**3GPP TSG-SA4 Meeting #130 *S4-241884***

**Orlando, United States, 18th Nov 2024 - 22nd Nov 2024 revision of S4aI240154**

|  |
| --- |
| *CR-Form-v12.3* |
| **CHANGE REQUEST** |
|  |
|  | **26.804** | **CR** | **0017** | **rev** | **3** | **Current version:** | **18.1.0** |  |
|  |
| *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\_AMD] Inband Signaling of QoS for 5G Media Streaming |
|  |  |
| ***Source to WG:*** | Qualcomm Germany, Lenovo |
| ***Source to TSG:*** | S4 |
|  |  |
| ***Work item code:*** | FS\_AMD |  | ***Date:*** | 2024-11-11 |
|  |  |  |  |  |
| ***Category:*** | **B** |  | ***Release:*** | Rel-19 |
|  | *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-17 (Release 17)Rel-18 (Release 18)Rel-19 (Release 19) Rel-20 (Release 20)* |
|  |  |
| ***Reason for change:*** | Inband QoS signaling is a means to quickly adapt to needs of the network without having to rely on reactiveness by the client or on throttling of traffic in the network.As an example, at recent IETF meetings, the issue of Secure Communication of Network Properties (SCONE-PRO) had been discussed. It was highlighted in several inputs that bandwidth is and remains a scarce resource, and that video is and will stay the dominant from of media on the Internet. Despite continuous capacity investments, it is hard to keep up with demand needed for video delivery. The Secure Communication of Network Properties (SCONEPRO) Working Group's primary objective is to specify a 'maximum achievable throughput' property for QUIC-based streaming video and an on-path protocol for securely communicating this property from a network device to a client endpoint.As 5G Media Streaming provides functionalities that are similar and aligned to the objectives of the IETF work, this key issue will address aspects that investigate how the requirements from SCONE-PRO are met by 5G Media streaming, to what extent SCONE-PRO can be combined with 5G Media Streaming and extensions to 5G Media Streaming would be suitable to address combination with SCONE-PRO. |
|  |  |
| ***Summary of change:*** | Addresses the work item objectives for this key issue* Documents the key issue in more detail, in particular how they relate to the 3GPP Media Delivery architecture and/or the MBS User Service architecture
* Studies collaboration scenarios between the Application Service Provider and the 5G System and for each of the key topics.
* Based on existing architectures, provides one or more deployment architectures that address the key topics and the collaboration models.
* Maps the key topics to basic functions and develop high-level call flows.
* Identifies the issues that need to be solved.
* Provides candidate solutions including call flows, protocols and APIs for each of the identified issues.

Identifies gaps and recommend potential normative work for stage-2 and stage-3, including which existing specifications would be impacted and/or if any new specifications would preferably be developed. |
|  |  |
| ***Consequences if not approved:*** |  |
|  |  |
| ***Clauses affected:*** | 2, 5.25 (new), 6.25 (new) |
|  |  |
|  | **Y** | **N** |  |  |
| ***Other specs*** |  | **X** |  Other core specifications  | TS/TR ... CR ...  |
| ***affected:*** |  | **X** |  Test specifications | TS/TR ... CR ...  |
| ***(show related CRs)*** |  | **X** |  O&M Specifications | TS/TR ... CR ...  |
|  |  |
| ***Other comments:*** |  |
|  |  |
| ***This CR's revision history:*** | This document merges S4-242047 |

## ===== CHANGE =====

## 2 References

[SCONE-PRO] Secure Communication of Network Properties (sconepro), <https://datatracker.ietf.org/group/sconepro/about/>.

[CMSD] CTA-5006: Web Application Video Ecosystem - Common Media Server Data

[SCONE] Standard Communication with Network Elements (scone), https://datatracker.ietf.org/wg/scone/about/

## ===== CHANGE (updates against latest submssion) =====

## 5.25 Inband Signaling of QoS for 5G Media Streaming

### 5.25.1 Description

#### 5.25.1.1 General

Video traffic is 70% of the overall traffic volume on the Internet and is expected to grow to 80% by 2028. Across developed and emerging markets video traffic forms 50-80% of traffic volume on mobile networks. New formats like short form videos have seen tremendous growth in recent years. These growth trends are likely to increase with new populations coming online on mobile-first markets.

Local access network conditions may constrain the maximum throughput for a given client, or be so volatile as to rapidly change the maximum throughput throughout the course of a session. In addition, despite capacity augmentation work such as deployment of new generations or new bands of spectrum, capacity augmentation efforts are not keeping pace with growth in demand. Hence, network operators have found it faster and less expensive to invest in shaping (also called throttling) of video traffic on a per-flow basis, which negatively affects video stream quality. However, network operators cannot explicitly measure the degradation to end user quality of experience (QoE) caused by traffic shaping, making this approach open loop.

Video traffic usually employs adaptive bitrate (ABR) schemes to dynamically adjust the video quality (and thus the data rate) in response to changing network conditions. Ideally, when a network operator performs traffic shaping, the ABR scheme should adapt the video quality in use to reflect the data rate allowed by shaping, and converge on a bitrate allowed by the shaper. In practice this convergence is extremely difficult to achieve while maintaining a good user experience. Application providers are even designing algorithms to detect the presence of such traffic shapers and estimate the targeted shaping rate, however, these algorithms are likely to be both inaccurate and complex.

Instead, it would be beneficial, for both the application provider and network operator, to signal network attributes to the application/media player to self-adapt its video traffic to conform to the specified characteristics. The application provider has the ability to measure end user QoE and therefore can self-adapt with QoE feedback.

Existing technologies and standardization efforts in the context of support inband signaling are provided below and discussed in more details.

#### 5.25.1.2 Secure Communication of Network Properties (SCONE-PRO)

At recent IETF meetings, the issue of Secure Communication of Network Properties (SCONE-PRO) [SCONE-PRO] had been discussed. It was highlighted in several inputs that bandwidth is and remains a scarce resource, and that video is and will stay the dominant from of media on the Internet. Despite continuous capacity investments, it is hard to keep up with demand needed for video delivery. In the following, a few key issues are highlighted that motivate the work with references to the material:

- ABR Video Shaping: <https://datatracker.ietf.org/meeting/119/materials/slides-119-sconepro-how-networks-shape-traffic-02>. This presentation introduces ABR video shaping, for which nowadays deep packet inspection and heuristics methods are used to throttle the video flow with a shaper or policer. It also addresses the downsides of policing and shaping and points to the lack of interoperability.

- How YouTube™ coordinates with some MNOs: <https://datatracker.ietf.org/meeting/119/materials/slides-119-sconepro-youtube-plan-aware-streaming-01>. This presentation provides insights how YouTube coordinates with some MNOs. An API exists documenting the maximum media rate, provided out-of-band from operator to service provider, and updates to this value be provided. The max bitrate is not exceeded by the format, but at the same time no policing/shaping is applied. The resulting reduced rates reduce costs and improve user experience.

- SCONE-PRO Problem Statement: <https://datatracker.ietf.org/meeting/120/materials/slides-120-sconepro-problem-statement-00>. The presentation also again highlights traffic shaping issues, including

- ABR schemes are not perfect and don’t converge quickly, causing poor user experience and stalling as it “ping pong” between qualities.

- Congestion Controllers are better suited to simple queueing and often make the “ping ponging” worse.

- CC's (and ABR's) bandwidth estimation often overshoot significantly due to the burst allowance of the TBF.

- The limit imposed by the TBF is artificial – it can support instantaneously more bandwidth, leading to periods of underutilization and difficulty for radio equipment to optimize spectrum usage.

The document further indicates that there are benefits that the video content provider receives maximum instantaneous throughput property from the network, while the shaper is removed or “dialed back”. This would result to move from a congestion-limited approach to an application-limited approach.

- An initial draft charter was provided in <https://datatracker.ietf.org/meeting/120/materials/slides-120-sconepro-draft-charter-v1-july-20-00>

- Video traffic is 70% of the overall traffic volume on the Internet and is expected to grow to 80% by 2028. Across developed and emerging markets video traffic forms 50-80% of traffic volume on mobile networks. New formats like short form videos have seen tremendous growth in recent years. These growth trends are likely to increase with new populations coming online on mobile-first markets.

- Local access network conditions may constrain the maximum throughput for a given client, or be so volatile as to rapidly change the maximum throughput throughout the course of a session. In addition, despite capacity augmentation work such as deployment of new generations or new bands of spectrum, capacity augmentation efforts are not keeping pace with growth in demand. These network operators have found it faster and less expensive to invest in shaping (also called throttling) of video traffic on a per-flow basis, which negatively affects video stream quality. This is done for both network management and business motivations. Network operators cannot explicitly measure the degradation to end user quality of experience (QoE) caused by traffic shaping, making this approach open loop.

- Video traffic usually employs adaptive bitrate (ABR) schemes to dynamically adjust the video quality (and thus the data rate) in response to changing network conditions. Ideally, when a network operator performs traffic shaping, the ABR scheme should adapt the video quality in use to reflect the data rate allowed by shaping, and converge on a bitrate allowed by the shaper. In practice this convergence is extremely difficult to achieve while maintaining a good user experience. Application providers are even designing algorithms to detect the presence of such traffic shapers and estimate the targeted shaping rate, however, these algorithms are likely to be both inaccurate and complex. Instead, it would be beneficial, for both the application provider and network operator, to signal network attributes to the application to self-adapt its video traffic to conform to the specified characteristics. The application provider has the ability to measure end user QoE and therefore can self-adapt with QoE feedback.

- The Secure Communication of Network Properties (SCONEPRO) Working Group's primary objective is to specify a 'maximum achievable throughput' property for QUIC-based streaming video and an on-path protocol for securely communicating this property from a network device to a client endpoint.

- Core solution characteristics are documented including:

- Flow associativity. The network communicates applicable properties as they relate to specific QUIC connections. This ensures that applications can authorize and apply actions on a per-QUIC connection basis.

- Single communication channel for both client initiation and network properties. The communication channel is initiated by a client device, just as the end-to-end application flows are also typically initiated by a client. The same communication channel is used to provide network properties to the client.

- Network properties sent from the network. The network provides the properties to the client. The client might communicate with the network but won't be providing network properties.

- On-path establishment. That is, no off-path element is needed to establish the communication channel between the entity communicating the properties and the client.

- Optionality. The communication channel is strictly optional for the functioning of application flows. A client's application flow must function even if the client does not establish the channel.

- Properties are not directives. A client is not mandated to act on properties received from the network, and the network is not mandated to act in conformance with the properties.

[As 5G Media Streaming provides functionalities that are similar and aligned to the objectives of the IETF work, this key issue will address aspects that investigate how the requirements from SCONE-PRO are met by 5G Media streaming, to what extent SCONE-PRO can be combined with 5G Media Streaming and extensions to 5G Media Streaming would be suitable to address combination with SCONE-PRO.

#### 5.25.1.3 Standard Communication with Network Elements (SCONE)

The work on Standard Communication with Network Elements (SCONE) follows on the SCONE-PRO work introduced in clause 5.25.1.2.

SCONE-PRO BoF sessions led to the creation of IETF SCONE WG in Web and Internet Transport with the first WG meeting taking place in November 2024 at IETF 121. SCONE WG charter maintains some of the objectives of the SCONE-PRO BoF sessions in order to support rate-adaptive applications deliver optimal user experience based on available network conditions for a given network UDP 4-tuple. In particular, as per SCONE WG charter, the following objectives are in scope:

- establish a mechanism for network elements capable of rate-limiting a UDP 4-tuple to communicate an upper bound on achievable bitrate, termed "throughput advice", to the sender of packets matching the UDP 4-tuple.

- allow an application through the mechanism to receive notifications containing throughput advice for both upstream and downstream traffic from any network elements capable of dropping or delaying packets on the path of a UDP 4-tuple

- enable the throughput advice as a guideline to enhance user experience given maximum bitrate manageable by a single network element for that user's current connection. The throughput advice is not a strict indicator of network congestion as is intended for adaptive bitrate applications and is not a replacement for congestion control algorithms.

- enable potential dynamic updates to the throughput advice by the network elements

- determine whether it is necessary for an endpoint to explicitly signal its capability of receiving throughput advice, and whether it is necessary for an endpoint to confirm its receipt of throughput advice.

The SCONE WG will focus initially on a solution for QUIC transport with a milestone to submit a standard track protocol to communicate "throughput advice"— from network elements to the endpoint to the IESG for publication within November 2025 timeline.

#### 5.25.1.4 Common Media Server Data (CMSD)

Common Media Server Data (CMSD) [CTA-5006] as introduced in Annex X.2 may be a candidate technology to signal inband QoS on application layer. Certain parameters may beneficially apply to support inband QoS, for example CMSD-Dynamic currentBitrate, bufferLevel, playbackPosition, or throughputEstimate.

However, Common Media Server Data is defined at the HTTP layer and it is not clear if it can be used by elements traversed in the network that operate on a lower level in the protocol stack and do not understand HTTP. This may make CMSD more suitable for example application server (AS), but not for elements in the network such as routers with rate limiting functionality. Another disadvantage of inband signalling of QoS using CMSD would be that it can only be applied to 5GMS streaming traffic and cannot be used generically for all types of traffic that may have different semantics.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Also, CMSD data is tightly linked to the media content and server-client connection. Throughput estimates, bufferlevel are linked to a single client and media presentation and do not apply generically to a link. It is for study if more generic information about the network connection is needed for inband QoS signalling.

### 5.25.2 Collaboration scenarios

This is for further study.

### 5.25.3 Architecture mapping

This is for further study.

### 5.25.4 High-level call flow

This is for further study.

### 5.25.5 Gap analysis and requirements

This is for further study.

### 5.25.6 Candidate solutions

This is for further study.

### 5.25.7 Summary and conclusions

Support of inband QoS signalling is for further study.

## ===== CHANGE =====

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |