23.1.2.1 Support of ECN marking for L4S

As described in RFC 9330 [X1], RFC 9331 [X2] and RFC 9332 [X3], the purpose of ECN marking for L4S (Low Latency, Low Loss and Scalable Throughput) is to inform a recipient host at the earliest opportunity that an IP packet has experienced network congestion at some point in its routing path. It exposes congestion information by marking ECN bits in the IP header of the user IP packets between the UE and the Application Server. This early notification may be used by the receiving application to report the congestion to its sending peer using a suitable Layer 4 feedback mechanism. Based on this feedback, the sender should reduce the sending bit rate. In the context of adaptive segmented media delivery, application layer rate adaptation may be needed in addition.

- For example, a media player consuming a media presentation that receives an ECN-marked downlink packet from a streaming media server may induce a reduction in the sender’s bit rate by switching to a representation of a lower bit rate. To support this functionality, the recipient host needs to support L4S feedback as described in RFC 9330 [X1].

- Furthermore, early notification of congestion reflected back to an Application Server may be used to improve the current delivery process or to facilitate session establishment for new clients. For example, if significant congestion is detected, the 5GMS AS may update the DASH presentation manifest to remove/add one or more representations.

\* \* \* 2nd Change \* \* \* \*
Call flows for L4S-on-request

#### 5.23.4.3 L4S-on-request for downlink media streaming

An Application Function may request L4S support from the 5G Network for a certain QoS Flow, e.g., by invoking the Nnef\_AfsessionWithQoS service. The concept of this solution is that an application only requests L4S support from the network when the application layer is able to adjust its load on the network depending on the L4S feedback (e.g. the percentage of congestion marks) it receives. The activation leverages the existing 5GMS Dynamic Policy invocation API, allowing the 5GMS-Aware Application to request L4S support as and when it is needed.

A high-level call flow for downlink media streaming is sketched in figure 5.23.4.3-1 below. The following is assumed:

- The service here is a unicast downlink media streaming service with dynamic policy support, as described in clause 5.7 of TS 26.501 [15].

- The Layer 4 protocol used for application flows is TCP and the TCP stack used supports L4S.

- The network supports L4S packet marking.

- The application has specifically requested ECN marking for its media delivery session.

- NG-RAN manipulates the ECN bits (per clause 5.37.3.2 of TS 23.501 [23]). It is equally possible that the PSA-UPF manipulates the ECN bits (per clause 5.37.3.3 of [23]).



Figure 5.23.4.3-1: Downlink media streaming call flow for L4S on request

The steps are as follows:

0: *Policy Template Provisioning.* A Policy Template is provisioned **with the requirement for L4S capability indicated by setting a flag**.

1: *Dynamic Policy activation.* The Media Session Handler within the 5GMSd Client obtains Service Access Information and triggers a dynamic policy activation. A Policy Template Binding is present within the Service Access Information for each provisioned Policy Template. **Policy Template Bindings suitable for L4S are indicated by an L4S capability requirement flag being set. The 5GMSd Client detects that an L4S-capable media transport stack is present and in use. The selected Policy Template is one configured with L4S capability.**

2: *QoS request.* The 5GMSd AF requests QoS handling using e.g. the Nnef\_AfSessionWithQoS service or the Npcf\_PolicyAuthorization service. **If the L4S capability requirement flag is set in the selected Policy Template, this indicates that the new QoS flow is required to be L4S-enabled.** The new QoS flow with the L4S indication setting propagates through the 5G System.

3: **If the L4S capability requirement flag is set in the Policy Template Binding for the selected Policy Template, the 5GMSd Client selects/enables the L4S capability of the used transport protocol.**

NOTE: This step may happen implicitly by selecting an L4S-supporting transport protocol stack.

4: The Media Player within the 5GMSd Client triggers the establishment of a TCP connection. The ECT(1) codepoint is set in the IP header, indicating an L4S-Capable Transport, and the SDAP entity ensures that the packet is forwarded via the matching QoS flow.

5: The 5GMSd AS responds to the TCP connection establishment request. The 5GMSd AS sets ECT(1) in the IP headers, indicating an L4S-Capable Transport.

6: The UPF finds the matching QoS flow identifier for the downlink packet and sends the packet via the according QoS flow to the UE. TCP Connection setup continues, with one ECT bit set in all packets.

7. When the RAN detects an upcoming congestion according to the congestion measurement (based on continuous congestion monitoring), the 5G System sets the CE (Congestion Experienced) codepoint in the IP header of the downlink packet.

**8. (Optional) The Media Player may generate a Congestion Notification including, for example, the percentage of congestion marks and the congestion trend, and notifies the Media Session Handler using the Network Assistance client API.**

**9. (Optional) The Media Player sends the congestion information to the 5GMS AS in a congestion report.**

10. **The** **TCP protocol stack used by the Media Player in the 5GMSd Client reflects the Early Congestion Notification to the TCP sender by setting the ECN-Echo (ECE) flag in the TCP header of an uplink PDU of the same TCP connection. The TCP sender reacts to the ECN-Echo accordingly (i.e., by reducing its sending congestion window).**

NOTE 1: The ECN-Echo flag is also acknowledged by the TCP sender setting the Congestion Window Reduced (CWR) flag in an outgoing TCP frame, but this acknowledgement is not illustrated in this call flow.

NOTE 2: Classic ECN [X6] requires an ECN signal to be treated as equivalent to a packet drop. L4S [X1] specifies a more fine-grained response and an early congestion signal triggers a less severe reaction. How a TCP sender behaves “accordingly” is not in scope of the specification.

11. Based on the CE indication received in step 7, or by detecting a reduced bit rate in the downlink application flow, the Media Player in the 5GMSd Client may react by, for example, changing the requested representation.

12. Based on the congestion report received in step 9, or the CE indication received in step 10, or detection of a reduced bit rate in the downlink application flow, the 5GMSd AS may react by, for example, updating the **DASH presentation** manifest to remove/add one or more representations.

#### 5.23.4.4 L4S-on-request for uplink media streaming

Support for uplink media streaming is very similar to that for downlink media streaming.

A high-level call flow for uplink media streaming is sketched in figure 5.23.4.3-1 below. The following is assumed:

- The service here is a unicast uplink media streaming service with dynamic policy support, as described in clause 6.9 of TS 26.501 [15].

- The Layer 4 protocol used for application flows is TCP and the TCP stack used supports L4S.

- The network supports L4S packet marking.

- The application has specifically requested ECN marking for its media delivery session.

- NG-RAN manipulates the ECN bits (per clause 5.37.3.2 of TS 23.501 [23]). It is equally possible that the PSA-UPF manipulates the ECN bits (per clause 5.37.3.3 of [23]).



Figure 5.23.4.4-1: Uplink media streaming call flow for L4S on request

The steps are as follows:

0: *Policy Template Provisioning.* A Policy Template is provisioned **with the requirement for L4S capability indicated by setting a flag**.

1: *Dynamic Policy activation.* The Media Session Handler within the 5GMSu Client obtains Service Access Information and triggers a dynamic policy activation. A Policy Template Binding is present within the Service Access Information for each provisioned Policy Template. **Policy Template Bindings suitable for L4S are indicated by an L4S capability requirement flag being set. The 5GMSu Client detects that an L4S-capable media transport stack is present and in use. The selected Policy Template is one configured with L4S capability.**

2: *QoS request.* The 5GMSu AF requests QoS handling using e.g. the Nnef\_AfSessionWithQoS service or the Npcf\_PolicyAuthorization service. **If the L4S capability requirement flag is set in the selected Policy Template, this indicates that the new QoS flow is required to be L4S-enabled.** The new QoS flow with the L4S indication setting propagates through the 5G System.

3: **If the L4S capability requirement flag is set in the Policy Template Binding for the selected Policy Template, the 5GMSu Client selects/enables the L4S capability of the used transport protocol.**

NOTE: This step may happen implicitly by selecting an L4S-supporting transport protocol stack.

4: The Media Streamer within the 5GMSu Client triggers the establishment of a TCP connection. The ECT(1) codepoint is set in the IP header, indicating an L4S-Capable Transport, and the SDAP entity ensures that the packet is forwarded via the matching QoS flow.

5: The 5GMSu AS responds to the TCP connection establishment request. The 5GMSu AS sets ECT(1) in the IP headers, indicating an L4S-Capable Transport.

6: The UPF finds the matching QoS flow identifier for the downlink packet and sends the packet via the according QoS flow to the UE. TCP Connection setup continues, with one ECT bit set in all packets.

7. When the RAN detects an upcoming congestion according to the congestion measurement (based on continuous congestion monitoring), the 5G System sets the CE (Congestion Experienced) codepoint in the IP header of the uplink packet.

8. The TCP protocol stack used by the 5GMSu AS reflects the Early Congestion Notification to the TCP sender by setting the ECN-Echo (ECE) flag in the TCP header of a downlink PDU of the same TCP connection. The TCP sender reacts to the ECN-Echo accordingly (i.e., by reducing its sending congestion window).

NOTE 1: The ECN-Echo flag is also acknowledged by the TCP sender setting the Congestion Window Reduced (CWR) flag in an outgoing TCP frame, but this acknowledgement is not illustrated in this call flow.

NOTE 2: Classic ECN [X6] requires an ECN signal to be treated as equivalent to a packet drop. L4S [X1] specifies a more fine-grained response and an early congestion signal triggers a less severe reaction. How a TCP sender behaves “accordingly” is not in scope of the specification.

9. Based on the CE indication received in step8, or by detecting a reduced bit rate in the uplink application flow, the Media Streamer in the 5GMSu Client may react by, for example, changing the requested representation.

10. Based on the CE indication received in step 7, or by detection of a reduced bit rate in the uplink application flow, the 5GMSu AS may react by, for example, updating the DASH presentation manifest to remove/add one or more representations.

\* \* \* 3rd Change \* \* \* \*
Gap analysis and requirements for L4S-on-request

5.23.5.3 L4S-on-request for downlink and uplink media streaming

Based on the call flow in clause 5.23.4.3, the following observations are made:

- L4S/ECN does not require modifications to the Media Player or a TCP based Media Streamer.

- The 5GMS AF needs to explicitly request L4S handling of packets by the 5G System by interacting with the PCF at reference point N5 (or else via the NEF at reference poiont N33).

- The Policy Template resource structure at reference point M1 needs to be extended to include an L4S capability requirement flag.

- The Policy Template Binding data structure carried in the Service Access Information resource at reference point M5 needs to be extended to reflect the value of the L4S capability requirement flag in the corresponding Policy Template.

- An L4S-capable transport protocol stack is required in both the 5GMS Client and at the 5GMS AS.

NOTE 1: When the transport protocol stack used on the UE or the Application Server does not support ECN marking, the ECT flags are set accordingly to explicitly indicate lack of support.

NOTE 2: Application processes such as the Media Player and 5GMS AS are not currently capable of obtaining L4S Congestion Notifications from the underlying Operating System. Future studies on API extensions to facilitate the usage of L4S are expected to bridge this gap.

- The Network Assistance client API needs to be extended to convey a new Congestion notification including, for example, the percentage of congestion marks and the congestion trend.

NOTE 3: The candidate solutions proposed below do not address this gap pending resolution of the gap identified in NOTE 2 above.

- The Media Session Handling API needs to be extended to specify the means to convey congestion information from the Media Session Handler to the 5GMS AS in a congestion report.

NOTE 4: The candidate solutions proposed below do not address this gap pending resolution of the gap identified in NOTE 2 above.

- Depending on the transport stack implementation, an explicit L4S activation is required at session start.

\* \* \* End Changes \* \* \* \*

# Proposal

We propose to document the content of section 4 in KI#12 as a modification of TR 26.804 **CR0007**, on adding congestion trend and status as parts of the network information exposure for 5G Media Delivery.