Workgroup: MOQ
Internet-Draft:
draft-defoy-moq-relay-network-handling-01
Published: 23 January 2023
Intended Status: Standards Track
Expires: 27 July 2023
Authors: X. de Foy
InterDigital
R. Krishna
InterDigital

MOQ Relays for Support of High-Throughput Low-Latency Traffic

Abstract

This document describes a mechanism to convey information about media frames. The information is used for specific handling in functions such as error recovery and congestion handling. These functions can be critical to improve energy efficiency and network capacity in some (especially wireless) networks. Due to end-to-end encryption, MOQ relays are expected to extract the metadata required by these functions. This document aims to enable a level of wireless network support for MOQ equivalent to what is possible for RTP.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 27 July 2023.

Copyright Notice

Table of Contents

1. Introduction

Wireless networks can be a challenging environment for applications with high-throughput and low latency requirements, such as video conferencing and Extended Reality (XR, presented for example in [I-D.draft-ietf-mops-ar-use-case]). Wireless networks can implement techniques to improve network capacity and energy efficiency, as well as reduce the impact of packet losses on users' quality of experience, using techniques such as (see Appendix A for additional details):

- The network can handle groups of packets based on how critical they are to the user's experience. Some groups of data packets hold application data units that are handled together (e.g., decoded) by the application. 3GPP defines the term "PDU set" to identify these groups of data packets carrying the payload of an application data unit [TR23.700-60], which can correspond to the data packets of an application layer data unit. Application data units can depend on other application data units to be handled or decoded by the application. The network can perform differentiated handling of groups of data packets, e.g., prioritize some groups over others in case of congestion. In congestion situations, the network can also selectively drop data packets that depend on an already lost application data unit.

- The network can limit the amount of time that the radio needs to stay awake to transmit and receive data. To achieve this this, the scheduler can benefit from information on the size and periodicity of traffic, as well as delay budget and expected jitter specific to the application.

Traffic handling of high-throughput low-latency traffic therefore includes differentiated handling of groups of packets (e.g., through

configuring of lower-layer scheduling). To perform this, a network node can act as, or communicate with, a MOQ relay to obtain the metadata that is associated with media data. It is also necessary for the media sender to identify application data units that correspond to different desired traffic handling (e.g., different layers within a frame), and provide associated metadata. Figure 1 describes a MOQ relay providing traffic handling control functionality in an access network (for example, for media streams sent by B towards A and C). For privacy and security, it is desirable that the MOQ relay, which can be operated by a network or service operator, does not have access to any media data or metadata that is not related to its operation. For interoperability, it is also desirable for these mechanisms to not be codec specific.

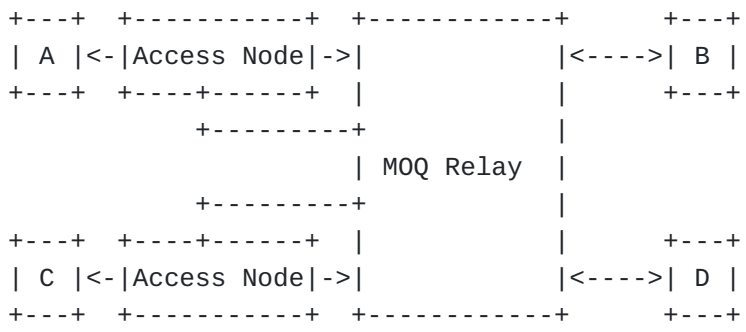

Figure 1: Traffic Handling by Access Networks using a MOQ Relay.

2. Traffic Handling of High-Throughput Low-Latency Traffic

Support of traffic handling of high-throughput low-latency in this document is based on the WARP protocol [I-D.draft-lcurley-warp], since it is the most active proposal in MOQ at this point. WARP is currently based on QUIC streams as a building block. This section may need to be adapted to also support datagram-based segments, if the MOQ protocol design evolves in this direction.

In WARP, a QUIC stream that transports an OBJECT message is the smallest unit of data that can be associated with a differentiated level of service by the network.

2.1. MOQ Relay Behavior

A MOQ relay at the ingress point of a wireless network will extract metadata associated with media segments, and associate this metadata, outside of the QUIC envelop, to packets it forwards to the radio access network. This metadata relates to the QUIC stream that carries the media segment, and to the group of packets carrying the QUIC stream. The list of metadata fields identified by 3GPP for XR support of RTP traffic [TR23.700-60] can be used as a starting

point, as it would enable feature parity for wireless network support of XR over RTP vs. XR over MOQ:

- *Identifier for the group of packets*: the relay can use the stream ID.

- *Start and end packet within the group*: the relay can obtain these indications from QUIC signaling (e.g., offset value 0 and FIN flag).

- *Packet sequence number within the group*: the relay can assign this number based on the packets it receives in order of STREAM frame offset. In case there are missing packets, the relay can use the STREAM frame offset and path MTU to determine the sequence number of the packet.

- *Size of the group* (in bytes or number of packets): the relay can use the Warp message length field of the OBJECT message. If a length in number of packets is needed by the RAN, the relay can estimate this value based on the Warp message length and the MTU. If the Warp message length is set to 0 (i.e., "continues until the end of the stream"), then the relay cannot extract this metadata and may provide a degraded service.

- *Importance of the group*: the relay can use the OBJECT message "delivery order" field set by the media sender.

For example, in a 5G system for downlink flows, a MOQ relay can be collocated with a UPF to extract metadata and provide it to the RAN over GTP-U (similarly to what will be done for RTP/SRTP traffic). For uplink flows, a MOQ relay on the device may extract metadata and provide it to the local networking stack, which will ultimately transmit the packet over the air. However this is not the only solution, since a MOQ application on the device could instead directly provide this metadata to the local networking stack of the device (which is outside of the scope of this document).

To enable different levels of service to be provided to different OBJECT messages, the relay should not coalesce data from different QUIC streams in a same UDP/IP packet, when forwarding towards the RAN.

2.2. Endpoint Behavior

To enable traffic handling of high-throughput low-latency, a MOQ client should set up a MOQ connection through a MOQ relay providing this functionality. Discovery of such relay is out of scope of this document.

Based on the metadata fields list established in Section 2.1, a media sender does not need to set extra metadata to enable XR support by a wireless network. Metadata described in [I-D.draft-lcurley-warp] is sufficient.

It is expected that a media sender will be aware of the nature of the traffic (e.g., AR/VR) and of the possibility for a wireless network to be used as an access network. In such case, the media sender SHOULD set the length field of OBJECT messages to a non-zero value to maximize the information available for the MOQ relay (otherwise, the wireless network support may be degraded).

3. Security Considerations

To enable support for the feature described in this document, the application exposes metadata to a MOQ relay under the control of a network or service operator. End-to-end encrypted media is not exposed to the MOQ relay. The level of exposure is similar to the Frame Marking RTP extension [I-D.draft-ietf-avtext-framemarking].

4. Acknowledgments

Thanks to Srinivas Gudumasu, who was a contributor to the first revision of this draft. Thanks to Jaya Rao, Ghyslain Pelletier, John Kaippallimalil, Sri Gundavelli and Hang Shi for discussions and comments about this draft.

5. Informative References

[I-D.draft-ietf-avtext-framemarking] Zanaty, M., Berger, E., and S. Nandakumar, "Frame Marking RTP Header Extension", Work in Progress, Internet-Draft, draft-ietf-avtext-framemarking-13, 11 November 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-avtext-framemarking-13>.

[I-D.draft-ietf-mops-ar-use-case] Krishna, R. and A. Rahman, "Media Operations Use Case for an Extended Reality Application on Edge Computing Infrastructure", Work in Progress, Internet-Draft, draft-ietf-mops-ar-use-case-09, 14 November 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-mops-ar-use-case-09>.

[I-D.draft-lcurley-warp] Curley, L., Pugin, K., Nandakumar, S., and V. Vasiliev, "Warp - Live Media Transport over QUIC", Work in Progress, Internet-Draft, draft-lcurley-warp-03,

18 January 2023, <https://datatracker.ietf.org/doc/html/draft-lcurley-warp-03>.

[TR23.700-60] 3GPP, "Study on architecture enhancement for XR and media services", 3GPP, 2022, <www.3gpp.org/dynareport/23700-60.htm>.

Appendix A. XR Traffic Handling in 5G Networks

As currently defined in the study report [TR23.700-60], a network function located at the ingress point of a wireless network, for example the User Plane Function (UPF) , can collect metadata from media flows and use it to handle XR traffic by proving the following functionality:

* Convey the collected metadata to the Radio Access Network (RAN), using GTP-U headers encapsulating the data packets it forwards to the RAN (e.g., for dynamic metadata such as packet sequence number within the group, priority/importance, dependency information, size, group ID). This makes it possible for the RAN to perform differentiated handling of the packets. The network can also convey some metadata related to a flow in control messages to the RAN (e.g., periodicity, delay budget). This makes it possible for the RAN to configure efficient scheduling for the flow.

* Use the collected metadata to perform some local processing (on the UPF or 5G device) such as locally prioritizing groups of packets in case of congestion.

Data plane metadata is expected to be obtained, for the time being, from RTP/SRTP and their extensions headers, or alternatively, from other methods not standardized by 3GPP.

* The following metadata was agreed to be used in the data plane:

  - ID of a group of packets that share similar handling by the network (a "PDU set")

  - Indication of the first and last data packet in a group

  - A sequence number of individual packets within the group

  - Size of a group in number of octets or data packets

  - Group importance relative to other groups

*The following metadata was agreed to be used in the control plane, where it is provisioned by the service operator.

-QoS parameters: delay budget and error rate for groups (PDU sets)

-Burst periodicity

-Whether all data packets are needed to process the application data unit carried by the group

Authors' Addresses

Xavier de Foy
InterDigital
Canada

Email: xavier.defoy@interdigital.com

Renan Krishna
InterDigital
Canada

Email: renan.krishna@interdigital.com