**3GPP TSG-SA4 Meeting #130 *S4-241971***

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

|  |
| --- |
| *CR-Form-v12.3* |
| **CHANGE REQUEST** |
|  |
|  | **26.804** | **CR** | **0019** | **rev** | **5** | **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 | **x** | Radio Access Network |  | Core Network |  |

|  |
| --- |
|  |
| ***Title:***  | [FS\_AMD] WT #13: New clause 5.24 QUIC-based Media Delivery |
|  |  |
| ***Source to WG:*** | Xiaomi |
| ***Source to TSG:*** | S4 |
|  |  |
| ***Work item code:*** | FS\_AMD |  | ***Date:*** | 2024-11-12 |
|  |  |  |  |  |
| ***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:*** | FS\_AMD includes the topic about opportunities with QUIC for segmented streaming (topic “m)”). The corresponding new clause “QUIC-based segmented media delivery” needs to be developped. |
|  |  |
| ***Summary of change:*** | Compared to the previous CR revision with TDoc n. S4aI240183, this one adds a figure illustrating the high-level protocol stack and a reference to a paper presenting a streaming approach using WebTransport. |
|  |  |
| ***Consequences if not approved:*** |  |
|  |  |
| ***Clauses affected:*** | 5.24 [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:*** | * CR-19rev2: [N/A]
* CR-19rev3: Populate clause 5.24
* CR-19rev4: Update clause 5.24.1 and Figures
* CR-19rev5: Added high-level protocol stack Figure
 |

Change #1

## 2 References

…

[2] Akamai Blog, “A QUICk Introduction to HTTP/3”, April 2020, <https://www.akamai.com/blog/developers/a-quick-introduction-http3>

…

[5] “”IETF IETF RFC 9114: "HTTP/3", June 2022.

…

[W3C-WT] W3C Working Draft: "WebTransport", May 2024, <https://www.w3.org/TR/webtransport>

[WT-H3] A. Frindell, E. Kinnear, V. Vasiliev: "WebTransport over HTTP/3", Work In Progress, Internet Draft, draft-ietf-webtrans-http3, <https://datatracker.ietf.org/doc/html/draft-ietf-webtrans-http3/>

[HTTP-DGRAM] IETF RFC 9297: "HTTP Datagrams and the Capsule Protocol", August 2022.

[QUIC-DGRAM] IETF RFC 9221: "An Unreliable Datagram Extension to QUIC", March 2022.

[HTTP-PRIO] IETF RFC 9218: "Extensible Prioritization Scheme for HTTP", June 2022.

[MoQ] L. Curly, K. Pugin, S. Nandakumar, V. Vasiliev, I. Swett: "Media over QUIC Transport", Work In Progress, Internet Draft, draft-ietf-moq-transport, https://datatracker.ietf.org/doc/draft-ietf-moq-transport/

[x6] Zhang, X., Jin, S., He, Y., Hassan, A., Mao, Z. M., Qian, F., & Zhang, Z. L., "QUIC is not Quick Enough over Fast Internet", in *Proceedings of the ACM on Web Conference 2024* (pp. 2713–2722), May 2024.

[x7] Nguyen, M., Nys, P., Pham, S., Silhavy, D., Arbanowski, S., & Steglich, S., "Toward WebTransport Support in HTTP Adaptive Streaming", in *Tenth International Conference on Communications and Electronics 2024* (ICCE) (pp. 96–101), IEEE, July 2024.

Change #2
(all new Text)

## 5.24 QUIC-based segmented media delivery

### 5.24.1 Description

#### 5.24.1.1 General

QUIC, specified in RFC 9000 [32], is a secure, reliable, multiplexed, connection-oriented transport protocol built on top of UDP. It is widely available and its impact on media streaming requires further study. In this clause we study QUIC-specific media streaming, not necessarily based on HTTP/3 [5] that is studied in clause 5.4.

A QUIC client establishes a connection with a server, and within this connection multiple concurrent streams can transport data. Thanks to a more efficient implementation of the TLS initial handshake, a QUIC connection is typically established faster than a TCP + TLS connection, therefore reducing initialization time. Additionally, by allowing QUIC streams to be multiplexed into a single QUIC connection, they can operate independently of each other, each with its own separate congestion window. Because packet retransmission occurs within a stream, a stall in one stream does not block the progress in the other streams. Similarly, packet loss in one stream does not affect the progress of data transfer in other streams multiplexed in the same QUIC connection. Finally, QUIC, like HTTP/2 [4], enables a sender to prioritise traffic at an individual stream level. Unlike HTTP/2, however, QUIC does not provide the means to signal this prioritisation to its connection peer, and so it is entirely up to the sending application to set appropriate priorities when multiplexing streams.

Even though QUIC solves many issues compared to TCP e.g. faster connection establishing and non-blocking multiplexing (using streams), there are some open issues and shortcomings when it comes to media content delivery.

A 5GMS Application Provider runs an adaptive media streaming service between a 5GMS AS and a 5GMS Client running on a UE using 5G Media Streaming protocols conveyed at reference points M2 and M4. However, only M4 is relevant for this key topic since it focuses on the media delivery to the UE and not on the ingest of the media itself. Also, since M5 is not meant to transport media, this reference point is also excluded from this key topic.

#### 5.24.1.2 Application access to QUIC protocol features

In a common scenario, QUIC is not directly exposed to the application, but it is more typically mediated through a QUIC-enabled application protocol such as HTTP/3 as specified in RFC 9114 [5] or Media over QUIC Transport [MoQ], or else through a protocol framework such as WebTransport [W3C-WT] mapped onto HTTP/3 [WT-H3]. For this reason, the set of QUIC features available to an application is the subset exposed by the chosen application protocol or framework (typically invoked via the public API of a client or server library). This approach facilitates the development of QUIC-based communication but at the cost of limiting the level of control an application has over the connection and stream management.

#### 5.24.1.3 Connection and stream management

Applications are encouraged to keep QUIC connections alive when it makes sense to do so, and HTTP/3 client libraries typically facilitate connection keep-alive behaviour for efficiency reasons. In addition, QUIC’s "0-RTT" connection establishment procedure allows a client to reconnect to a server it has previously connected to, and to reuse a security context negotiated during a previous connection and cached by both parties to send application payload data to the server in the first UDP datagram of the new QUIC connection.

In the context of HTTP/3 [5], the mechanism for handling streams is tightly specified by the HTTP/3 protocol that is each HTTP request-response transaction consumes one stream in the QUIC connection.

In contrast, WebTransport [W3C-WT] allows the application a greater degree of control over QUIC connections and streams when layered on top of HTTP/3 [WT-H3]. However, the use of WebTransport for segmented media delivery is not a well-studied area. In one implementation of WebTransport for segmented media streaming [x7] the client sends metrics to the server, the server performs throughput estimation and then sends the respective segments to the client.

QUIC allows both the client and the server to initiate the opening of a new stream in a connection and the WebTransport API exposes this capability to applications. This feature is not directly supported by HTTP/3 [5] where a server push is always stimulated by a client-initiated request.

The protocol extension in RFC 9221 [QUIC-DGRAM] specifies how peers in a QUIC connection can exchange unreliable application messages using a special-purpose QUIC datagram frame. This type of frame is acknowledged by the QUIC recipient, but is never retransmitted if left unacknowledged. Section 2 of RFC 9297 [HTTP-DGRAM] specifies how QUIC datagrams can be used to support the exchange of unreliable HTTP messages in HTTP/3. Furthermore, [WT-H3] specifies how this feature can be exposed in the WebTransport framework when that is layered on top of HTTP/3 (and hence on top of QUIC).

Finally, QUIC supports different congestion control algorithms and control over this is exposed to the application in the WebTransport API [W3C-WT].

An overview of the protocol stack is illustrated in figure 5.24.1.1-1. The white boxes indicating protocol layers and the grey boxes non-protocol layers, such as client/server frameworks. The ovals are instantiations of the API defined by a given block, with red colour indicating a standardized API and black indicating non-standardized implementation-specific APIs (e.g. library APIs). Everything in user space is potentially accessible from the application and must rely on the implementation of the API, while the layers in kernel space are accessed via Operating System calls (e.g. UDP can typically only be accessed via an API provided by the OS-provided socket library).



Figure 5.24.1.1-1: QUIC high-level protocol stack

Note in figure 5.24.1.1-1that the WebTransport framework intiallly uses HTTP/3 for connection establishment and afterwards switches to interfacing with the QUIC layer for data transfer.

The Media over QUIC Transport layer [MoQ] typically uses QUIC directly, but can also be layered on top of WebTransport.

#### 5.24.1.4 Stream prioritisation

By design, QUIC does not provide the means for connection peers to signal the priority of a stream, leaving decision-making over stream prioritisation to the application. To support this, section 2.3 of RFC 9000 [32] recommends that QUIC implementations should expose the means for applications to set relative stream priorities, thus providing the ability to influence the scheduling of packets for transmission in a QUIC connection. Thus, stream prioritisation is typically handled on the higher application layers, but neither the network nor the recipient is aware of the applied mechanism. Relevant standardization efforts include an extensible prioritisation scheme for HTTP (including HTTP/3) defined in RFC 9218 [HTTP-PRIO], and prioritisation scheme for the Media over QUIC (MoQ) transport protocol [MoQ] currently being developed by IETF.

#### 5.24.1.5 Computational overhead of QUIC endpoints

Adoption of QUIC-based application protocols such as HTTP/3 [5], Media over QUIC Transport [MoQ] and application frameworks such as WebTransport [W3C-WT] can have impact on the UEs due to the processing required for the underlying QUIC protocol itself. In TCP, most functions are executed in kernel space, which over the years have been heavily optimised including, in some hosts, hardware offload. In QUIC, by contrast, most protocol functions are executed in user space.

There are two identified sources for this computational overhead [x6]:

1. The in-kernel UDP stack issues many packet reads; this is because each datagram arriving in the link layer is forwarded for processing to the transport layer i.e., there are no offload mechanisms used like the sender-side Generic Segmentation Offload (GSO), or the generic receive offload mechanism (GRO) used by the link layer module to combine datagrams into a mega datagram before forwarding it to the transport layer.

2. There is increased overhead of processing packets (and, as a result, in generating responses) due to increased user space processing required as a result of 1 above (i.e. all the packets need to be processed individually) combined with managing QUIC acknowledgements in user space (instead of kernel space as with TCP in HTTP/2).

These potential performance effects should be taken in consideration because they can negatively impact the overall Quality of Service that can be obtained from a QUIC connection, and can increase the energy consumption of the UE.

#### 5.24.1.6 Key Issue objectives

Editor’s Note: Focus the preceding set of open issues into a set of clear questions that this Key Issue aims to address.

The objectives listed below are targeting QUIC-specific delivery aspects:

• Application access to QUIC connections (e.g. via QUIC libraries).

• Variability of QUIC implementations.

Editor’s Note: Further content to be provided.

### 5.24.2 Collaboration scenarios

#### 5.24.2.1 General

For the purpose of describing the following scenarios, it is only assumed that the 5GMS Client supports the QUIC protocol. Additionally, the 5GMS Client may support higher level protocols based on QUIC (for instance HTTP/3 or WebTransport) but this is not required, and the analysis of those collaboration scenarios is still applicable to any of those cases.

#### 5.24.2.2 QUIC-agnostic 5GMS Client

In this scenario, the Media Stream Handler of the 5GMS Client operates a QUIC session over reference point M4 but the 5GMS Client has no specific feature regarding QUIC. This has the advantage that 5GMS Client is generic and implements the same logic whether or not QUIC is used for the delivery of the media.



Figure 5.24.2.2-1: QUIC-agnostic 5GMS Client

#### 5.24.2.3 Media-independent QUIC-aware 5GMS Client

In this scenario, the Media Stream Handler of the 5GMS Client operates a QUIC session over reference point M4 and the 5GMS Client is able to detect whether QUIC is used and, in case it is used, the 5GMS Client can apply different logic.



Figure 5.24.2.3-1: Media-independent QUIC-aware 5GMS Client

In this case, the QUIC client implementation is not specifically optimised for media transport (e.g. a generic off-the-shelf QUIC client library) and the set of QUIC protocol features exposed to the Media Stream Handler is limited by the richness of its API.

With some limited control over the QUIC streams, such a 5GMS Client would typically be able to:

* Set relative priorities between the different QUIC streams. Relative stream priorities can be useful to differentiate audio and video, base layer and enhancement layer, etc.
* Receive updates sent proactively by the 5GMS AS. For example, a 5GMSd AS could send MPD updates to a Media Player using this mechanism.

Editor’s note: More possible general features to be added.

#### 5.24.2.4 Media-optimised QUIC-aware 5GMS Client

In this scenario, the Media Stream Handler of the 5GMS Client operates a QUIC session over reference point M4 and the 5GMS Client is able to control the delivery of the media within the QUIC session.



Figure 5.24.2.4-1: Media-optimised QUIC-aware 5GMS Client

In this case, the QUIC client implementation is optimised for media transport and the set of QUIC protocol features exposed to the Media Stream Handler is therefore unlimited. Hence, the media-optimised, QUIC-aware 5GMS Client provides the finest control over the delivery of media within the QUIC session.

With fine control over the QUIC streams, such a 5GMS Client would typically be able to:

* Set relative priorities between the different QUIC streams. Relative stream priorities can be useful to differentiate audio and video, base layer and enhancement layer, etc.
* Receive updates sent proactively by the 5GMS AS. For example, a 5GMSd AS could send MPD updates to a Media Player using this mechanism.
* Use one QUIC stream for all the media segments of a given component (e.g. per CMAF Track).

Editor’s note: More possible special features to be added.

### 5.24.3 Architecture mapping

#### 5.24.3.1 General

In the 5GMS architecture, the Media Stream Handler in the 5GMS Client is connected to the 5GMS Application Server (5GMS AS) via reference point M4. Therefore, both the client and the server on that interface need to be compatible in terms of protocols and functionalities. In the clause, the QUIC Client module which is part of the 5GMS Client will thus be connected to a QUIC Server connected or part of the 5GMS AS. Three mappings are considered, one for each collaboration scenario.

#### 5.24.3.2 Mapping with a QUIC-agnostic 5GMS Client

In this mapping, both 5GMS Client and 5GMS AS are agnostic to the QUIC protocol and the usage of QUIC as a network protocol is thus transparent for system. The integration of the QUIC Client in the UE and the QUIC Server with the 5GMS AS is considered to be out of scope of the 5GMS System.



Figure 5.24.3.2-1: Architecture with QUIC-agnostic 5GMS Client and 5GMS AS

#### 5.24.3.3 Mapping with a Media-independent QUIC-aware 5GMS Client

In this mapping, the 5GMS Client and the 5GMS AS Server are integrated with, respectively, a QUIC Client and QUIC Server. The QUIC Client and the QUIC Server are external to the systems which means that any software implementation may be used. The integrations of the QUIC Client in the UE and the QUIC Server with the 5GMS AS are achieved via the exposed APIs by both the QUIC Client and QUIC Server. However, since the QUIC Client and Server are not a standardised part of the 5GMS System, the APIs exposed by the QUIC Client and the QUIC Server may thus also differ in capabilities.

In addition, the QUIC Client and the QUIC Server need to support a common QUIC protocol version. The QUIC protocol does not provide the means for negotiating the protocol version when establishing the connection, but rather enables a QUIC Server to reject unsupported version and propose alternative versions to a QUIC Client. This step of version selection may be influenced by Media Stream Handler and the 5GMS AS through those APIs.



Figure 5.24.3.3-1: Architecture with QUIC-aware 5GMS Client and 5GMS AS

#### 5.24.3.4 Mapping with a Media-optimised QUIC-aware 5GMS Client

In this mapping, the 5GMS Client and the 5GMS AS are integrated with, respectively, a QUIC Client and QUIC Server. The QUIC Client and the QUIC Server are part of the 5GMS System which means that their software implementation follows requirement and functionalities supporting the 5GMS Client’s needs. The integrations of the QUIC Client with the Media Stream Handler in the UE and the QUIC Server with the 5GMS AS are achieved via the exposed APIs by both the QUIC Client and QUIC Server. Since the functionality of the QUIC Client and Server are standardised as part of the 5GMS System in this mapping, the APIs exposed by the QUIC Client and the QUIC Server are compatible in terms of functionalities. In addition, the QUIC Client and the QUIC Server support a commonly agreed QUIC version to guarantee the establishment of the QUIC connection. In addition, those client and server APIs enables the 5GMS Client to optimise the media delivery.



Figure 5.24.3.4-1: Architecture with media-optimised QUIC-aware 5GMS Client and 5GMS AS

### 5.24.4 High-level call flows

#### 5.24.4.1 General

Editor’s Note: Further content to be provided.

### 5.24.5 Gap analysis and requirements

An analysis of the gaps identified in clause 5.24.4 and the derivation of requirements from these are for further study.

### 5.24.6 Candidate solutions

Editor’s Note: Provide candidate solutions (including call flows) for each of the identified issues.

### 5.24.7 Summary and conclusions

The study of this Key Issue has explored the ways in which QUIC can be deployed to support the 5G Media Streaming architecture, and the potential open issues arising from this deployment.

Editor’s Note: Further content to be provided.

END OF CHANGES