

Independent Submission
Internet-Draft
Intended status: Informational
Expires: 23 April 2022

L. Dong
K. Makhijani
R. Li
Futurewei Technologies Inc.
20 October 2021

A Use Case of Packets' Significance Difference with Media Scalability
draft-dong-usecase-packet-significance-diff-01

Abstract

This document introduces a use case of packets' significance difference embedded with media scalability. With the dominance of video traffic on the Internet, selectively dropping packets or parts of packets from competing media streams becomes a complementary mechanism when dealing with network congestion.

The document describes the characteristics of media scalability, some limitations of existing end-to-end congestion control mechanisms through rate control and adaptation, explains why current ways of entire packet dropping at the traffic class level using in-network active queue management are not most appropriate to meet end users' Quality of Service expectations. The document identifies that there exists "significance difference" among packets or even among parts of the packets within a flow, and brings out a new set of requirements for application and network to support packet significance difference to improve the Quality of Experience of end users.

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 23 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terms and Abbreviations	3
3. Media Scalability and Congestion Control	4
4. Packet Dropping	5
5. Significance Difference Among Packets and Within Packets . .	6
6. New Requirements	7
7. IANA Considerations	8
8. Security Considerations	8
9. Acknowledgements	8
10. Informative References	8
Authors' Addresses	11

1. Introduction

Recent studies [CiscoNetworkingIndex] show that IP video traffic will be 82 percent of all consumer Internet traffic by 2021 in a global scale, up from 73 percent in 2016. Live video has grown 15-fold from 2016 to 2021, accounts for 13 percent of Internet video traffic by 2021. VR (Virtual Reality) and AR (Augmented Reality) traffic has increased 20-fold between 2016 and 2021, at a CAGR (Compound Annual Growth Rate) of 82 percent. With the rapid growth of multimedia streaming traffic, it is increasingly likely that multiple streaming flows share a bottleneck link, which would inevitably cause network congestion. Today's transport protocols and Internet protocols are oblivious to multimedia streaming applications or end users' QoE (Quality of Experience) expectations. From the perspective of user experience and user expectation, the following two observations could be made.

- * It is very likely that a user may prefer to acquire the media content in a somewhat degraded quality that is above the tolerance threshold rather than getting nothing at all for a few seconds.

- * A user may be particularly interested in certain group of blocks belonging to the interested objects in the media content (i.e., Region of Interest, RoI). It is necessary to prevent the RoI blocks from being lost during transmission.

At the beginning of this document, the different types of scalability are discussed in current video codecs, facilitating the rate control and adaptation mechanisms carried out in video segments when dealing with network congestion during the media streaming. It is acknowledged that such mechanisms have efficiently improved users' QoE. However, the packets on the wire cannot avoid the possibility of being entirely dropped when the bottleneck network nodes cannot retain them due to buffer overflowing during congestion. Thanks to the scalability characteristics designed to the video codecs, it is not hard to find out that the importance or significance of different packets within a media streaming flow or even different parts of the single packet could vary for their usefulness in decoding and recovering the media content to meet receiver's expectation. The document highlights the requirements of making the user's preference and application context aware to the network to help further improve the QoE of media streaming. Accordingly, the network could treat the packets or different parts of the packets according to the characteristics of the packets and end users' preferences.

2. Terms and Abbreviations

The terms and abbreviations used in this document are listed below.

- * AR: Augmented Reality
- * CAGR: Compound Annual Growth Rate
- * DASH: Dynamic Adaptive Streaming over HTTP
- * GOP: Group of Picture
- * HAS: HTTP Adaptive Stream
- * HTTP: Hypertext Transfer Protocol
- * QoE: Quality of Experience
- * QoS: Quality of Service
- * SNR: Signal-to-Noise Ratio
- * SVC: Scalable Video Coding

* VR: Virtual Reality

The above terminology is defined in greater details in the remainder of this document.

3. Media Scalability and Congestion Control

A visual scene is represented in digital form by sampling the real scene spatially on a rectangular grid in the video image plane and sampling temporally at regular time intervals as a sequence of still frames. Correspondingly, modern media codec [Conklin2001] [Kim2001] incorporates three types of "Scalability": i.e., temporal scalability, spatial scalability, and quality scalability, which adapt the media bitstream by adding or removing some portions to/from it in order to match the different needs or preferences of end users as well as to the network conditions.

Temporal scalability refers to scalability designed to allow the frame rate of the video bitstream to be varied using interlayer prediction. Spatial scalability represents the spatial resolution variations with respect to the original image frame. The lower layer provides the basic spatial resolution. The enhancement layer employs the spatially interpolated lower layers and constructs the source video in its full spatial resolution. Quality scalability is also commonly referred to as fidelity or SNR (Signal-to-Noise Ratio) scalability. Each spatial layer could have many quality layers. For example, SVC (Scalable Video Coding) [SVC] is an H.264 [H.264] extension that divides a single video bitstream into multiple representations or layers. This hierarchical layered structure comprises a base layer and two enhancement layers. The media may be scaled up by adding the enhancement layer(s) or scaled down by dropping the enhancement layer(s). The levels of scalability included in the media stream affect the quality of media presented to the end users' devices.

Bursty loss and longer-than-expected delay have catastrophic effect on QoE to end-users in media streaming. They are usually caused by network congestion. Despite all kinds of congestion control mechanisms developed in the community over the decades [Saadi2019] [Adams2013], they often target different goals, e.g., link utilization improvement, loss reduction, fairness enhancement. By leveraging the flexibility and variety of media qualities provided by different types of media scalability, for media streaming, minimizing the possibility of network congestion can often be achieved by rate control and media adaptation methods.

Existing rate control and adaptation methods [Bentaleb2019] [Wu2001] can be at source-side and receiver-side, which are carried at end devices and servers, respectively.

- * In source-based schemes [Wu2000], source regulates the sending rate to maintain the packet loss ratio below a threshold by employing the feedback from probing experiments, or source determines the sending rate through a TCP-friendly model. However, some constraints exist, media codecs can usually only adjust their output rates in a much more coarse-grained fashion than, for example, TCP. Users' QoE would also suffer if encoding rates are switched too frequently.
- * HTTP (Hypertext Transfer Protocol)-based dynamic video adaptation methods [Kua2017] could be driven by source. The server collects the feedback from the network and client (e.g., dynamic variation of network bandwidth and receiving buffer capacity of the client), and accordingly, the video quality will be adapted and streamed. On the other hand, adaptation techniques are also proposed at receiver-side, which mainly use DASH (Dynamic Adaptive Streaming over HTTP) [MPEG-DASH-SAND] [MPEG-DASH] and HAS (HTTP Adaptive Stream) for streaming adapted video data.
- * The receiver-based rate control [McCannel1996] is typically used in multicasting scalable media content, which is split into multiple layers, with each layer corresponding to one channel in the multicast tree. Receivers could regulate their own receiving rates by adding/dropping channels. Thus receiver-based rate has its limited usage in unicasting. All these techniques consider full quality while streaming from sender to receivers; hence, they consume more resources in the network.

4. Packet Dropping

Acknowledging the benefits offered by various congestion control and congestion avoidance mechanisms, we would like to point out that the feedback and rate adaption might not be prompt enough to cope with the dropping of packets on the wire.

In the current Internet, a packet is treated as the minimal, independent, and self-sufficient unit that gets classified, forwarded, or dropped completely by a network node, according to the local configuration and congestion condition. Although congestion discard can be mitigated by a mixture of ingress traffic shaping and active queue management mechanisms [Thiruchelvi2008] [Adams2013] to avoid any network resource overdrawn, it is not feasible to be deployed on a large scale, meanwhile wastes network resources preparing for the worst possible scenario.

DiffServ [RFC2475] is used to manage resources such as bandwidth and queuing buffers on a per-hop basis between different classes of traffic. The Internet traffic may be separated into different classes with differentiated priorities. This allows preferential treatment for latency or loss sensitive traffic over more tolerant applications, for example those that can afford retransmission. However, with video traffic dominating Internet traffic, flows of media streaming applications with the same class still compete for network resources when encountering bottleneck links and fighting network congestion, preference decided on traffic class would not be effective to eliminate the possibility of degraded service levels or packet drops due to collisions with each other.

The routers treat every bit/byte in the packet payload equally, which means every bit/byte has the same significance to the routers. Each to-be-dropped packet is discarded completely. If the transport layer protocol is TCP, after timeout or duplicate acknowledgements received at the sender, the sender may re-try to send the dropped packet before the maximum number of re-transmissions reaches. Retransmission of packets wastes network resources, reduces the overall throughput of the connection and causes longer latency for the packet delivery. The study [RFC8836] has shown that a loss rate of 1% is tolerable to users while a loss rate of 3% is intolerable to most users who found the quality to be annoying (or worse), according to the subjective opinions of the effects of packet loss on media quality. Therefore, the current way of handling network congestion by discarding the packet entirely and retransmitting the packets in a blind-of-application-context manner is not very suitable for media streaming.

5. Significance Difference Among Packets and Within Packets

With the various scalability implemented in the media codec, some bits of an encoded media stream are more important than others. Bits belonging to base layer usually are more significant to the decoder than bits belonging to enhancement layers. For example, I-frames hold complete picture data [Orosz2015] and is frequently referenced by the subsequent frames. It is inserted by the encoder when the scene changes. Losing the first I-frame in the GOP (Group of Pictures) would cause video picture even missing for few seconds, because P- and B-frames referencing to the I-frame would not be decoded nor displayed either. Thus, I-frames are most essential in the media stream, which have the most effect on perceived video quality, and such effect can last through the whole GOP. P- and B-frames are inserted at appropriate places to reduce the video size or bitrate and are tuned to maintain a certain video quality level. P-frame stands for Predicted Frame and allows macroblocks to be compressed using temporal prediction in addition to spatial

prediction. A P-frame might be referenced by a P frame after it, or a B frame before or after it. B-frame stands for bi-directional frame, which can be predicted using backward prediction and forward prediction. A B-frame can act as a reference, and if so, it is termed as a reference B-frame. If a B-frame is not to be used as a reference, it is called a non-reference B-frame. Video scenes with a low level of movement are less sensitive to both B-frame and P-frame packet loss, alternatively video scenes with a high level of movement are more sensitive to both B-frame and P-frame packet loss. A lost P-frame can impact the remaining part of the GOP. A lost B-frame has only local effects in a slowly moving content or with large static background. In a scene of a dynamically moving content, losing B-frame has more dramatic impact and its scale can be as far-reaching as a P-frame loss.

As another example, macroblocks that are identified to represent the objects in RoI are likely more important than other macroblocks of non-RoI regions. For packets carrying RoI macroblocks in the media stream need to have higher priority to be retained compared to other packets carrying non-RoI macroblocks.

According to the characteristics of frames contained in the video packet payload, namely: frame type, whether the frames are referenced by other frames, movement level of the pictures, whether the picture contained in the packet belongs to RoI or not, etc., significance difference could present among packets for the video decoding at the receiver side and the QoE improvement of end users. The dropping priority is possibly implemented at packet level in the network.

On the other hand, let's say that the end-users can reveal their preferences to the network, e.g., degree of tolerance to the decoded media content' quality degradation, which might reflect visually such as resolution reduction, missing objects in non-RoI regions, the network could selectively drop packets in a differentiated manner according to such information. This avoids retransmission or delay of those packets with higher significance, reduce the experienced end-to-end latency of end users, and maintain the continuous streaming of the media. This is achieved at the cost of dropping lower-significance packets.

6. New Requirements

We have discussed in the previous sections that due to the various types of scalability implemented in the media codecs, "significance difference" exists among packets or even among parts of the packets. In other words, some packets containing the more important macroblocks (e.g., RoI macroblocks, base layer macroblocks) show higher significance than other packets for the media decoding at the

receiver side and the improvement of QoE of end users. In order for the network be able to treat the packets of media streams in a differentiated manner and at finer granularity than DiffServ, the application shall reveal some information to the network to enable selective packet dropping or partial packet dropping. For example, an API could be implemented to input such information or metadata from the application, which might be mapped to IPv6 extension header, IPv4 options or a dedicated metadata field in the IP header. Some examples of such information or metadata are listed below:

- * Receiving end user's preference on media quality, e.g. tolerable quality degradation regarding for example resolution.
- * Characteristics of media content contained in the packets, e.g., frame type, whether the packet contains frames that are referenced by other frames, movement level of the video sample contained in the packet.
- * Labeling of the packets or some parts of the packets that correspond to receiver's interested objects as RoI.

Correspondingly, the network shall be able to leverage the above information revealed by the application, and selectively drop packets or parts of the packets from competing media streaming flows with precedence order when network congestion happens. The retransmission could be maximumly eliminated. The receiving end user is able to consume the delivered packets as many as possible in-time with acceptable quality.

7. IANA Considerations

This document requires no actions from IANA.

8. Security Considerations

This document introduces no new security issues.

9. Acknowledgements

10. Informative References

[Adams2013]

Adams, R., "Active Queue Management: A Survey", IEEE Communications Surveys and Tutorials, vol. 15, no. 3, pp. 1425-1476, 2013, <<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6329367>>.

- [Bentaleb2019] Bentaleb, A., Taani, B., Begen, A. C., Timmerer, C., and R. Zimmermann, "A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP", IEEE Communications Surveys and Tutorials, vol. 21, no. 1, pp. 562-585, 2019, <<https://ieeexplore.ieee.org/document/8424813>>.
- [CiscoNetworkingIndex] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016 to 2021", June 2017, <<https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>>.
- [Conklin2001] Conklin, G. J., Greenbaum, G. S., Lillevold, K. O., Lippman, A. F., and Y. A. Reznik, "Video Coding for Streaming Media Delivery on the Internet", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 3, pp. 269-281, 2001, <<https://ieeexplore.ieee.org/document/911155>>.
- [H.264] ITU-T, "H.264 : Advanced Video Coding for Generic Audiovisual Services", 2019, <<https://www.itu.int/rec/T-REC-H.264-201906-I/en>>.
- [Kim2001] Kim, T., "Scalable video Streaming Over Internet", Ph.D. Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, January 2005, <<https://smartech.gatech.edu/handle/1853/6829>>.
- [Kua2017] Kua, J., Armitage, G., and P. Branch, "A Survey of Rate Adaptation Techniques for Dynamic Adaptive Streaming Over HTTP", IEEE Communications Surveys and Tutorials, vol. 19, no. 3, pp. 1842-1866, 2017, <<https://ieeexplore.ieee.org/document/7884970>>.
- [McCanne1996] McCanne, S., Jacobson, V., and M. Vetterli, "Receiver-Driven Layered Multicast", ACM Sigcomm, pp. 117-130, 1996, <<http://www.cs.toronto.edu/syslab/courses/csc2209/06au/papers/recmc.pdf>>.
- [MPEG-DASH] ISO/IEC, "23009-1:2019, Dynamic Adaptive Streaming over HTTP (DASH) - Part 1: Media Presentation Description and Segment Formats", 2019, <<https://www.iso.org/standard/79329.html>>.

- [MPEG-DASH-SAND] ISO/IEC, "23009-5:2017, Dynamic Adaptive Streaming over HTTP (DASH) - Part 5: Server and Network Assisted DASH (SAND)", February 2017, <<https://www.iso.org/standard/69079.html>>.
- [Orosz2015] Orosz, P., Skopko, T., and P. Varga, "Towards Estimating Video QoE Based on Frame Loss Statistics of the Video Streams", DOI: 10.1109/INM.2015.7140482, IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 1282-1285, 2015, <<https://ieeexplore.ieee.org/document/7140482>>.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998, <<https://datatracker.ietf.org/doc/html/rfc2475>>.
- [RFC8836] Jesup, R. and Z. Sarker, "Congestion Control Requirements for Interactive Real-Time Media", RFC 8836, January 2001, <<https://datatracker.ietf.org/doc/html/rfc8836>>.
- [Saadi2019] Al-Saadi, R., Armitage, G., But, J., and P. Branch, "A Survey of Delay-Based and Hybrid TCP Congestion Control Algorithms", IEEE Communications Surveys and Tutorials, vol. 21, no. 4, pp. 3609-3638, 2019, <<https://ieeexplore.ieee.org/document/8668433>>.
- [SVC] Schwarz, H., Marpe, D., and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 9, 1103-1120, 2007, <<https://ieeexplore.ieee.org/document/4317636>>.
- [Thiruchelvi2008] Thiruchelvi, G. and J. Raja, "A Survey On Active Queue Management Mechanisms", International Journal of Computer Science and Network Security, vol. 8, 2008, <https://www.researchgate.net/publication/310468829_A_Survey_on_Active_Queue_Management_Techniques>.
- [Wu2000] Wu, D., Hou, Y., and Y. Zhang, "Transporting Real-Time Video Over the Internet: Challenges and approaches", Proceedings of the IEEE, vol. 88, no. 12, 1855-1875, 2000, <http://www.wu.ece.ufl.edu/mypapers/ProcIEEE_camera.pdf>.

- [Wu2001] Wu, D., Hou, Y., Zhu, W., Zhang, Y., and J. Peha,
"Streaming Video Over the Internet: Approaches and
Directions", IEEE Transactions on Circuits and Systems for
Video Technology, vol. 11, no. 3, pp. 282-300, 2001,
<<https://ieeexplore.ieee.org/document/911156>>.

Authors' Addresses

Lijun Dong
Futurewei Technologies Inc.

Email: lijun.dong@futurewei.com

Kiran Makhijani
Futurewei Technologies Inc.

Email: kiran.ietf@gmail.com

Richard Li
Futurewei Technologies Inc.

Email: richard.li@futurewei.com

MOPS
Internet-Draft
Intended status: Informational
Expires: 7 September 2022

R. Krishna
InterDigital Europe Limited
A. Rahman
InterDigital Communications, LLC
6 March 2022

Media Operations Use Case for an Augmented Reality Application on Edge
Computing Infrastructure
draft-ietf-mops-ar-use-case-04

Abstract

This document explores the issues involved in the use of Edge Computing resources to operationalize media use cases that involve Extended Reality (XR) applications. In particular, we discuss those applications that run on devices having different form factors and need Edge computing resources to mitigate the effect of problems such as a need to support interactive communication requiring low latency, limited battery power, and heat dissipation from those devices. The intended audience for this document are network operators who are interested in providing edge computing resources to operationalize the requirements of such applications.

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 7 September 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Use Case	3
3.1. Processing of Scenes	4
3.2. Generation of Images	5
4. Requirements	5
5. AR Network Traffic and Interaction with TCP	8
6. Informative References	8
Authors' Addresses	12

1. Introduction

Extended Reality (XR) is a term that includes Augmented Reality (AR), Virtual Reality (VR) and Mixed Reality (MR) [XR]. AR combines the real and virtual, is interactive and is aligned to the physical world of the user [AUGMENTED_2]. On the other hand, VR places the user inside a virtual environment generated by a computer [AUGMENTED]. MR merges the real and virtual world along a continuum that connects completely real environment at one end to a completely virtual environment at the other end. In this continuum, all combinations of the real and virtual are captured [AUGMENTED].

XR applications will bring several requirements for the network and the mobile devices running these applications. Some XR applications such as AR require a real-time processing of video streams to recognize specific objects. This is then used to overlay information on the video being displayed to the user. In addition XR applications such as AR and VR will also require generation of new video frames to be played to the user. Both the real-time processing of video streams and the generation of overlay information are computationally intensive tasks that generate heat [DEV_HEAT_1], [DEV_HEAT_2] and drain battery power [BATT_DRAIN] on the mobile device running the XR application. Consequently, in order to run applications with XR characteristics on mobile devices, computationally intensive tasks need to be offloaded to resources provided by Edge Computing.

Edge Computing is an emerging paradigm where computing resources and storage are made available in close network proximity at the edge of the Internet to mobile devices and sensors [EDGE_1], [EDGE_2]. These edge computing devices use cloud technologies that enable them to support offloaded XR applications. In particular, the edge devices deploy cloud computing implementation techniques such as disaggregation (breaking vertically integrated systems into independent components with open interfaces using SDN), virtualization (being able to run multiple independent copies of those components such as SDN Controller apps, Virtual Network Functions on a common hardware platform) and commoditization (being able to elastically scale those virtual components across commodity hardware as the workload dictates) [EDGE_3]. Such techniques enable XR applications requiring low-latency and high bandwidth to be delivered by mini-clouds running on proximate edge devices

In this document, we discuss the issues involved when edge computing resources are offered by network operators to operationalize the requirements of XR applications running on devices with various form factors. Examples of such form factors include Head Mounted Displays (HMD) such as Optical-see through HMDs and video-see-through HMDs and Hand-held displays. Smart phones with video cameras and GPS are another example of such devices. These devices have limited battery capacity and dissipate heat when running. Besides as the user of these devices moves around as they run the XR application, the wireless latency and bandwidth available to the devices fluctuates and the communication link itself might fail. As a result algorithms such as those based on adaptive-bit-rate techniques that base their policy on heuristics or models of deployment perform sub-optimally in such dynamic environments.[ABR_1]. We motivate these issues with a use-case that we present in the following sections.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Use Case

We now describe a use case that involves an application with AR systems' characteristics. Consider a group of tourists who are being conducted in a tour around the historical site of the Tower of London. As they move around the site and within the historical buildings, they can watch and listen to historical scenes in 3D that are generated by the AR application and then overlaid by their AR headsets onto their real-world view. The headset then continuously updates their view as they move around.

The AR application first processes the scene that the walking tourist is watching in real-time and identifies objects that will be targeted for overlay of high resolution videos. It then generates high resolution 3D images of historical scenes related to the perspective of the tourist in real-time. These generated video images are then overlaid on the view of the real-world as seen by the tourist.

We now discuss this processing of scenes and generation of high resolution images in greater detail.

3.1. Processing of Scenes

The task of processing a scene can be broken down into a pipeline of three consecutive subtasks namely tracking, followed by an acquisition of a model of the real world, and finally registration [AUGMENTED].

Tracking: This includes tracking of the three dimensional coordinates and six dimensional pose (coordinates and orientation) of objects in the real world[AUGMENTED]. The AR application that runs on the mobile device needs to track the pose of the user's head, eyes and the objects that are in view. This requires tracking natural features that are then used in the next stage of the pipeline.

Acquisition of a model of the real world: The tracked natural features are used to develop an annotated point cloud based model that is then stored in a database. To ensure that this database can be scaled up, techniques such as combining a client side simultaneous tracking and mapping and a server-side localization are used [SLAM_1], [SLAM_2], [SLAM_3], [SLAM_4].

Registration: The coordinate systems, brightness, and color of virtual and real objects need to be aligned in a process called registration [REG]. Once the natural features are tracked as discussed above, virtual objects are geometrically aligned with those features by geometric registration. This is followed by resolving occlusion that can occur between virtual and the real objects [OCCL_1], [OCCL_2]. The AR application also applies photometric registration [PHOTO_REG] by aligning the brightness and color between the virtual and real objects. Additionally, algorithms that calculate global illumination of both the virtual and real objects [GLB_ILLUM_1], [GLB_ILLUM_2] are executed. Various algorithms to deal with artifacts generated by lens distortion [LENS_DIST], blur [BLUR], noise [NOISE] etc are also required.

3.2. Generation of Images

The AR application must generate a high-quality video that has the properties described in the previous step and overlay the video on the AR device's display- a step called situated visualization. This entails dealing with registration errors that may arise, ensuring that there is no visual interference [VIS_INTERFERE], and finally maintaining temporal coherence by adapting to the movement of user's eyes and head.

4. Requirements

The components of AR applications perform tasks such as real-time generation and processing of high-quality video content that are computationally intensive. As a result, on AR devices such as AR glasses excessive heat is generated by the chip-sets that are involved in the computation [DEV_HEAT_1], [DEV_HEAT_2]. Additionally, the battery on such devices discharges quickly when running such applications [BATT_DRAIN].

A solution to the heat dissipation and battery drainage problem is to offload the processing and video generation tasks to the remote cloud. However, running such tasks on the cloud is not feasible as the end-to-end delays must be within the order of a few milliseconds. Additionally, such applications require high bandwidth and low jitter to provide a high QoE to the user. In order to achieve such hard timing constraints, computationally intensive tasks can be offloaded to Edge devices.

Another requirement for our use case and similar applications such as 360 degree streaming is that the display on the AR/VR device should synchronize the visual input with the way the user is moving their head. This synchronization is necessary to avoid motion sickness that results from a time-lag between when the user moves their head and when the appropriate video scene is rendered. This time lag is often called "motion-to-photon" delay. Studies have shown [PER_SENSE], [XR], [OCCL_3] that this delay can be at most 20ms and preferably between 7-15ms in order to avoid the motion sickness problem. Out of these 20ms, display techniques including the refresh rate of write displays and pixel switching take 12-13ms [OCCL_3], [CLOUD]. This leaves 7-8ms for the processing of motion sensor inputs, graphic rendering, and RTT between the AR/VR device and the Edge. The use of predictive techniques to mask latencies has been considered as a mitigating strategy to reduce motion sickness [PREDICT]. In addition, Edge Devices that are proximate to the user might be used to offload these computationally intensive tasks. Towards this end, the 3GPP requires and supports an Ultra Reliable Low Latency of 0.1ms to 1ms for communication between an Edge server and User Equipment (UE) [URLLC].

Note that the Edge device providing the computation and storage is itself limited in such resources compared to the Cloud. So, for example, a sudden surge in demand from a large group of tourists can overwhelm that device. This will result in a degraded user experience as their AR device experiences delays in receiving the video frames. In order to deal with this problem, the client AR applications will need to use Adaptive Bit Rate (ABR) algorithms that choose bit-rates policies tailored in a fine-grained manner to the resource demands and playback the videos with appropriate QoE metrics as the user moves around with the group of tourists.

However, heavy-tailed nature of several operational parameters make prediction-based adaptation by ABR algorithms sub-optimal [ABR_2]. This is because with such distributions, law of large numbers works too slowly, the mean of sample does not equal the mean of distribution, and as a result standard deviation and variance are unsuitable as metrics for such operational parameters [HEAVY_TAIL_1], [HEAVY_TAIL_2]. Other subtle issues with these distributions include the "expectation paradox" [HEAVY_TAIL_1] where the longer we have waited for an event the longer we have to wait and the issue of mismatch between the size and count of events [HEAVY_TAIL_1]. This makes designing an algorithm for adaptation error-prone and challenging. Such operational parameters include but are not limited to buffer occupancy, throughput, client-server latency, and variable transmission times. In addition, edge devices and communication links may fail and logical communication relationships between various software components change frequently as the user moves around with their AR device [UBICOMP].

Thus, once the offloaded computationally intensive processing is completed on the Edge Computing, the video is streamed to the user with the help of an ABR algorithm which needs to meet the following requirements [ABR_1]:

- * Dynamically changing ABR parameters: The ABR algorithm must be able to dynamically change parameters given the heavy-tailed nature of network throughput. This, for example, may be accomplished by AI/ML processing on the Edge Computing on a per client or global basis.
- * Handling conflicting QoE requirements: QoE goals often require high bit-rates, and low frequency of buffer refills. However in practice, this can lead to a conflict between those goals. For example, increasing the bit-rate might result in the need to fill up the buffer more frequently as the buffer capacity might be limited on the AR device. The ABR algorithm must be able to handle this situation.
- * Handling side effects of deciding a specific bit rate: For example, selecting a bit rate of a particular value might result in the ABR algorithm not changing to a different rate so as to ensure a non-fluctuating bit-rate and the resultant smoothness of video quality. The ABR algorithm must be able to handle this situation.

5. AR Network Traffic and Interaction with TCP

In addition to the requirements for ABR algorithms, there are other operational issues that need to be considered for AR use cases such as the one described above. In a study [AR_TRAFFIC] conducted to characterize multi-user AR over cellular networks, the following issues were identified:

- * The uploading of data from an AR device to a remote server for processing dominates the end-to-end latency.
- * A lack of visual features in the grid environment can cause increased latencies as the AR device uploads additional visual data for processing to the remote server.
- * AR applications tend to have large bursts that are separated by significant time gaps. As a result, the TCP congestion window enters slow start before the large bursts of data arrive increasing the perceived user latency. The study [AR_TRAFFIC] shows that segmentation latency at 4G LTE (Long Term Evolution)'s RAN (Radio Access Network)'s RLC (Radio Link Control) layer impacts TCP's performance during slow-start.

6. Informative References

- [ABR_1] Mao, H., Netravali, R., and M. Alizadeh, "Neural Adaptive Video Streaming with Pensieve", In Proceedings of the Conference of the ACM Special Interest Group on Data Communication, pp. 197-210, 2017.
- [ABR_2] Yan, F., Ayers, H., Zhu, C., Fouladi, S., Hong, J., Zhang, K., Levis, P., and K. Winstein, "Learning in situ: a randomized experiment in video streaming", In 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20), pp. 495-511, 2020.
- [AR_TRAFFIC] Apichartttrisorn, K., Balasubramanian, B., Chen, J., Sivaraj, R., Tsai, Y., Jana, R., Krishnamurthy, S., Tran, T., and Y. Zhou, "Characterization of Multi-User Augmented Reality over Cellular Networks", In 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 1-9. IEEE, 2020.
- [AUGMENTED] Schmalstieg, D. S. and T.H. Hollerer, "Augmented Reality", Addison Wesley, 2016.

- [AUGMENTED_2] Azuma, R. T., "A Survey of Augmented Reality.", Presence:Teleoperators and Virtual Environments 6.4, pp. 355-385., 1997.
- [BATT_DRAIN] Seneviratne, S., Hu, Y., Nguyen, T., Lan, G., Khalifa, S., Thilakarathna, K., Hassan, M., and A. Seneviratne, "A survey of wearable devices and challenges.", In IEEE Communication Surveys and Tutorials, 19(4), p.2573-2620., 2017.
- [BLUR] Kan, P. and H. Kaufmann, "Physically-Based Depth of Field in Augmented Reality.", In Eurographics (Short Papers), pp. 89-92., 2012.
- [CLOUD] Corneo, L., Eder, M., Mohan, N., Zavodovski, A., Bayhan, S., Wong, W., Gunningberg, P., Kangasharju, J., and J. Ott, "Surrounded by the Clouds: A Comprehensive Cloud Reachability Study.", In Proceedings of the Web Conference 2021, pp. 295-304, 2021.
- [DEV_HEAT_1] LiKamWa, R., Wang, Z., Carroll, A., Lin, F., and L. Zhong, "Draining our Glass: An Energy and Heat characterization of Google Glass", In Proceedings of 5th Asia-Pacific Workshop on Systems pp. 1-7, 2013.
- [DEV_HEAT_2] Matsushashi, K., Kanamoto, T., and A. Kurokawa, "Thermal model and countermeasures for future smart glasses.", In Sensors, 20(5), p.1446., 2020.
- [EDGE_1] Satyanarayanan, M., "The Emergence of Edge Computing", In Computer 50(1) pp. 30-39, 2017.
- [EDGE_2] Satyanarayanan, M., Klas, G., Silva, M., and S. Mangiante, "The Seminal Role of Edge-Native Applications", In IEEE International Conference on Edge Computing (EDGE) pp. 33-40, 2019.
- [EDGE_3] Peterson, L. and O. Sunay, "5G mobile networks: A systems approach.", In Synthesis Lectures on Network Systems., 2020.

- [GLB_ILLUM_1] Kan, P. and H. Kaufmann, "Differential irradiance caching for fast high-quality light transport between virtual and real worlds.", In IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 133-141, 2013.
- [GLB_ILLUM_2] Franke, T., "Delta voxel cone tracing.", In IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 39-44, 2014.
- [HEAVY_TAIL_1] Crovella, M. and B. Krishnamurthy, "Internet measurement: infrastructure, traffic and applications", John Wiley and Sons Inc., 2006.
- [HEAVY_TAIL_2] Taleb, N., "The Statistical Consequences of Fat Tails", STEM Academic Press, 2020.
- [I-D.ietf-mops-streaming-opcons] Holland, J., Begen, A., and S. Dawkins, "Operational Considerations for Streaming Media", Work in Progress, Internet-Draft, draft-ietf-mops-streaming-opcons-09, 1 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-mops-streaming-opcons-09>>.
- [LENS_DIST] Fuhrmann, A. and D. Schmalstieg, "Practical calibration procedures for augmented reality.", In Virtual Environments 2000, pp. 3-12. Springer, Vienna, 2000.
- [NOISE] Fischer, J., Bartz, D., and W. Straßer, "Enhanced visual realism by incorporating camera image effects.", In IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 205-208., 2006.
- [OCCL_1] Breen, D.E., Whitaker, R.T., and M. Tuceryan, "Interactive Occlusion and automatic object placement for augmented reality", In Computer Graphics Forum, vol. 15, no. 3 , pp. 229-238, Edinburgh, UK: Blackwell Science Ltd, 1996.
- [OCCL_2] Zheng, F., Schmalstieg, D., and G. Welch, "Pixel-wise closed-loop registration in video-based augmented reality", In IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 135-143, 2014.

- [OCCL_3] Lang, B., "Oculus Shares 5 Key Ingredients for Presence in Virtual Reality.", <https://www.roadtovr.com/oculus-shares-5-key-ingredients-for-presence-in-virtual-reality/>, 2014.
- [PER_SENSE] Mania, K., Adelstein, B.D., Ellis, S.R., and M.I. Hill, "Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity.", In Proceedings of the 1st Symposium on Applied perception in graphics and visualization pp. 39-47., 2004.
- [PHOTO_REG] Liu, Y. and X. Granier, "Online tracking of outdoor lighting variations for augmented reality with moving cameras", In IEEE Transactions on visualization and computer graphics, 18(4), pp.573-580, 2012.
- [PREDICT] Buker, T. J., Vincenzi, D.A., and J.E. Deaton, "The effect of apparent latency on simulator sickness while using a see-through helmet-mounted display: Reducing apparent latency with predictive compensation..", In Human factors 54.2, pp. 235-249., 2012.
- [REG] Holloway, R. L., "Registration error analysis for augmented reality.", In Presence:Teleoperators and Virtual Environments 6.4, pp. 413-432., 1997.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [SLAM_1] Ventura, J., Arth, C., Reitmayr, G., and D. Schmalstieg, "A minimal solution to the generalized pose-and-scale problem", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 422-429, 2014.
- [SLAM_2] Sweeny, C., Fragoso, V., Hollerer, T., and M. Turk, "A scalable solution to the generalized pose and scale problem", In European Conference on Computer Vision, pp. 16-31, 2014.

- [SLAM_3] Gauglitz, S., Sweeny, C., Ventura, J., Turk, M., and T. Hollerer, "Model estimation and selection towards unconstrained real-time tracking and mapping", In IEEE transactions on visualization and computer graphics, 20(6), pp. 825-838, 2013.
- [SLAM_4] Pirchheim, C., Schmalstieg, D., and G. Reitmayr, "Handling pure camera rotation in keyframe-based SLAM", In 2013 IEEE international symposium on mixed and augmented reality (ISMAR), pp. 229-238, 2013.
- [UBICOMP] Bardram, J. and A. Friday, "Ubiquitous Computing Systems", In Ubiquitous Computing Fundamentals pp. 37-94. CRC Press, 2009.
- [URLLC] 3GPP, "3GPP TR 23.725: Study on enhancement of Ultra-Reliable Low-Latency Communication (URLLC) support in the 5G Core network (5GC).", <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3453>, 2019.
- [VIS_INTERFERE] Kalkofen, D., Mendez, E., and D. Schmalstieg, "Interactive focus and context visualization for augmented reality.", In 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 191-201., 2007.
- [XR] 3GPP, "3GPP TR 26.928: Extended Reality (XR) in 5G.", <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3534>, 2020.

Authors' Addresses

Renan Krishna
InterDigital Europe Limited
64, Great Eastern Street
London
EC2A 3QR
United Kingdom
Email: renan.krishna@interdigital.com

Akbar Rahman
InterDigital Communications, LLC
1000 Sherbrooke Street West
Montreal H3A 3G4
Canada
Email: Akbar.Rahman@InterDigital.com

MOPS
Internet-Draft
Intended status: Informational
Expires: 23 October 2022

J. Holland
Akamai Technologies, Inc.
A. Begen
Networked Media
S. Dawkins
Tencent America LLC
21 April 2022

Operational Considerations for Streaming Media
draft-ietf-mops-streaming-opcons-10

Abstract

This document provides an overview of operational networking issues that pertain to quality of experience when streaming video and other high-bitrate media over the Internet.

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 23 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
1.1. Notes for Contributors and Reviewers	4
1.1.1. Venues for Contribution and Discussion	4
2. Our Focus on Streaming Video	5
3. Bandwidth Provisioning	6
3.1. Scaling Requirements for Media Delivery	6
3.1.1. Video Bitrates	6
3.1.2. Virtual Reality Bitrates	6
3.2. Path Bandwidth Constraints	7
3.2.1. Recognizing Changes from an Expected Baseline	8
3.3. Path Requirements	9
3.4. Caching Systems	9
3.5. Predictable Usage Profiles	11
3.6. Unpredictable Usage Profiles	11
3.7. Extremely Unpredictable Usage Profiles	12
4. Latency Considerations	14
4.1. Ultra Low-Latency	14
4.2. Low-Latency Live	15
4.3. Non-Low-Latency Live	16
4.4. On-Demand	16
5. Adaptive Encoding, Adaptive Delivery, and Measurement Collection	17
5.1. Overview	17
5.2. Adaptive Encoding	18
5.3. Adaptive Segmented Delivery	18
5.4. Advertising	18
5.5. Bitrate Detection Challenges	20
5.5.1. Idle Time between Segments	21
5.5.2. Head-of-Line Blocking	21
5.5.3. Wide and Rapid Variation in Path Capacity	22
5.6. Measurement Collection	22
6. Evolution of Transport Protocols and Transport Protocol Behaviors	23
6.1. UDP and Its Behavior	24
6.2. TCP and Its Behavior	25
6.3. QUIC and Its Behavior	26
7. Streaming Encrypted Media	28
7.1. General Considerations for Media Encryption	29
7.2. Considerations for "Hop-by-Hop" Media Encryption	30
7.3. Considerations for "End-to-End" Media Encryption	32
8. Further Reading and References	32
9. IANA Considerations	33
10. Security Considerations	33
11. Acknowledgments	33
12. Informative References	33
Authors' Addresses	41

1. Introduction

This document examines networking and transport protocol issues as they relate to quality of experience (QOE) in Internet media delivery, especially focusing on capturing characteristics of streaming video delivery that have surprised network designers or transport experts who lack specific video expertise, since streaming media highlights key differences between common assumptions in existing networking practices and observations of video delivery issues encountered when streaming media over those existing networks.

This document specifically focuses on streaming applications and defines streaming as follows:

- * Streaming is transmission of a continuous media from a server to a client and its simultaneous consumption by the client.
- * Here, "continuous media" refers to media and associated streams such as video, audio, metadata, etc. In this definition, the critical term is "simultaneous", as it is not considered streaming if one downloads a video file and plays it after the download is completed, which would be called download-and-play.

This has two implications.

- * First, the server's transmission rate must (loosely or tightly) match to client's consumption rate in order to provide uninterrupted playback. That is, the client must not run out of data (buffer underrun) or accept more data than it can buffer before playback (buffer overrun) as any excess media that cannot be buffered is simply discarded.
- * Second, the client's consumption rate is limited not only by bandwidth availability, but also media availability. The client cannot fetch media that is not available from a server yet.

This document contains

- * A short description of streaming video characteristics in Section 2, to set the stage for the rest of the document,
- * General guidance on bandwidth provisioning (Section 3) and latency considerations (Section 4) for streaming video delivery,

- * A description of adaptive encoding and adaptive delivery techniques in common use for streaming video, along with a description of the challenges media senders face in detecting the bitrate available between the media sender and media receiver, and collection of measurements by a third party for use in analytics (Section 5),
- * A description of existing transport protocols used for video streaming, and the issues encountered when using those protocols, along with a description of the QUIC transport protocol [RFC9000] that we expect to be used for streaming media (Section 6),
- * A description of implications when streaming encrypted media (Section 7), and
- * A number of useful pointers for further reading on this rapidly changing subject (Section 8).

Making specific recommendations on operational practices aimed at mitigating the issues described in this document is out of scope, though some existing mitigations are mentioned in passing. The intent is to provide a point of reference for future solution proposals to use in describing how new technologies address or avoid existing observed problems.

1.1. Notes for Contributors and Reviewers

Note to RFC Editor: Please remove this section and its subsections before publication.

This section is to provide references to make it easier to review the development and discussion on the draft so far.

1.1.1. Venues for Contribution and Discussion

This document is in the Github repository at:

<https://github.com/ietf-wg-mops/draft-ietf-mops-streaming-opcons>
(<https://github.com/ietf-wg-mops/draft-ietf-mops-streaming-opcons>)

Readers are welcome to open issues and send pull requests for this document.

Substantial discussion of this document should take place on the MOPS working group mailing list (mops@ietf.org).

- * Join: <https://www.ietf.org/mailman/listinfo/mops>
(<https://www.ietf.org/mailman/listinfo/mops>)

- * Search: <https://mailarchive.ietf.org/arch/browse/mops/>
(<https://mailarchive.ietf.org/arch/browse/mops/>)

2. Our Focus on Streaming Video

As the internet has grown, an increasingly large share of the traffic delivered to end users has become video. The most recent available estimates found that 75% of the total traffic to end users was video in 2019. At that time, the share of traffic that was video had been growing for years and was projected to continue growing (Appendix D of [CVNI]).

A substantial part of this growth is due to increased use of streaming video, although the amount of video traffic in real-time communications (for example, online videoconferencing) has also grown significantly. While both streaming video and videoconferencing have real-time delivery and latency requirements, these requirements vary from one application to another. For additional discussion of latency requirements, see Section 4.

In many contexts, video traffic can be handled transparently as generic application-level traffic. However, as the volume of video traffic continues to grow, it is becoming increasingly important to consider the effects of network design decisions on application-level performance, with considerations for the impact on video delivery.

Much of the focus of this document is on reliable media using HTTP. HTTP is widely used because

- * support for HTTP is widely available in a wide range of operating systems,
- * HTTP is also used in a wide variety of other applications,
- * HTTP has been demonstrated to provide acceptable performance over the open Internet,
- * HTTP includes state of the art standardized security mechanisms, and
- * HTTP can make use of already-deployed caching infrastructure such as CDNs (Content Delivery Networks), local proxies, and browser caches.

Various HTTP versions have been used for media delivery. HTTP/1.0, HTTP/1.1 and HTTP/2 are carried over TCP, and TCP's transport behavior is described in Section 6.2. HTTP/3 is carried over QUIC, and QUIC's transport behavior is described in Section 6.3.

Unreliable media delivery using RTP and other UDP-based protocols is also discussed in Section 4.1, Section 6.1, and Section 7.2, but it is difficult to give general guidance for these applications. For instance, when loss occurs, the most appropriate response may depend on the type of codec being used.

3. Bandwidth Provisioning

3.1. Scaling Requirements for Media Delivery

3.1.1. Video Bitrates

Video bitrate selection depends on many variables including the resolution (height and width), frame rate, color depth, codec, encoding parameters, scene complexity and amount of motion. Generally speaking, as the resolution, frame rate, color depth, scene complexity and amount of motion increase, the encoding bitrate increases. As newer codecs with better compression tools are used, the encoding bitrate decreases. Similarly, a multi-pass encoding generally produces better quality output compared to single-pass encoding at the same bitrate, or delivers the same quality at a lower bitrate.

Here are a few common resolutions used for video content, with typical ranges of bitrates for the two most popular video codecs [Encodings].

Name	Width x Height	H.264	H.265
DVD	720 x 480	1.0 Mbps	0.5 Mbps
720p (1K)	1280 x 720	3-4.5 Mbps	2-4 Mbps
1080p (2K)	1920 x 1080	6-8 Mbps	4.5-7 Mbps
2160p (4k)	3840 x 2160	N/A	10-20 Mbps

Table 1

3.1.2. Virtual Reality Bitrates

The bitrates given in Section 3.1.1 describe video streams that provide the user with a single, fixed, point of view - so, the user has no "degrees of freedom", and the user sees all of the video image that is available.

Even basic virtual reality (360-degree) videos that allow users to look around freely (referred to as "three degrees of freedom", or 3DoF) require substantially larger bitrates when they are captured and encoded as such videos require multiple fields of view of the scene. Yet, due to smart delivery methods such as viewport-based or tiled-based streaming, we do not need to send the whole scene to the user. Instead, the user needs only the portion corresponding to its viewpoint at any given time ([Survey360o]).

In more immersive applications, where limited user movement ("three degrees of freedom plus", or 3DoF+) or full user movement ("six degrees of freedom", or 6DoF) is allowed, the required bitrate grows even further. In this case, immersive content is typically referred to as volumetric media. One way to represent the volumetric media is to use point clouds, where streaming a single object may easily require a bitrate of 30 Mbps or higher. Refer to [MPEGI] and [PCC] for more details.

3.2. Path Bandwidth Constraints

Even when the bandwidth requirements for video streams along a path are well understood, additional analysis is required to understand the constraints on bandwidth at various points in the network. This analysis is necessary because media servers may react to bandwidth constraints using two independent feedback loops:

- * Media servers often respond to application-level feedback from the media player that indicates a bottleneck link somewhere along the path, by adjusting the amount of media that the media server will send to the media player in a given timeframe. This is described in greater detail in Section 5.
- * Media servers also typically implement transport protocols with capacity-seeking congestion controllers that probe for bandwidth, and adjust the sending rate based on transport mechanisms. This is described in greater detail in Section 6.

The result is that these two (potentially competing) "helpful" mechanisms each respond to the same bottleneck with no coordination between themselves, so that each is unaware of actions taken by the other, and this can result in QOE for users that is significantly lower than what could have been achieved.

In one example, if a media server overestimates the available bandwidth to the media player,

- * the transport protocol detects loss due to congestion, and reduces its sending window size per round trip,

- * the media server adapts to application-level feedback from the media player, and reduces its own sending rate,
- * the transport protocol sends media at the new, lower rate, and confirms that this new, lower rate is "safe", because no transport-level loss is occurring, but
- * because the media server continues to send at the new, lower rate, the transport protocol's maximum sending rate is now limited by the amount of information the media server queues for transmission, so
- * the transport protocol can't probe for available path bandwidth by sending at a higher rate.

In order to avoid these types of situations, which can potentially affect all the users whose streaming media traverses a bottleneck link, there are several possible mitigations that streaming operators can use, but the first step toward mitigating a problem is knowing when that problem occurs.

3.2.1. Recognizing Changes from an Expected Baseline

There are many reasons why path characteristics might change suddenly, for example,

- * "cross traffic" that traverses part of the path, especially if this traffic is "inelastic", and does not, itself, respond to indications of path congestion.
- * routing changes, which can happen in normal operation, especially if the new path now includes path segments that are more heavily loaded, offer lower total bandwidth, or simply cover more distance.

In order to recognize that a path carrying streaming media is "not behaving the way it normally does", having an expected baseline that describes "the way it normally does" is fundamental. Analytics that aid in that recognition can be more or less sophisticated, and can be as simple as noticing that the apparent round trip times for media traffic carried over TCP transport on some paths are suddenly and significantly longer than usual. Passive monitors can detect changes in the elapsed time between the acknowledgements for specific TCP segments from a TCP receiver, since TCP octet sequence numbers and acknowledgements for those sequence numbers are "carried in the clear", even if the TCP payload itself is encrypted. See Section 6.2 for more information.

As transport protocols evolve to encrypt their transport header fields, one side effect of increasing encryption is that the kind of passive monitoring, or even "performance enhancement" ([RFC3135]) that was possible with the older transport protocols (UDP, described in Section 6.1 and TCP, described in Section 6.2) is no longer possible with newer transport protocols such as QUIC (described in Section 6.3). The IETF has specified a "latency spin bit" mechanism in Section 17.4 of [RFC9000] to allow passive latency monitoring from observation points on the network path throughout the duration of a connection, but currently chartered work in the IETF is focusing on end-point monitoring and reporting, rather than on passive monitoring.

One example is the "qlog" mechanism [I-D.ietf-quic-qlog-main-schema], a protocol-agnostic mechanism used to provide better visibility for encrypted protocols such as QUIC ([I-D.ietf-quic-qlog-quic-events]) and for HTTP/3 ([I-D.ietf-quic-qlog-h3-events]).

3.3. Path Requirements

The bitrate requirements in Section 3.1 are per end-user actively consuming a media feed, so in the worst case, the bitrate demands can be multiplied by the number of simultaneous users to find the bandwidth requirements for a router on the delivery path with that number of users downstream. For example, at a node with 10,000 downstream users simultaneously consuming video streams, approximately 80 Gbps might be necessary in order for all of them to get typical content at 1080p resolution.

However, when there is some overlap in the feeds being consumed by end users, it is sometimes possible to reduce the bandwidth provisioning requirements for the network by performing some kind of replication within the network. This can be achieved via object caching with delivery of replicated objects over individual connections, and/or by packet-level replication using multicast.

To the extent that replication of popular content can be performed, bandwidth requirements at peering or ingest points can be reduced to as low as a per-feed requirement instead of a per-user requirement.

3.4. Caching Systems

When demand for content is relatively predictable, and especially when that content is relatively static, caching content close to requesters, and pre-loading caches to respond quickly to initial requests is often useful (for example, HTTP/1.1 caching is described in [I-D.ietf-httpbis-cache]). This is subject to the usual considerations for caching - for example, how much data must be

cached to make a significant difference to the requester, and how the benefits of caching and pre-loading caches balances against the costs of tracking "stale" content in caches and refreshing that content.

It is worth noting that not all high-demand content is "live" content. One relevant example is when popular streaming content can be staged close to a significant number of requesters, as can happen when a new episode of a popular show is released. This content may be largely stable, so low-cost to maintain in multiple places throughout the Internet. This can reduce demands for high end-to-end bandwidth without having to use mechanisms like multicast.

Caching and pre-loading can also reduce exposure to peering point congestion, since less traffic crosses the peering point exchanges if the caches are placed in peer networks, especially when the content can be pre-loaded during off-peak hours, and especially if the transfer can make use of "Lower-Effort Per-Hop Behavior (LE PHB) for Differentiated Services" [RFC8622], "Low Extra Delay Background Transport (LEDBAT)" [RFC6817], or similar mechanisms.

All of this depends, of course, on the ability of a content provider to predict usage and provision bandwidth, caching, and other mechanisms to meet the needs of users. In some cases (Section 3.5), this is relatively routine, but in other cases, it is more difficult (Section 3.6, Section 3.7).

And as with other parts of the ecosystem, new technology brings new challenges. For example, with the emergence of ultra-low-latency streaming, responses have to start streaming to the end user while still being transmitted to the cache, and while the cache does not yet know the size of the object. Some of the popular caching systems were designed around cache footprint and had deeply ingrained assumptions about knowing the size of objects that are being stored, so the change in design requirements in long-established systems caused some errors in production. Incidents occurred where a transmission error in the connection from the upstream source to the cache could result in the cache holding a truncated segment and transmitting it to the end user's device. In this case, players rendering the stream often had the video freeze until the player was reset. In some cases the truncated object was even cached that way and served later to other players as well, causing continued stalls at the same spot in the video for all players playing the segment delivered from that cache node.

3.5. Predictable Usage Profiles

Historical data shows that users consume more videos and at a higher bit rate than they did in the past on their connected devices. Improvements in the codecs that help with reducing the encoding bitrates with better compression algorithms could not have offset the increase in the demand for the higher quality video (higher resolution, higher frame rate, better color gamut, better dynamic range, etc.). In particular, mobile data usage has shown a large jump over the years due to increased consumption of entertainment as well as conversational video.

3.6. Unpredictable Usage Profiles

Although TCP/IP has been used with a number of widely used applications that have symmetric bandwidth requirements (similar bandwidth requirements in each direction between endpoints), many widely-used Internet applications operate in client-server roles, with asymmetric bandwidth requirements. A common example might be an HTTP GET operation, where a client sends a relatively small HTTP GET request for a resource to an HTTP server, and often receives a significantly larger response carrying the requested resource. When HTTP is commonly used to stream movie-length video, the ratio between response size and request size can become arbitrarily large.

For this reason, operators may pay more attention to downstream bandwidth utilization when planning and managing capacity. In addition, operators have been able to deploy access networks for end users using underlying technologies that are inherently asymmetric, favoring downstream bandwidth (e.g. ADSL, cellular technologies, most IEEE 802.11 variants), assuming that users will need less upstream bandwidth than downstream bandwidth. This strategy usually works, except when it fails because application bandwidth usage patterns have changed in ways that were not predicted.

One example of this type of change was when peer-to-peer file sharing applications gained popularity in the early 2000s. To take one well-documented case ([RFC5594]), the Bittorrent application created "swarms" of hosts, uploading and downloading files to each other, rather than communicating with a server. Bittorrent favored peers who uploaded as much as they downloaded, so that new Bittorrent users had an incentive to significantly increase their upstream bandwidth utilization.

The combination of the large volume of "torrents" and the peer-to-peer characteristic of swarm transfers meant that end user hosts were suddenly uploading higher volumes of traffic to more destinations than was the case before Bittorrent. This caused at least one large

Internet service provider (ISP) to attempt to "throttle" these transfers in order to to mitigate the load that these hosts placed on their network. These efforts were met by increased use of encryption in Bittorrent, and complaints to regulators calling for regulatory action.

The BitTorrent case study is just one example, but the example is included here to make it clear that unpredicted and unpredictable massive traffic spikes may not be the result of natural disasters, but they can still have significant impacts.

Especially as end users increase use of video-based social networking applications, it will be helpful for access network providers to watch for increasing numbers of end users uploading significant amounts of content.

3.7. Extremely Unpredictable Usage Profiles

The causes of unpredictable usage described in Section 3.6 were more or less the result of human choices, but we were reminded during a post-IETF 107 meeting that humans are not always in control, and forces of nature can cause enormous fluctuations in traffic patterns.

In his talk, Sanjay Mishra [Mishra] reported that after the CoViD-19 pandemic broke out in early 2020,

- * Comcast's streaming and web video consumption rose by 38%, with their reported peak traffic up 32% overall between March 1 to March 30,
- * AT&T reported a 28% jump in core network traffic (single day in April, as compared to pre stay-at-home daily average traffic), with video accounting for nearly half of all mobile network traffic, while social networking and web browsing remained the highest percentage (almost a quarter each) of overall mobility traffic, and
- * Verizon reported similar trends with video traffic up 36% over an average day (pre COVID-19).

We note that other operators saw similar spikes during this time period. Craig Labowitz [Labovitz] reported

- * Weekday peak traffic increases over 45%-50% from pre-lockdown levels,
- * A 30% increase in upstream traffic over their pre-pandemic levels, and

- * A steady increase in the overall volume of DDoS traffic, with amounts exceeding the pre-pandemic levels by 40%. (He attributed this increase to the significant rise in gaming-related DDoS attacks ([LabovitzDDoS]), as gaming usage also increased.)

Subsequently, the Internet Architecture Board (IAB) held a COVID-19 Network Impacts Workshop [IABcovid] in November 2020. Given a larger number of reports and more time to reflect, the following observations from the draft workshop report are worth considering.

- * Participants describing different types of networks reported different kinds of impacts, but all types of networks saw impacts.
- * Mobile networks saw traffic reductions and residential networks saw significant increases.
- * Reported traffic increases from ISPs and Internet Exchange Points (IXP) over just a few weeks were as big as the traffic growth over the course of a typical year, representing a 15-20% surge in growth to land at a new normal that was much higher than anticipated.
- * At DE-CIX Frankfurt, the world's largest Internet Exchange Point in terms of data throughput, the year 2020 has seen the largest increase in peak traffic within a single year since the IXP was founded in 1995.
- * The usage pattern changed significantly as work-from-home and videoconferencing usage peaked during normal work hours, which would have typically been off-peak hours with adults at work and children at school. One might expect that the peak would have had more impact on networks if it had happened during typical evening peak hours for video streaming applications.
- * The increase in daytime bandwidth consumption reflected both significant increases in "essential" applications such as videoconferencing and virtual private networks (VPN), and entertainment applications as people watched videos or played games.
- * At the IXP level, it was observed that physical link utilization increased. This phenomenon could probably be explained by a higher level of uncacheable traffic such as videoconferencing and VPNs from residential users as they stopped commuting and switched to work-at-home.

4. Latency Considerations

Streaming media latency refers to the "glass-to-glass" time duration, which is the delay between the real-life occurrence of an event and the streamed media being appropriately displayed on an end user's device. Note that this is different from the network latency (defined as the time for a packet to cross a network from one end to another end) because it includes video encoding/decoding and buffering time, and for most cases also ingest to an intermediate service such as a CDN or other video distribution service, rather than a direct connection to an end user.

Streaming media can be usefully categorized according to the application's latency requirements into a few rough categories:

- * ultra low-latency (less than 1 second)
- * low-latency live (less than 10 seconds)
- * non-low-latency live (10 seconds to a few minutes)
- * on-demand (hours or more)

4.1. Ultra Low-Latency

Ultra low-latency delivery of media is defined here as having a glass-to-glass delay target under one second.

Some media content providers aim to achieve this level of latency for live media events. This introduces new challenges relative to less-restricted levels of latency requirements because this latency is the same scale as commonly observed end-to-end network latency variation (for example, due to effects such as bufferbloat ([CoDel]), Wi-Fi error correction, or packet reordering). These effects can make it difficult to achieve this level of latency for the general case, and may require tradeoffs in relatively frequent user-visible media artifacts. However, for controlled environments or targeted networks that provide mitigations against such effects, this level of latency is potentially achievable with the right provisioning.

Applications requiring ultra low latency for media delivery are usually tightly constrained on the available choices for media transport technologies, and sometimes may need to operate in controlled environments to reliably achieve their latency and quality goals.

Most applications operating over IP networks and requiring latency this low use the Real-time Transport Protocol (RTP) [RFC3550] or WebRTC [RFC8825], which uses RTP for the media transport as well as several other protocols necessary for safe operation in browsers.

Worth noting is that many applications for ultra low-latency delivery do not need to scale to more than a few users at a time, which simplifies many delivery considerations relative to other use cases.

Recommended reading for applications adopting an RTP-based approach also includes [RFC7656]. For increasing the robustness of the playback by implementing adaptive playout methods, refer to [RFC4733] and [RFC6843].

Applications with further-specialized latency requirements are out of scope for this document.

4.2. Low-Latency Live

Low-latency live delivery of media is defined here as having a glass-to-glass delay target under 10 seconds.

This level of latency is targeted to have a user experience similar to traditional broadcast TV delivery. A frequently cited problem with failing to achieve this level of latency for live sporting events is the user experience failure from having crowds within earshot of one another who react audibly to an important play, or from users who learn of an event in the match via some other channel, for example social media, before it has happened on the screen showing the sporting event.

Applications requiring low-latency live media delivery are generally feasible at scale with some restrictions. This typically requires the use of a premium service dedicated to the delivery of live video, and some tradeoffs may be necessary relative to what is feasible in a higher latency service. The tradeoffs may include higher costs, or delivering a lower quality video, or reduced flexibility for adaptive bitrates, or reduced flexibility for available resolutions so that fewer devices can receive an encoding tuned for their display. Low-latency live delivery is also more susceptible to user-visible disruptions due to transient network conditions than higher latency services.

Implementation of a low-latency live video service can be achieved with the use of low-latency extensions of HLS (called LL-HLS) [I-D.draft-pantos-hls-rfc8216bis] and DASH (called LL-DASH) [LL-DASH]. These extensions use the Common Media Application Format (CMAF) standard [MPEG-CMAF] that allows the media to be packaged into

and transmitted in units smaller than segments, which are called chunks in CMAF language. This way, the latency can be decoupled from the duration of the media segments. Without a CMAF-like packaging, lower latencies can only be achieved by using very short segment durations. However, shorter segments means more frequent intra-coded frames and that is detrimental to video encoding quality. CMAF allows us to still use longer segments (improving encoding quality) without penalizing latency.

While a LL-HLS client retrieves each chunk with a separate HTTP GET request, a LL-DASH client uses the chunked transfer encoding feature of the HTTP [CMAF-CTE] which allows the LL-DASH client to fetch all the chunks belonging to a segment with a single GET request. An HTTP server can transmit the CMAF chunks to the LL-DASH client as they arrive from the encoder/packager. A detailed comparison of LL-HLS and LL-DASH is given in [MMSP20].

4.3. Non-Low-Latency Live

Non-low-latency live delivery of media is defined here as a live stream that does not have a latency target shorter than 10 seconds.

This level of latency is the historically common case for segmented video delivery using HLS [RFC8216] and DASH [MPEG-DASH]. This level of latency is often considered adequate for content like news or pre-recorded content. This level of latency is also sometimes achieved as a fallback state when some part of the delivery system or the client-side players do not have the necessary support for the features necessary to support low-latency live streaming.

This level of latency can typically be achieved at scale with commodity CDN services for HTTP(s) delivery, and in some cases the increased time window can allow for production of a wider range of encoding options relative to the requirements for a lower latency service without the need for increasing the hardware footprint, which can allow for wider device interoperability.

4.4. On-Demand

On-Demand media streaming refers to playback of pre-recorded media based on a user's action. In some cases on-demand media is produced as a by-product of a live media production, using the same segments as the live event, but freezing the manifest after the live event has finished. In other cases, on-demand media is constructed out of pre-recorded assets with no streaming necessarily involved during the production of the on-demand content.

On-demand media generally is not subject to latency concerns, but other timing-related considerations can still be as important or even more important to the user experience than the same considerations with live events. These considerations include the startup time, the stability of the media stream's playback quality, and avoidance of stalls and video artifacts during the playback under all but the most severe network conditions.

In some applications, optimizations are available to on-demand video that are not always available to live events, such as pre-loading the first segment for a startup time that doesn't have to wait for a network download to begin.

5. Adaptive Encoding, Adaptive Delivery, and Measurement Collection

5.1. Overview

A simple model of video playback can be described as a video stream consumer, a buffer, and a transport mechanism that fills the buffer. The consumption rate is fairly static and is represented by the content bitrate. The size of the buffer is also commonly a fixed size. The fill process needs to be at least fast enough to ensure that the buffer is never empty, however it also can have significant complexity when things like personalization or ad workflows are introduced.

The challenges in filling the buffer in a timely way fall into two broad categories: 1. content selection and 2. content variation. Content selection comprises all of the steps needed to determine which content variation to offer the client. Content variation is the number of content options that exist at any given selection point. A common example, easily visualized, is Adaptive BitRate (ABR), described in more detail below. The mechanism used to select the bitrate is part of the content selection, and the content variation are all of the different bitrate renditions.

ABR is a sort of application-level response strategy in which the streaming client attempts to detect the available bandwidth of the network path by observing the successful application-layer download speed, then chooses a bitrate for each of the video, audio, subtitles and metadata (among the limited number of available options) that fits within that bandwidth, typically adjusting as changes in available bandwidth occur in the network or changes in capabilities occur during the playback (such as available memory, CPU, display size, etc.).

5.2. Adaptive Encoding

Media servers can provide media streams at various bitrates because the media has been encoded at various bitrates. This is a so-called "ladder" of bitrates, that can be offered to media players as part of the manifest that describes the media being requested by the media player, so that the media player can select among the available bitrate choices.

The media server may also choose to alter which bitrates are made available to players by adding or removing bitrate options from the ladder delivered to the player in subsequent manifests built and sent to the player. This way, both the player, through its selection of bitrate to request from the manifest, and the server, through its construction of the bitrates offered in the manifest, are able to affect network utilization.

5.3. Adaptive Segmented Delivery

ABR playback is commonly implemented by streaming clients using HLS [RFC8216] or DASH [MPEG-DASH] to perform a reliable segmented delivery of media over HTTP. Different implementations use different strategies [ABRSurvey], often relying on proprietary algorithms (called rate adaptation or bitrate selection algorithms) to perform available bandwidth estimation/prediction and the bitrate selection.

Many systems will do an initial probe or a very simple throughput speed test at the start of a video playback. This is done to get a rough sense of the highest video bitrate in the ABR ladder that the network between the server and player will likely be able to provide under initial network conditions. After the initial testing, clients tend to rely upon passive network observations and will make use of player side statistics such as buffer fill rates to monitor and respond to changing network conditions.

The choice of bitrate occurs within the context of optimizing for one or more metrics monitored by the client, such as highest achievable video quality or lowest chances for a rebuffering event (playback stall).

5.4. Advertising

A variety of business models exist for producers of streaming media. Some content providers derive the majority of the revenue associated with streaming media directly from consumer subscriptions or one-time purchases. Others derive the majority of their streaming media associated revenue from advertising. Many content providers derive income from a mix of these and other sources of funding. The

inclusion of advertising alongside or interspersed with streaming media content is therefore common in today's media landscape.

Some commonly used forms of advertising can introduce potential user experience issues for a media stream. This section provides a very brief overview of a complex and evolving space, but a complete coverage of the potential issues is out of scope for this document.

The same techniques used to allow a media player to switch between renditions of different bitrates at segment or chunk boundaries can also be used to enable the dynamic insertion of advertisements (hereafter referred to as "ads").

Ads may be inserted either with Client Side Ad Insertion (CSAI) or Server Side Ad Insertion (SSAI). In CSAI, the ABR manifest will generally include links to an external ad server for some segments of the media stream, while in SSAI the server will remain the same during advertisements, but will include media segments that contain the advertising. In SSAI, the media segments may or may not be sourced from an external ad server like with CSAI.

In general, the more targeted the ad request is, the more requests the ad service needs to be able to handle concurrently. If connectivity is poor to the ad service, this can cause rebuffering even if the underlying video assets (both content and ads) are able to be accessed quickly. The less targeted, the more likely the ad requests can be consolidated and can leverage the same caching techniques as the video content.

In some cases, especially with SSAI, advertising space in a stream is reserved for a specific advertiser and can be integrated with the video so that the segments share the same encoding properties such as bitrate, dynamic range, and resolution. However, in many cases ad servers integrate with a Supply Side Platform (SSP) that offers advertising space in real-time auctions via an Ad Exchange, with bids for the advertising space coming from Demand Side Platforms (DSPs) that collect money from advertisers for delivering the advertisements. Most such Ad Exchanges use application-level protocol specifications published by the Interactive Advertising Bureau [IAB-ADS], an industry trade organization.

This ecosystem balances several competing objectives, and integrating with it naively can produce surprising user experience results. For example, ad server provisioning and/or the bitrate of the ad segments might be different from that of the main video, either of which can sometimes result in video stalls. For another example, since the inserted ads are often produced independently they might have a different base volume level than the main video, which can make for a jarring user experience.

Additionally, this market historically has had incidents of ad fraud (misreporting of ad delivery to end users for financial gain). As a mitigation for concerns driven by those incidents, some SSPs have required the use of players with features like reporting of ad delivery, or providing information that can be used for user tracking. Some of these and other measures have raised privacy concerns for end users.

In general this is a rapidly developing space with many considerations, and media streaming operators engaged in advertising may need to research these and other concerns to find solutions that meet their user experience, user privacy, and financial goals. For further reading on mitigations, [BAP] has published some standards and best practices based on user experience research.

5.5. Bitrate Detection Challenges

This kind of bandwidth-measurement system can experience trouble in several ways that are affected by networking and transport protocol issues. Because adaptive application-level response strategies are often using rates as observed by the application layer, there are sometimes inscrutable transport-level protocol behaviors that can produce surprising measurement values when the application-level feedback loop is interacting with a transport-level feedback loop.

A few specific examples of surprising phenomena that affect bitrate detection measurements are described in the following subsections. As these examples will demonstrate, it is common to encounter cases that can deliver application level measurements that are too low, too high, and (possibly) correct but varying more quickly than a lab-tested selection algorithm might expect.

These effects and others that cause transport behavior to diverge from lab modeling can sometimes have a significant impact on bitrate selection and on user QOE, especially where players use naive measurement strategies and selection algorithms that don't account for the likelihood of bandwidth measurements that diverge from the true path capacity.

5.5.1. Idle Time between Segments

When the bitrate selection is chosen substantially below the available capacity of the network path, the response to a segment request will typically complete in much less absolute time than the duration of the requested segment, leaving significant idle time between segment downloads. This can have a few surprising consequences:

- * TCP slow-start when restarting after idle requires multiple RTTs to re-establish a throughput at the network's available capacity. When the active transmission time for segments is substantially shorter than the time between segments, leaving an idle gap between segments that triggers a restart of TCP slow-start, the estimate of the successful download speed coming from the application-visible receive rate on the socket can thus end up much lower than the actual available network capacity. This in turn can prevent a shift to the most appropriate bitrate. [RFC7661] provides some mitigations for this effect at the TCP transport layer, for senders who anticipate a high incidence of this problem.
- * Mobile flow-bandwidth spectrum and timing mapping can be impacted by idle time in some networks. The carrier capacity assigned to a link can vary with activity. Depending on the idle time characteristics, this can result in a lower available bitrate than would be achievable with a steadier transmission in the same network.

Some receiver-side ABR algorithms such as [ELASTIC] are designed to try to avoid this effect.

Another way to mitigate this effect is by the help of two simultaneous TCP connections, as explained in [MMSys11] for Microsoft Smooth Streaming. In some cases, the system-level TCP slow-start restart can also be disabled, for example as described in [OReilly-HPBN].

5.5.2. Head-of-Line Blocking

In the event of a lost packet on a TCP connection with SACK support (a common case for segmented delivery in practice), loss of a packet can provide a confusing bandwidth signal to the receiving application. Because of the sliding window in TCP, many packets may be accepted by the receiver without being available to the application until the missing packet arrives. Upon arrival of the one missing packet after retransmit, the receiver will suddenly get access to a lot of data at the same time.

To a receiver measuring bytes received per unit time at the application layer, and interpreting it as an estimate of the available network bandwidth, this appears as a high jitter in the goodput measurement, presenting as a stall, followed by a sudden leap that can far exceed the actual capacity of the transport path from the server when the hole in the received data is filled by a later retransmission.

It is worth noting that more modern transport protocols such as QUIC have mitigation of head-of-line blocking as a protocol design goal. See Section 6.3 for more details.

5.5.3. Wide and Rapid Variation in Path Capacity

As many end devices have moved to wireless connectivity for the final hop (Wi-Fi, 5G, or LTE), new problems in bandwidth detection have emerged from radio interference and signal strength effects.

Each of these technologies can experience sudden changes in capacity as the end user device moves from place to place and encounters new sources of interference. Microwave ovens, for example, can cause a throughput degradation of more than a factor of 2 while active [Micro]. 5G and LTE likewise can easily see rate variation by a factor of 2 or more over a span of seconds as users move around.

These swings in actual transport capacity can result in user experience issues that can be exacerbated by insufficiently responsive ABR algorithms.

5.6. Measurement Collection

Media players use measurements to guide their segment-by-segment adaptive streaming requests, but may also provide measurements to streaming media providers.

In turn, providers may base analytics on these measurements, to guide decisions such as whether adaptive encoding bitrates in use are the best ones to provide to media players, or whether current media content caching is providing the best experience for viewers.

To that effect, the Consumer Technology Association (CTA) who owns the Web Application Video Ecosystem (WAVE) project has published two important specifications.

- * CTA-2066: Streaming Quality of Experience Events, Properties and Metrics

[CTA-2066] specifies a set of media player events, properties, QOE metrics and associated terminology for representing streaming media QOE across systems, media players and analytics vendors. While all these events, properties, metrics and associated terminology is used across a number of proprietary analytics and measurement solutions, they were used in slightly (or vastly) different ways that led to interoperability issues. CTA-2066 attempts to address this issue by defining a common terminology as well as how each metric should be computed for consistent reporting.

* CTA-5004: Common Media Client Data (CMCD)

Many assume that the CDNs have a holistic view into the health and performance of the streaming clients. However, this is not the case. The CDNs produce millions of log lines per second across hundreds of thousands of clients and they have no concept of a "session" as a client would have, so CDNs are decoupled from the metrics the clients generate and report. A CDN cannot tell which request belongs to which playback session, the duration of any media object, the bitrate, or whether any of the clients have stalled and are rebuffering or are about to stall and will rebuffer. The consequence of this decoupling is that a CDN cannot prioritize delivery for when the client needs it most, prefetch content, or trigger alerts when the network itself may be underperforming. One approach to couple the CDN to the playback sessions is for the clients to communicate standardized media-relevant information to the CDNs while they are fetching data. [CTA-5004] was developed exactly for this purpose.

6. Evolution of Transport Protocols and Transport Protocol Behaviors

Because networking resources are shared between users, a good place to start our discussion is how contention between users, and mechanisms to resolve that contention in ways that are "fair" between users, impact streaming media users. These topics are closely tied to transport protocol behaviors.

As noted in Section 5, ABR response strategies such as HLS [RFC8216] or DASH [MPEG-DASH] are attempting to respond to changing path characteristics, and underlying transport protocols are also attempting to respond to changing path characteristics.

For most of the history of the Internet, these transport protocols, described in Section 6.1 and Section 6.2, have had relatively consistent behaviors that have changed slowly, if at all, over time. Newly standardized transport protocols like QUIC [RFC9000] can behave differently from existing transport protocols, and these behaviors may evolve over time more rapidly than currently-used transport protocols.

For this reason, we have included a description of how the path characteristics that streaming media providers may see are likely to evolve over time.

6.1. UDP and Its Behavior

For most of the history of the Internet, we have trusted UDP-based applications to limit their impact on other users. One of the strategies used was to use UDP for simple query-response application protocols, such as DNS, which is often used to send a single-packet request to look up the IP address for a DNS name, and return a single-packet response containing the IP address. Although it is possible to saturate a path between a DNS client and DNS server with DNS requests, in practice, that was rare enough that DNS included few mechanisms to resolve contention between DNS users and other users (whether they are also using DNS, or using other application protocols that share the same pathways).

In recent times, the usage of UDP-based applications that were not simple query-response protocols has grown substantially, and since UDP does not provide any feedback mechanism to senders to help limit impacts on other users, application-level protocols such as RTP [RFC3550] have been responsible for the decisions that TCP-based applications have delegated to TCP - what to send, how much to send, and when to send it. Because UDP itself has no transport-layer feedback mechanisms, UDP-based applications that send and receive substantial amounts of information are expected to provide their own feedback mechanisms, and to respond to the feedback the application receives. This expectation is most recently codified in Best Current Practice [RFC8085].

In contrast to adaptive segmented delivery over a reliable transport as described in Section 5.3, some applications deliver streaming media using an unreliable transport, and rely on a variety of approaches, including:

- * raw MPEG Transport Stream ("MPEG-TS")-formatted video [MPEG-TS] over UDP, which makes no attempt to account for reordering or loss in the transport,
- * RTP [RFC3550], which can notice loss and repair some limited reordering,
- * SCTP [RFC4960], which can use partial reliability [RFC3758] to recover from some loss, but can abandon recovery to limit head-of-line blocking, and

- * SRT [SRT], which can use forward error correction and time-bound retransmission to recover from loss within certain limits, but can abandon recovery to limit head-of-line blocking.

Under congestion and loss, approaches like the above generally experiences transient video artifacts more often and delay of playback effects less often, as compared with reliable segment transport. Often one of the key goals of using a UDP-based transport that allows some unreliability is to reduce latency and better support applications like videoconferencing, or for other live-action video with interactive components, such as some sporting events.

Congestion avoidance strategies for deployments using unreliable transport protocols vary widely in practice, ranging from being entirely unresponsive to congestion, to using feedback signaling to change encoder settings (as in [RFC5762]), to using fewer enhancement layers (as in [RFC6190]), to using proprietary methods to detect QOE issues and turn off video in order to allow less bandwidth-intensive media such as audio to be delivered.

RTP relies on RTCP Sender and Receiver Reports [RFC3550] as its own feedback mechanism, and even includes Circuit Breakers for Unicast RTP Sessions [RFC8083] for situations when normal RTP congestion control has not been able to react sufficiently to RTP flows sending at rates that result in sustained packet loss.

The notion of "Circuit Breakers" has also been applied to other UDP applications in [RFC8084], such as tunneling packets over UDP that are potentially not congestion-controlled (for example, "Encapsulating MPLS in UDP", as described in [RFC7510]). If streaming media is carried in tunnels encapsulated in UDP, these media streams may encounter "tripped circuit breakers", with resulting user-visible impacts.

6.2. TCP and Its Behavior

For most of the history of the Internet, we have trusted TCP to limit the impact of applications that sent a significant number of packets, in either or both directions, on other users. Although early versions of TCP were not particularly good at limiting this impact [RFC0793], the addition of Slow Start and Congestion Avoidance, as described in [RFC2001], were critical in allowing TCP-based applications to "use as much bandwidth as possible, but to avoid using more bandwidth than was possible". Although dozens of RFCs have been written refining TCP decisions about what to send, how much to send, and when to send it, since 1988 [Jacobson-Karels] the signals available for TCP senders remained unchanged - end-to-end acknowledgements for packets that were successfully sent and

received, and packet timeouts for packets that were not.

The success of the largely TCP-based Internet is evidence that the mechanisms TCP used to achieve equilibrium quickly, at a point where TCP senders do not interfere with other TCP senders for sustained periods of time, have been largely successful. The Internet continued to work even when the specific mechanisms used to reach equilibrium changed over time. Because TCP provides a common tool to avoid contention, as some TCP-based applications like FTP were largely replaced by other TCP-based applications like HTTP, the transport behavior remained consistent.

In recent times, the TCP goal of probing for available bandwidth, and "backing off" when a network path is saturated, has been supplanted by the goal of avoiding growing queues along network paths, which prevent TCP senders from reacting quickly when a network path is saturated. Congestion control mechanisms such as COPA [COPA18] and BBR [I-D.cardwell-iccr-g-bbr-congestion-control] make these decisions based on measured path delays, assuming that if the measured path delay is increasing, the sender is injecting packets onto the network path faster than the receiver can accept them, so the sender should adjust its sending rate accordingly.

Although TCP behavior has changed over time, the common practice of implementing TCP as part of an operating system kernel has acted to limit how quickly TCP behavior can change. Even with the widespread use of automated operating system update installation on many end-user systems, streaming media providers could have a reasonable expectation that they could understand TCP transport protocol behaviors, and that those behaviors would remain relatively stable in the short term.

6.3. QUIC and Its Behavior

The QUIC protocol, developed from a proprietary protocol into an IETF standards-track protocol [RFC9000], turns many of the statements made in Section 6.1 and Section 6.2 on their heads.

Although QUIC provides an alternative to the TCP and UDP transport protocols, QUIC is itself encapsulated in UDP. As noted elsewhere in Section 7.1, the QUIC protocol encrypts almost all of its transport parameters, and all of its payload, so any intermediaries that network operators may be using to troubleshoot HTTP streaming media performance issues, perform analytics, or even intercept exchanges in current applications will not work for QUIC-based applications without making changes to their networks. Section 7 describes the implications of media encryption in more detail.

While QUIC is designed as a general-purpose transport protocol, and can carry different application-layer protocols, the current standardized mapping is for HTTP/3 [I-D.ietf-quic-http], which describes how QUIC transport features are used for HTTP. The convention is for HTTP/3 to run over UDP port 443 [Port443] but this is not a strict requirement.

When HTTP/3 is encapsulated in QUIC, which is then encapsulated in UDP, streaming operators (and network operators) might see UDP traffic patterns that are similar to HTTP(S) over TCP. Since earlier versions of HTTP(S) rely on TCP, UDP ports may be blocked for any port numbers that are not commonly used, such as UDP 53 for DNS. Even when UDP ports are not blocked and HTTP/3 can flow, streaming operators (and network operators) may severely rate-limit this traffic because they do not expect to see legitimate high-bandwidth traffic such as streaming media over the UDP ports that HTTP/3 is using.

As noted in Section 5.5.2, because TCP provides a reliable, in-order delivery service for applications, any packet loss for a TCP connection causes "head-of-line blocking", so that no TCP segments arriving after a packet is lost will be delivered to the receiving application until the lost packet is retransmitted, allowing in-order delivery to the application to continue. As described in [RFC9000], QUIC connections can carry multiple streams, and when packet losses do occur, only the streams carried in the lost packet are delayed.

A QUIC extension currently being specified ([I-D.ietf-quic-datagram]) adds the capability for "unreliable" delivery, similar to the service provided by UDP, but these datagrams are still subject to the QUIC connection's congestion controller, providing some transport-level congestion avoidance measures, which UDP does not.

As noted in Section 6.2, there is an increasing interest in transport protocol behaviors that respond to delay measurements, instead of responding to packet loss. These behaviors may deliver improved user experience, but in some cases have not responded to sustained packet loss, which exhausts available buffers along the end-to-end path that may affect other users sharing that path. The QUIC protocol provides a set of congestion control hooks that can be used for algorithm agility, and [RFC9002] defines a basic algorithm with transport behavior that is roughly similar to TCP NewReno [RFC6582]. However, QUIC senders can and do unilaterally choose to use different algorithms such as loss-based CUBIC [RFC8312], delay-based COPA or BBR, or even something completely different.

The Internet community does have experience with deploying new congestion controllers without melting the Internet. As noted in [RFC8312], both the CUBIC congestion controller and its predecessor BIC have significantly different behavior from Reno-style congestion controllers such as TCP NewReno [RFC6582], but both CUBIC and BIC were added to the Linux kernel in order to allow experimentation and analysis, and both were then selected as the default TCP congestion controllers in Linux, and both were deployed globally.

The point mentioned in Section 6.2 about TCP congestion controllers being implemented in operating system kernels is different with QUIC. Although QUIC can be implemented in operating system kernels, one of the design goals when this work was chartered was "QUIC is expected to support rapid, distributed development and testing of features", and to meet this expectation, many implementers have chosen to implement QUIC in user space, outside the operating system kernel, and to even distribute QUIC libraries with their own applications. It is worth noting that streaming operators using HTTP/3, carried over QUIC, can expect more frequent deployment of new congestion controller behavior than has been the case with HTTP/1 and HTTP/2, carried over TCP.

It is worth considering that if TCP-based HTTP traffic and UDP-based HTTP/3 traffic are allowed to enter operator networks on roughly equal terms, questions of fairness and contention will be heavily dependent on interactions between the congestion controllers in use for TCP-based HTTP traffic and UDP-based HTTP/3 traffic.

7. Streaming Encrypted Media

"Encrypted Media" has at least three meanings:

- * Media encrypted at the application layer, typically using some sort of Digital Rights Management (DRM) system, and typically remaining encrypted "at rest", when senders and receivers store it.
- * Media encrypted by the sender at the transport layer, and remaining encrypted until it reaches the ultimate media consumer (in this document, referred to as "end-to-end media encryption").
- * Media encrypted by the sender at the transport layer, and remaining encrypted until it reaches some intermediary that is not the ultimate media consumer, but has credentials allowing decryption of the media content. This intermediary may examine and even transform the media content in some way, before forwarding re-encrypted media content (in this document referred to as "hop-by-hop media encryption").

In this document, we will focus on media encrypted at the transport layer, whether encrypted "hop-by-hop" or "end-to-end". Because media encrypted at the application layer will only be processed by application-level entities, this encryption does not have transport-layer implications. Of course, both "hop-by-hop" and "end-to-end" encrypted transport may carry media that is, in addition, encrypted at the application layer.

Each of these encryption strategies is intended to achieve a different goal. For instance, application-level encryption may be used for business purposes, such as avoiding piracy or enforcing geographic restrictions on playback, while transport-layer encryption may be used to prevent media stream manipulation or to protect manifests.

This document does not take a position on whether those goals are "valid" (whatever that might mean).

Both "end-to-end" and "hop-by-hop" media encryption have specific implications for streaming operators. These are described in Section 7.2 and Section 7.3.

7.1. General Considerations for Media Encryption

The use of strong encryption does provide confidentiality for encrypted streaming media, from the sender to either an intermediary or the ultimate media consumer, and this does prevent Deep Packet Inspection by any intermediary that does not possess credentials allowing decryption. However, even encrypted content streams may be vulnerable to traffic analysis. An intermediary that can identify an encrypted media stream without decrypting it, may be able to "fingerprint" the encrypted media stream of known content, and then match the targeted media stream against the fingerprints of known content. This protection can be lessened if a media provider is repeatedly encrypting the same content. [CODASPY17] is an example of what is possible when identifying HTTPS-protected videos over TCP transport, based either on the length of entire resources being transferred, or on characteristic packet patterns at the beginning of a resource being transferred.

If traffic analysis is successful at identifying encrypted content and associating it with specific users, this breaks privacy as certainly as examining decrypted traffic.

Because HTTPS has historically layered HTTP on top of TLS, which is in turn layered on top of TCP, intermediaries do have access to unencrypted TCP-level transport information, such as retransmissions, and some carriers exploited this information in attempts to improve

transport-layer performance [RFC3135]. The most recent standardized version of HTTPS, HTTP/3 [I-D.ietf-quic-http], uses the QUIC protocol [RFC9000] as its transport layer. QUIC relies on the TLS 1.3 initial handshake [RFC8446] only for key exchange [RFC9001], and encrypts almost all transport parameters itself, with the exception of a few invariant header fields. In the QUIC short header, the only transport-level parameter which is sent "in the clear" is the Destination Connection ID [RFC8999], and even in the QUIC long header, the only transport-level parameters sent "in the clear" are the Version, Destination Connection ID, and Source Connection ID. For these reasons, HTTP/3 is significantly more "opaque" than HTTPS with HTTP/1 or HTTP/2.

[I-D.ietf-quic-manageability] discusses manageability of the QUIC transport protocol that is used to encapsulate HTTP/3, focusing on the implications of QUIC's design and wire image on network operations involving QUIC traffic. It discusses what network operators can consider in some detail.

More broadly, RFC 9065 [RFC9065], "Considerations around Transport Header Confidentiality, Network Operations, and the Evolution of Internet Transport Protocols" describes the impact of increased encryption of transport headers in general terms.

7.2. Considerations for "Hop-by-Hop" Media Encryption

Although the IETF has put considerable emphasis on end-to-end streaming media encryption, there are still important use cases that require the insertion of intermediaries.

There are a variety of ways to involve intermediaries, and some are much more intrusive than others.

From a content provider's perspective, a number of considerations are in play. The first question is likely whether the content provider intends that intermediaries are explicitly addressed from endpoints, or whether the content provider is willing to allow intermediaries to "intercept" streaming content transparently, with no awareness or permission from either endpoint.

If a content provider does not actively work to avoid interception by intermediaries, the effect will be indistinguishable from "impersonation attacks", and endpoints cannot be assumed of any level of privacy.

Assuming that a content provider does intend to allow intermediaries to participate in content streaming, and does intend to provide some level of privacy for endpoints, there are a number of possible tools, either already available or still being specified. These include

- * Server And Network assisted DASH [MPEG-DASH-SAND] - this specification introduces explicit messaging between DASH clients and network elements or between various network elements for the purpose of improving the efficiency of streaming sessions by providing information about real-time operational characteristics of networks, servers, proxies, caches, CDNs, as well as DASH client's performance and status.
- * "Double Encryption Procedures for the Secure Real-Time Transport Protocol (SRTP)" [RFC8723] - this specification provides a cryptographic transform for the Secure Real-time Transport Protocol that provides both hop-by-hop and end-to-end security guarantees.
- * Secure Media Frames [SFRAME] - [RFC8723] is closely tied to SRTP, and this close association impeded widespread deployment, because it could not be used for the most common media content delivery mechanisms. A more recent proposal, Secure Media Frames [SFRAME], also provides both hop-by-hop and end-to-end security guarantees, but can be used with other transport protocols beyond SRTP.

The choice of whether to involve intermediaries sometimes requires careful consideration. As an example, when ABR manifests were commonly sent unencrypted some networks would modify manifests during peak hours by removing high-bitrate renditions in order to prevent players from choosing those renditions, thus reducing the overall bandwidth consumed for delivering these media streams and thereby improving the network load and the user experience for their customers. Now that ubiquitous encryption typically prevents this kind of modification, in order to maintain the same level of network health and user experience across networks whose users would have benefitted from this intervention a media streaming operator sometimes needs to choose between adding intermediaries who are authorized to change the manifests or adding significant extra complexity to their service.

Some resources that might inform other similar considerations are further discussed in [RFC8824] (for WebRTC) and [I-D.ietf-quic-manageability] (for HTTP/3 and QUIC).

7.3. Considerations for "End-to-End" Media Encryption

"End-to-end" media encryption offers the potential of providing privacy for streaming media consumers, with the idea being that if an unauthorized intermediary can't decrypt streaming media, the intermediary can't use Deep Packet Inspection to examine HTTP request and response headers and identify the media content being streamed.

"End-to-end" media encryption has become much more widespread in the years since the IETF issued "Pervasive Monitoring Is an Attack" [RFC7258] as a Best Current Practice, describing pervasive monitoring as a much greater threat than previously appreciated. After the Snowden disclosures, many content providers made the decision to use HTTPS protection - HTTP over TLS - for most or all content being delivered as a routine practice, rather than in exceptional cases for content that was considered "sensitive".

Unfortunately, as noted in [RFC7258], there is no way to prevent pervasive monitoring by an "attacker", while allowing monitoring by a more benign entity who "only" wants to use DPI to examine HTTP requests and responses in order to provide a better user experience. If a modern encrypted transport protocol is used for end-to-end media encryption, intermediary streaming operators are unable to examine transport and application protocol behavior. As described in Section 7.2, only an intermediary streaming operator who is explicitly authorized to examine packet payloads, rather than intercepting packets and examining them without authorization, can continue these practices.

[RFC7258] said that "The IETF will strive to produce specifications that mitigate pervasive monitoring attacks", so streaming operators should expect the IETF's direction toward preventing unauthorized monitoring of IETF protocols to continue for the foreseeable future.

8. Further Reading and References

The Media Operations community maintains a list of references and resources for further reading at this location:

- * <https://github.com/ietf-wg-mops/draft-ietf-mops-streaming-opcons/blob/main/living-doc-mops-streaming-opcons.md>
(<https://github.com/ietf-wg-mops/draft-ietf-mops-streaming-opcons/blob/main/living-doc-mops-streaming-opcons.md>)

Editor's note: The link above might or might not be changed during IESG Evaluation. See <https://github.com/ietf-wg-mops/draft-ietf-mops-streaming-opcons/issues/114> (<https://github.com/ietf-wg-mops/draft-ietf-mops-streaming-opcons/issues/114>) for updates.

9. IANA Considerations

This document requires no actions from IANA.

10. Security Considerations

Security is an important matter for streaming media applications and it was briefly touched on in Section 7.1. This document itself introduces no new security issues.

11. Acknowledgments

Thanks to Alexandre Gouaillard, Aaron Falk, Chris Lemmons, Dave Oran, Eric Vyncke, Glenn Deen, Kyle Rose, Leslie Daigle, Lucas Pardue, Mark Nottingham, Matt Stock, Mike English, Renan Krishna, Roni Even, Sanjay Mishra, and Will Law for very helpful suggestions, reviews and comments.

12. Informative References

[ABRSurvey]

Taani, B., Begen, A. C., Timmerer, C., Zimmermann, R., and A. Bentaleb et al, "A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP", IEEE Communications Surveys & Tutorials , 2019, <<https://ieeexplore.ieee.org/abstract/document/8424813>>.

[BAP]

"The Coalition for Better Ads", n.d., <<https://www.betterads.org/>>.

[CMAF-CTE]

Law, W., "Ultra-Low-Latency Streaming Using Chunked-Encoded and Chunked Transferred CMAF", October 2018, <<https://www.akamai.com/us/en/multimedia/documents/white-paper/low-latency-streaming-cmaf-whitepaper.pdf>>.

[CODASPY17]

Reed, A. and M. Kranch, "Identifying HTTPS-Protected Netflix Videos in Real-Time", ACM CODASPY , March 2017, <<https://dl.acm.org/doi/10.1145/3029806.3029821>>.

[CoDel]

Nichols, K. and V. Jacobson, "Controlling Queue Delay", Communications of the ACM, Volume 55, Issue 7, pp. 42-50 , July 2012.

[COPA18]

Arun, V. and H. Balakrishnan, "Copa: Practical Delay-Based Congestion Control for the Internet", USENIX NSDI , April 2018, <<https://web.mit.edu/copa/>>.

- [CTA-2066] Consumer Technology Association, "Streaming Quality of Experience Events, Properties and Metrics", March 2020, <<https://shop.cta.tech/products/streaming-quality-of-experience-events-properties-and-metrics>>.
- [CTA-5004] CTA, "Common Media Client Data (CMCD)", September 2020, <<https://shop.cta.tech/products/web-application-video-ecosystem-common-media-client-data-cta-5004>>.
- [CVNI] "Cisco Visual Networking Index: Forecast and Trends, 2017-2022 White Paper", 27 February 2019, <<https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html>>.
- [ELASTIC] De Cicco, L., Caldaralo, V., Palmisano, V., and S. Mascolo, "ELASTIC: A client-side controller for dynamic adaptive streaming over HTTP (DASH)", Packet Video Workshop , December 2013, <<https://ieeexplore.ieee.org/document/6691442>>.
- [Encodings]
Apple, Inc, "HLS Authoring Specification for Apple Devices", June 2020, <https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices>.
- [I-D.cardwell-iccr-g-bbr-congestion-control]
Cardwell, N., Cheng, Y., Yeganeh, S. H., Swett, I., and V. Jacobson, "BBR Congestion Control", Work in Progress, Internet-Draft, draft-cardwell-iccr-g-bbr-congestion-control-02, 7 March 2022, <<https://datatracker.ietf.org/doc/html/draft-cardwell-iccr-g-bbr-congestion-control-02>>.
- [I-D.draft-pantos-hls-rfc8216bis]
Pantos, R., "HTTP Live Streaming 2nd Edition", Work in Progress, Internet-Draft, draft-pantos-hls-rfc8216bis-10, 8 November 2021, <<https://datatracker.ietf.org/doc/html/draft-pantos-hls-rfc8216bis-10>>.
- [I-D.ietf-httpbis-cache]
Fielding, R. T., Nottingham, M., and J. Reschke, "HTTP Caching", Work in Progress, Internet-Draft, draft-ietf-httpbis-cache-19, 12 September 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-httpbis-cache-19>>.

[I-D.ietf-quic-datagram]

Pauly, T., Kinnear, E., and D. Schinazi, "An Unreliable Datagram Extension to QUIC", Work in Progress, Internet-Draft, draft-ietf-quic-datagram-10, 4 February 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-datagram-10>>.

[I-D.ietf-quic-http]

Bishop, M., "Hypertext Transfer Protocol Version 3 (HTTP/3)", Work in Progress, Internet-Draft, draft-ietf-quic-http-34, 2 February 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-http-34>>.

[I-D.ietf-quic-manageability]

Kuehlewind, M. and B. Trammell, "Manageability of the QUIC Transport Protocol", Work in Progress, Internet-Draft, draft-ietf-quic-manageability-16, 6 April 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-manageability-16>>.

[I-D.ietf-quic-qlog-h3-events]

Marx, R., Niccolini, L., and M. Seemann, "HTTP/3 and QPACK qlog event definitions", Work in Progress, Internet-Draft, draft-ietf-quic-qlog-h3-events-01, 7 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-qlog-h3-events-01>>.

[I-D.ietf-quic-qlog-main-schema]

Marx, R., Niccolini, L., and M. Seemann, "Main logging schema for qlog", Work in Progress, Internet-Draft, draft-ietf-quic-qlog-main-schema-02, 7 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-qlog-main-schema-02>>.

[I-D.ietf-quic-qlog-quic-events]

Marx, R., Niccolini, L., and M. Seemann, "QUIC event definitions for qlog", Work in Progress, Internet-Draft, draft-ietf-quic-qlog-quic-events-01, 7 March 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-quic-qlog-quic-events-01>>.

[IAB-ADS] "IAB", n.d., <<https://www.iab.com/>>.

[IABcovid] Arkko, J., Farrel, S., Kuehlewind, M., and C. Perkins, "Report from the IAB COVID-19 Network Impacts Workshop 2020", November 2020, <<https://datatracker.ietf.org/doc/draft-iab-covid19-workshop/>>.

- [Jacobson-Karels]
Jacobson, V. and M. Karels, "Congestion Avoidance and Control", November 1988,
<<https://ee.lbl.gov/papers/congavoid.pdf>>.
- [Labovitz] Labovitz, C., "Network traffic insights in the time of COVID-19: April 9 update", April 2020,
<<https://www.nokia.com/blog/network-traffic-insights-time-covid-19-april-9-update/>>.
- [LabovitzDDoS]
Takahashi, D., "Why the game industry is still vulnerable to DDoS attacks", May 2018,
<<https://venturebeat.com/2018/05/13/why-the-game-industry-is-still-vulnerable-to-distributed-denial-of-service-attacks/>>.
- [LL-DASH] DASH-IF, "Low-latency Modes for DASH", March 2020,
<<https://dashif.org/docs/CR-Low-Latency-Live-r8.pdf>>.
- [Micro] Taher, T. M., Misurac, M. J., LoCicero, J. L., and D. R. Ucci, "Microwave Oven Signal Interference Mitigation For Wi-Fi Communication Systems", 2008 5th IEEE Consumer Communications and Networking Conference 5th IEEE, pp. 67-68 , 2008.
- [Mishra] Mishra, S. and J. Thibault, "An update on Streaming Video Alliance", April 2020,
<<https://datatracker.ietf.org/meeting/interim-2020-mops-01/materials/slides-interim-2020-mops-01-sessa-april-15-2020-mops-interim-an-update-on-streaming-video-alliance>>.
- [MMSP20] Durak, K. and et al, "Evaluating the performance of Apple's low-latency HLS", IEEE MMSP , September 2020,
<<https://ieeexplore.ieee.org/document/9287117>>.
- [MMSys11] Akhshabi, S., Begen, A. C., and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP", ACM MMSys , February 2011,
<<https://dl.acm.org/doi/10.1145/1943552.1943574>>.
- [MPEG-CMAF]
"ISO/IEC 23000-19:2020 Multimedia application format (MPEG-A) - Part 19: Common media application format (CMAF) for segmented media", March 2020,
<<https://www.iso.org/standard/79106.html>>.

- [MPEG-DASH] "ISO/IEC 23009-1:2019 Dynamic adaptive streaming over HTTP (DASH) - Part 1: Media presentation description and segment formats", December 2019, <<https://www.iso.org/standard/79329.html>>.
- [MPEG-DASH-SAND] "ISO/IEC 23009-5:2017 Dynamic adaptive streaming over HTTP (DASH) - Part 5: Server and network assisted DASH (SAND)", February 2017, <<https://www.iso.org/standard/69079.html>>.
- [MPEG-TS] "H.222.0 : Information technology - Generic coding of moving pictures and associated audio information: Systems", 29 August 2018, <<https://www.itu.int/rec/T-REC-H.222.0>>.
- [MPEGI] Boyce, J. M. and et al, "MPEG Immersive Video Coding Standard", Proceedings of the IEEE , n.d., <<https://ieeexplore.ieee.org/document/9374648>>.
- [OReilly-HPBN] "High Performance Browser Networking (Chapter 2: Building Blocks of TCP)", May 2021, <<https://hpbn.co/building-blocks-of-tcp/>>.
- [PCC] Schwarz, S. and et al, "Emerging MPEG Standards for Point Cloud Compression", IEEE Journal on Emerging and Selected Topics in Circuits and Systems , March 2019, <<https://ieeexplore.ieee.org/document/8571288>>.
- [Port443] "Service Name and Transport Protocol Port Number Registry", April 2021, <<https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.txt>>.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<https://www.rfc-editor.org/rfc/rfc793>>.
- [RFC2001] Stevens, W., "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", RFC 2001, DOI 10.17487/RFC2001, January 1997, <<https://www.rfc-editor.org/rfc/rfc2001>>.

- [RFC3135] Border, J., Kojo, M., Griner, J., Montenegro, G., and Z. Shelby, "Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations", RFC 3135, DOI 10.17487/RFC3135, June 2001, <<https://www.rfc-editor.org/rfc/rfc3135>>.
- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, DOI 10.17487/RFC3550, July 2003, <<https://www.rfc-editor.org/rfc/rfc3550>>.
- [RFC3758] Stewart, R., Ramalho, M., Xie, Q., Tuexen, M., and P. Conrad, "Stream Control Transmission Protocol (SCTP) Partial Reliability Extension", RFC 3758, DOI 10.17487/RFC3758, May 2004, <<https://www.rfc-editor.org/rfc/rfc3758>>.
- [RFC4733] Schulzrinne, H. and T. Taylor, "RTP Payload for DTMF Digits, Telephony Tones, and Telephony Signals", RFC 4733, DOI 10.17487/RFC4733, December 2006, <<https://www.rfc-editor.org/rfc/rfc4733>>.
- [RFC4960] Stewart, R., Ed., "Stream Control Transmission Protocol", RFC 4960, DOI 10.17487/RFC4960, September 2007, <<https://www.rfc-editor.org/rfc/rfc4960>>.
- [RFC5594] Peterson, J. and A. Cooper, "Report from the IETF Workshop on Peer-to-Peer (P2P) Infrastructure, May 28, 2008", RFC 5594, DOI 10.17487/RFC5594, July 2009, <<https://www.rfc-editor.org/rfc/rfc5594>>.
- [RFC5762] Perkins, C., "RTP and the Datagram Congestion Control Protocol (DCCP)", RFC 5762, DOI 10.17487/RFC5762, April 2010, <<https://www.rfc-editor.org/rfc/rfc5762>>.
- [RFC6190] Wenger, S., Wang, Y.-K., Schierl, T., and A. Eleftheriadis, "RTP Payload Format for Scalable Video Coding", RFC 6190, DOI 10.17487/RFC6190, May 2011, <<https://www.rfc-editor.org/rfc/rfc6190>>.
- [RFC6582] Henderson, T., Floyd, S., Gurtov, A., and Y. Nishida, "The NewReno Modification to TCP's Fast Recovery Algorithm", RFC 6582, DOI 10.17487/RFC6582, April 2012, <<https://www.rfc-editor.org/rfc/rfc6582>>.

- [RFC6817] Shalunov, S., Hazel, G., Iyengar, J., and M. Kuehlewind, "Low Extra Delay Background Transport (LEDBAT)", RFC 6817, DOI 10.17487/RFC6817, December 2012, <<https://www.rfc-editor.org/rfc/rfc6817>>.
- [RFC6843] Clark, A., Gross, K., and Q. Wu, "RTP Control Protocol (RTCP) Extended Report (XR) Block for Delay Metric Reporting", RFC 6843, DOI 10.17487/RFC6843, January 2013, <<https://www.rfc-editor.org/rfc/rfc6843>>.
- [RFC7258] Farrell, S. and H. Tschofenig, "Pervasive Monitoring Is an Attack", BCP 188, RFC 7258, DOI 10.17487/RFC7258, May 2014, <<https://www.rfc-editor.org/rfc/rfc7258>>.
- [RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <<https://www.rfc-editor.org/rfc/rfc7510>>.
- [RFC7656] Lennox, J., Gross, K., Nandakumar, S., Salgueiro, G., and B. Burman, Ed., "A Taxonomy of Semantics and Mechanisms for Real-Time Transport Protocol (RTP) Sources", RFC 7656, DOI 10.17487/RFC7656, November 2015, <<https://www.rfc-editor.org/rfc/rfc7656>>.
- [RFC7661] Fairhurst, G., Sathiaselalan, A., and R. Secchi, "Updating TCP to Support Rate-Limited Traffic", RFC 7661, DOI 10.17487/RFC7661, October 2015, <<https://www.rfc-editor.org/rfc/rfc7661>>.
- [RFC8083] Perkins, C. and V. Singh, "Multimedia Congestion Control: Circuit Breakers for Unicast RTP Sessions", RFC 8083, DOI 10.17487/RFC8083, March 2017, <<https://www.rfc-editor.org/rfc/rfc8083>>.
- [RFC8084] Fairhurst, G., "Network Transport Circuit Breakers", BCP 208, RFC 8084, DOI 10.17487/RFC8084, March 2017, <<https://www.rfc-editor.org/rfc/rfc8084>>.
- [RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <<https://www.rfc-editor.org/rfc/rfc8085>>.
- [RFC8216] Pantos, R., Ed. and W. May, "HTTP Live Streaming", RFC 8216, DOI 10.17487/RFC8216, August 2017, <<https://www.rfc-editor.org/rfc/rfc8216>>.

- [RFC8312] Rhee, I., Xu, L., Ha, S., Zimmermann, A., Eggert, L., and R. Scheffenegger, "CUBIC for Fast Long-Distance Networks", RFC 8312, DOI 10.17487/RFC8312, February 2018, <<https://www.rfc-editor.org/rfc/rfc8312>>.
- [RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <<https://www.rfc-editor.org/rfc/rfc8446>>.
- [RFC8622] Bless, R., "A Lower-Effort Per-Hop Behavior (LE PHB) for Differentiated Services", RFC 8622, DOI 10.17487/RFC8622, June 2019, <<https://www.rfc-editor.org/rfc/rfc8622>>.
- [RFC8723] Jennings, C., Jones, P., Barnes, R., and A.B. Roach, "Double Encryption Procedures for the Secure Real-Time Transport Protocol (SRTP)", RFC 8723, DOI 10.17487/RFC8723, April 2020, <<https://www.rfc-editor.org/rfc/rfc8723>>.
- [RFC8824] Minaburo, A., Toutain, L., and R. Andreasen, "Static Context Header Compression (SCHC) for the Constrained Application Protocol (CoAP)", RFC 8824, DOI 10.17487/RFC8824, June 2021, <<https://www.rfc-editor.org/rfc/rfc8824>>.
- [RFC8825] Alvestrand, H., "Overview: Real-Time Protocols for Browser-Based Applications", RFC 8825, DOI 10.17487/RFC8825, January 2021, <<https://www.rfc-editor.org/rfc/rfc8825>>.
- [RFC8999] Thomson, M., "Version-Independent Properties of QUIC", RFC 8999, DOI 10.17487/RFC8999, May 2021, <<https://www.rfc-editor.org/rfc/rfc8999>>.
- [RFC9000] Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/rfc/rfc9000>>.
- [RFC9001] Thomson, M., Ed. and S. Turner, Ed., "Using TLS to Secure QUIC", RFC 9001, DOI 10.17487/RFC9001, May 2021, <<https://www.rfc-editor.org/rfc/rfc9001>>.
- [RFC9002] Iyengar, J., Ed. and I. Swett, Ed., "QUIC Loss Detection and Congestion Control", RFC 9002, DOI 10.17487/RFC9002, May 2021, <<https://www.rfc-editor.org/rfc/rfc9002>>.

- [RFC9065] Fairhurst, G. and C. Perkins, "Considerations around Transport Header Confidentiality, Network Operations, and the Evolution of Internet Transport Protocols", RFC 9065, DOI 10.17487/RFC9065, July 2021, <<https://www.rfc-editor.org/rfc/rfc9065>>.
- [SFRAME] "Secure Media Frames Working Group (Home Page)", n.d., <<https://datatracker.ietf.org/doc/charter-ietf-sframe/>>.
- [SRT] Sharabayko, M., "Secure Reliable Transport (SRT) Protocol Overview", 15 April 2020, <<https://datatracker.ietf.org/meeting/interim-2020-mops-01/materials/slides-interim-2020-mops-01-sessa-april-15-2020-mops-interim-an-update-on-streaming-video-alliance>>.
- [Survey360o] Yaqoob, A., Bi, T., and G. Muntean, "A Survey on Adaptive 360° Video Streaming: Solutions, Challenges and Opportunities", IEEE Communications Surveys & Tutorials, July 2020, <<https://ieeexplore.ieee.org/document/9133103>>.

Authors' Addresses

Jake Holland
Akamai Technologies, Inc.
150 Broadway
Cambridge, MA 02144,
United States of America
Email: jakeholland.net@gmail.com

Ali Begen
Networked Media
Turkey
Email: ali.begen@networked.media

Spencer Dawkins
Tencent America LLC
United States of America
Email: spencerdawkins.ietf@gmail.com