Evaluating 5G uplink performance in low latency video streaming

Mikko Uitto and Antti Heikkinen
VTT Technical Research Centre of Finland Ltd
Kaitoväylä 1, 90570 Oulu, Finland
Email: firstname.surname@vtt.fi

***Abstract*—In this paper, we evaluate the efficiency of low latency video streaming in a real standalone 5G test environment with a particular focus on uplink latency using UDP and TCP transport protocols. Furthermore, the evaluation comprise also congested scenario in which additional traffic is brought to the uplink reducing the network capacity, which is a common phenomenon in live mobile streaming from the field. Although 5G and beyond networking has brought higher capacity and reduced latency, the uplink is usually less prioritized in terms of network share. This limitation can generate a bottleneck in live video streaming use cases requiring low latency, such as delivering real time environmental data between vehicles in V2X scenarios or performing remote operations from a distance. The results gathered from the extensive set of evaluation with test cases indicate that 5G standalone has improved potential for low latency streaming and delay variation stays relatively satisfying level even in congested network scenarios.**

***Keywords*—Video streaming; Low latency; RTSP; 5G SA**

I. Introduction

The fifth generation (5G) mobile networking has started actualizing their environment from Non-Standalone (NSA) towards Standalone (SA) architecture, which has been promised to provide improved performance, energy efficiency, and support for dedicated network slices [1]. The commercial operators are still mainly using NSA, but the SA tests and partial deployment have already started to take place. Many of the latest user equipment (UE) devices have the support for SA, but connectivity and/or performance issues still occurs especially when working with research oriented test networks, which may not use the existing mobile network operator codes in conjunction with mobile country codes. In this paper, we are using one of the research networks for performing our evaluative low latency video streaming measurements with SA connectivity, and deploy 5G test network (5GTN) [2].

The improved reliability and performance of mobile networks has started the shift from wired towards wireless infrastructure in various industry sectors following the path by Industry 4.0 [3]. These include not only the use cases specific for the smart manufacturing and warehousing or vehicle-to-everything (V2X), but 5G brings several advantages for less dedicated environments, such as in mines or harbour areas. These environments tend to form closed surroundings benefiting of private mobile network, which is utilised by the area services. As these services are already evolving towards unmanned vehicles with remote operations, video streaming with extreme low latency play the key role of these scenarios. The variation of delay (aka jitter) is sometimes even more important and characteristics for low latency applications.

Uplink (UL) congestion and limited capacity in public 5G networks is another challenge for remote operations in mobile streaming scenarios. Commercial networks usually comply with the predetermined country policies where uplink/downlink (DL) share is fixed with 1/4 or 3/7 ratio, which means DL holds practically higher throughput capacity. These policies can lead easily to congested situation in the UL channel if multiple users (e.g. with cameras) are consuming the resources simultaneously. The earlier conducted studies using NSA and transmission control protocol (TCP) indicate that this will also lead to higher UL delay even at lower traffic rates [4]. Industrial private networks usually require much higher throughput to UL than DL, because the area services are usually likely producing data rather than consuming. In this work, we will take the next step with SA and evaluate its usefulness for low latency communication especially in terms of latency and jitter in conjunction with the UL channel usage.

HTTP streaming has been dominating the non-latency critical video transmission in several service platforms, but it is more based on serving the high number of DL users with high quality and enabling adaptation against network fluctuations. In the streaming scenarios requiring low latency with high performance, user datagram protocol (UDP) instead of TCP can lead to decreased latency, especially in the UL. Traditionally UDP is used in live streaming due to its less complex and lightweight nature with less IP overhead. Thus, use of UDP can usually demand preliminary work in the setup including i.e. processing of network address translation (NAT) with firewall rules as well as opening the ports in the receiver. On the other hand, TCP guarantees the transmission with high reliability, but has a slightly higher cost in terms of latency.

In this paper, we concentrate on performing deep evaluation using real-time streaming protocol (RTSP) over UDP and TCP for delivering live video stream over 5G SA with and without UL congestion in a real network environment. 1/4 and 3/7 UL/DL share configurations are evaluated. The emphasis will be on analyzing the network delays and jitter, as well as end to end (E2E) latency. The term 'delay' is defined as one way latency between network nodes, and term latency concerns the whole transmission chain starting from the camera capture ending to display in the playback device.

This paper is organised in the following way. First in Section II the system architecture is represented from the video streaming as well as from the connectivity point of view.

Section III illustrates the evaluation setup followed by the results in Section IV. Finally, Section V concludes the paper.

II. System architecture

A. Low latency video streaming architecture

The low latency video streaming system captures, encodes, and transmits video from one node to another over a network. The client-server video streaming architecture may have two or three nodes shown in Fig. 1. In the system with two nodes Fig. 1a) a video source, encoder, and a streaming server locate in the transmitter side and a video player in the receiver side. In the three nodes system Fig. 1b) the video source and the encoder in transmitter side push the encoded video over the network to the video server and video server streams the video over the network to the video player. There are several components in the low latency video streaming chain and each component adds a certain amount of delay to the chain and the sum of the delays forms the E2E latency [5].

Live video source is usually a camera which captures the real environment or graphic engine that creates synthetic video or augmented reality overlay to the live video. Video is captured or created in certain framerate and raw format such as YUV. The delay caused by the camera consists of temporal sampling and processing. If sampling rate of the camera is 30 Hz (30 fps), it means the frame period is 33 ms. This means if event occurs just before sampling the effect of sampling rate on the delay is quite small. On the other hand, if event occurs just after the sampling the delay is almost the entire frame period (33 ms). A large part of the processing delay may originate from transmission of video frames from the camera to the encoder. For instance, if the camera is connected to the computer via USB 3.0 (5 Gbit/s) the transmission delay for raw YUV 4:2:0 1920x1080 frame is about 5 ms.

Video encoder compress the raw video from video source by removing redundant information to format that conforms to a standard such as H.264, HEVC, which target on providing audiovisual services for human consumption. The latest standards in video coding have also enabled so called 2-D video for machines, which targets on utilising the video with lower bitrates in machine analysis and machine vision requiring even lower latency [6]. The low latency streaming encoders use rate-control to produce constant bitrate. The video encoders have options for low latency video streaming, which e.g. sets the buffer size as small as possible and disables b-frames. The low latency options reduces the latency, but the video quality may be lower especially in case of network fluctuations. In addition, video encoders can use graphics processing unit (GPU) acceleration to reduce latency particularly when higher resolutions and bitrates are required to be processed in real-time.

The streaming server receives the video from the encoder and delivers it to the clients over a network. The streaming server may repack or encapsulate the encoded video in a suitable format for transmission over the network. The streaming server establishes and possibly controls streaming sessions between the server and the client. For the low latency streaming video servers are based on RTSP, real-time messaging protocol (RTMP), or web real-time communication (WebRTC) protocols. The RTSP server can use either UDP or TCP protocols to receive the live video feed from the encoder and serve it to the video player.

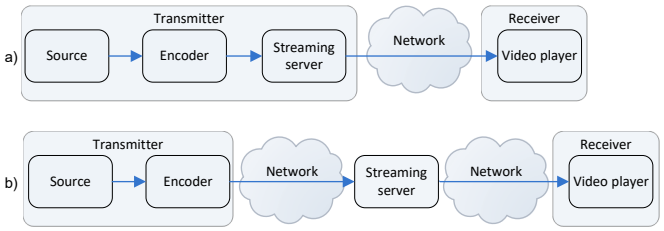

Fig. 1: Low latency video streaming architectures.

The video player receives the video from the streaming server and feeds it to the decoder. The buffer in the player handles the differences between the video bitrate and the network transmission rate. If video bitrate is constant and network throughput does not vary, the buffer may be smaller whereby the latency is smaller. For low latency video streaming, the video player's buffer should be as small as possible. The video player, the streaming server, or the encoder can perform rate adaptation to fit the video to the network throughput.

B. Low latency video streaming in 5G network

Low latency video streaming from vehicles or places without a fixed network require a high-speed, mobile, and reliable network infrastructure. Public and private 5G networks provide a viable option for video transmission. Especially uplink performance and edge computing are key factors in low latency video streaming. 5G networks can be configured to better support uplink transmission and 5G SA architecture provides more optimized network for different use cases and even improves the performance compared to NSA. In addition, bringing services closer to user such as deploying video server to the near edge of the network, can reduce latency [7].

5G has evolved and operators has started shifting from NSA to SA [1]. In the 5G NSA network the UE is connected to LTE and 5G network at the same time. The control plane that UE uses is anchored to LTE and utilizes the 4G core network (EPC). The UE uses 5G network as a data plane. In the 5G SA network UE is directly connected to the 5G core (5GC) network and does not need EPC, which can also decrease the latency in the core network.

In time divisional dublex (TDD), one block of spectrum is time divided between the uplink and downlink, which provides some flexibility in the allocation of resources between the UL and DL. However in LTE or 5G NSA TDD networks, operating in the same area with the same frequency, the bands should be synchronised. Base stations aka small cells should be transmitting at one time periods and UEs should be transmitting at another time periods. Hence it is recommended to use a common frame structure to avoid cross-link interference (CLI). The frame structure define UL/DL shares and in the public networks it is e.g. 1/4 where uplink performance is not optimal for use cases that require high throughput and low latency. In our previous work we have been evaluated live video streaming performance in 5G NSA network [4]. The uplink delay in the 5G NSA network with 3/7 frame structure was about 12 ms.

5G-Advanced as defined in 3GPP Release 18 should bring new solutions for increasing TDD UL performance. The CLI mitigation techniques allow to use frame structures more freely and enable so-called uplink-heavy UL/DL configurations [8]. Probably first 5G networks that support uplink-heavy use cases would be private 5G SA networks empowered by multi access edge computing.

III. Evaluation

This section presents the test environment, different test cases as well as testing tools. The emphasis for the evaluative measurements in this paper was to keep 5G SA in a fixed configuration during the evaluation by setting the desired UL/DL share 1/4 or 3/7.

A. Test environment

5GTN was used during the live evaluation. This research infra provides the possibility to test the latest available mobile HW and SW against different configurations and use cases. 5GTN comprises currently both 5G NSA and SA architectures as well as mmWave for testing the higher frequencies in URLLC scenarios. The research environment comprises also the novel UEs, which are acquired flexibly according to the feature support and need for different experiments. The point of our interest, 5G SA, operating at band 78 @ 3.5 GHz, can be paired with several different core instances, of which Open5GS [9] was used in the experiments for this paper. Indoor SA cell was applied in the evaluation with 1/4 and 3/7 UL/DL share in the frame configuration. The video equipment as well as 5G modems was placed approximately 5 m distance from the SA cell, line of sight, in order to have good signal coverage and strength. Identical 5G modems (Telewell) were deployed both for the UL and DL, as well as for the traffic generator.

B. Test setup

The test setup for evaluating low latency video streaming is presented in Fig. 2. In the test setup all the computers run the Ubuntu 20.04 operating system. At the transmitter side we have Intel NUC Core i7 mini PC, which contains a Video Encoder, an iPerf3 client, and a Qosium Probe. A webcam (Logitech BRIO 4K Ultra HD Pro) is connected to the mini PC with a USB cable. The Video Encoder is modified to change the video target bitrate at runtime without interrupting the encoding. The encoder is based on the FFmpeg version 4.2.4. It captures 4:2:0 YUV video frames from camera at 1920x1080 resolution, 30 frames/s, and encodes the video using x264 encoder with the low latency settings. The mini PC is connected to the 5G network using 5G modem. The Encoder pushes the video to uplink to the Video Server. The Video Server is a RTSP server (rtsp-simple-server), which receives live video from the Encoder, publish and serve it to the end users. The Encoder can push the video to the RTSP server either using UDP or TCP protocol. At the receiver side the Video Player is running on a Core i9 laptop connected to the 5G network with 5G modem or wired network using ethernet cable depending on the test case. The MPV video player application is used to play the video from the RTSP server using either UDP or TCP protocol. MPV is also used with low latency settings.

C. Test tools

Two types of testing methods and tools were used. First, Qosium [10] SW tool for measuring the network performance was deployed in the essential network nodes. The Qosium Probes were installed in the Encoder, Video Server, and Video Player according to the Fig. 2 in order to measure UL and DL network delays. DL delays were not illustrated as results for this paper. Precision time protocol (PTP) was used in each of the network nodes by syncing the clocks with the master, which was a server located in 5GTN. The clock synchronization was done from another network interface (1 GB LAN connection), but the measured data used the 5G modem interface (USB 3.0). With PTP, accurate time synchronization for achieving reliable measurements was possible.

Second, we measured E2E latency or glass-to-glass latency which is the time it takes for an image that camera captures from a screen 1 encoded, transmitted and played to a screen 2. On the screen 1 we have a running time in ms which the camera captures. We used a Linux terminal command: (*while true; do echo -ne "'date +%H:%M:%S:%N'"; done*) to show the time in the screen 1. The video player plays the streamed video in the screen 2 that is placed next to the screen 1. We recorded a slow-motion video at 240 fps of the two screens. The E2E latency is calculated from the recorded video using the frame by frame forwarding feature in the MPV player. From the recorded video we calculated how many frames it takes until the same time value was identified on the screen 1 is seen on the screen 2. For instance, if it takes 40 frames when the screen 1 time is displayed on the screen 2 the end-to-end latency is 1/240 x 40 = 167 ms. The precision of the measurement method is 4.2 ms.

Finally, iPerf3 [11] was used for generating extra traffic into the mobile network and to model congestion in the UL. In this paper, we did not focus on evaluating the DL capacity and performance, since our use cases and scenarios require more functionality on the UL, from the source to the edge.

IV. Results

This Section represents the results of our test cases, depicted in Table I. We illustrate our findings in terms of E2E latency, as well as according to the network measurements with Qosium. All the tests were repeated five times after which averages were calculated. Three main KPIs are illustrated: delay, jitter, and throughput. The jitter value is the absolute value, defined as the absolute value of difference of delays between sequential packets. Fig. 3 and 4 focus on E2E latency, while as the other figures concentrate on the network KPIs.

We measured the E2E latency in the test cases 1 and 2. In the low latency video streaming setup, the E2E latency is the total delay caused by camera, encoder, video transmission to the UL to the video server, processing at the video server, video streaming to DL to the video player, decoding and video playback on screen 2. In test case 1 we performed the test five times and each time we recorded 30 s a slow-motion video. In the slow-motion video we calculated E2E latency in three points at approximately 5 s, 15 s, and 25 s. In test case 2 we performed the test five times and each time we recorded a slow-motion video of the test. In the slow-motion video we calculated E2E latency in two points of each bitrate.

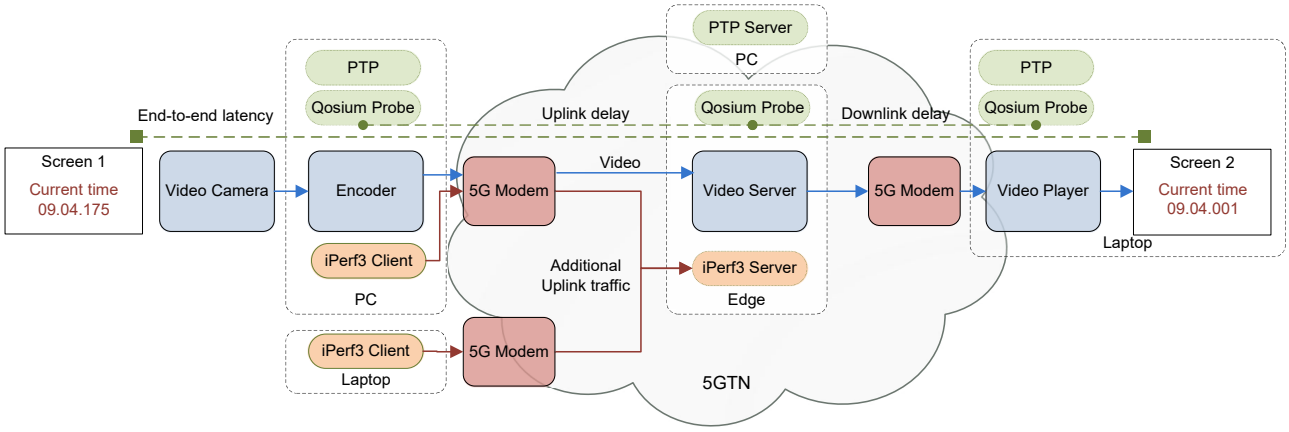

Fig. 2: Evaluation setup for low latency video streaming.

We measured the E2E latency for low latency RTSP video streaming in the test case 1. We present the results comparing the E2E latency of between LAN and 5G network (1/4 share) when using either UDP or TCP protocols. The average values for the E2E latency are presented in Fig. 3. We can see from the result that 5G network increases the E2E latency 10-16 ms taking the margin of error into account (4.2ms).

In test case 2 we wanted to measure how the bitrate of the video affects the E2E latency. The uplink from the encoder to the server uses a 5G connection and downlink from the video server to the player uses LAN. The average values for the E2E latency at different bitrates are presented in Fig. 4. We can see from the results that E2E latency increases gradually as the bitrate increases. Especially in 5G 1/4 share a larger increase is seen when the bitrate is 40 Mbit/s which is quite close to network UL maximum throughput. There is no significant difference in E2E latency between the UDP and TCP protocols.

Fig. 5 and Fig. 6 represents the 5G network delays (UL and DL) and jitters when using 4 Mbit/s video stream, which is sufficient rate for achieving good quality using H.264 encoding. The measured curves are connected to test cases 1.3-1.6 according to the Table I. The corresponding LAN curves

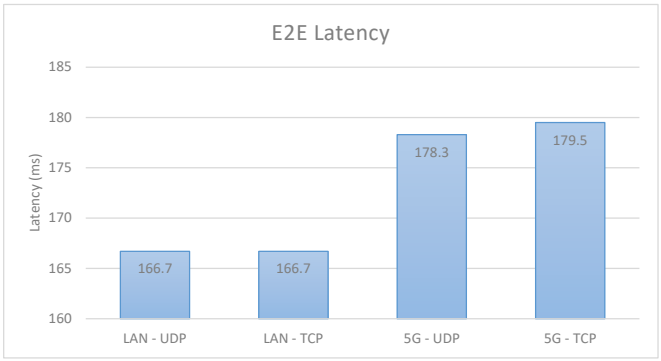

Fig. 3: Measured average E2E latency for the RTSP video stream.

are left outside the scope of this paper, because 5G findings are more substantial. According to the figures, UDP outperforms slightly better in terms of delay and jitter in the UL, but the difference is only 2 ms in favour of UDP, on average. DL

TABLE I: Test cases.

Test case	Description	UL	DL	RTSP over	5G UL/DL share	Evaluation
1	**Baseline RTSP video streaming**					
1.1	1920x1080, 30 fps, 4 Mbit/s	LAN	LAN	TCP	-	E2E
1.2	1920x1080, 30 fps, 4 Mbit/s	LAN	LAN	UDP	-	E2E
1.3	1920x1080, 30 fps, 4 Mbit/s	5G	5G	TCP	1/4	E2E, UL
1.4	1920x1080, 30 fps, 4 Mbit/s	5G	5G	UDP	1/4	E2E, UL
1.5	1920x1080, 30 fps, 4 Mbit/s	5G	5G	TCP	3/7	UL
1.6	1920x1080, 30 fps, 4 Mbit/s	5G	5G	UDP	3/7	UL
2	**RTSP video at different bitrates (Mbit/s)**					
2.1	1, 2, 4, 6, 10, 15, 20, 25, 30, 35, 40	5G	LAN	TCP	1/4	E2E, UL
2.2	1, 2, 4, 6, 10, 15, 20, 25, 30, 35, 40	5G	LAN	UDP	1/4	E2E, UL
2.3	1, 2, 4, 6, 10, 15, 20, 25, 30, 35, 40	5G	LAN	TCP	3/7	E2E, UL
2.4	1, 2, 4, 6, 10, 15, 20, 25, 30, 35, 40	5G	LAN	UDP	3/7	E2E, UL
3	**Traffic UE1: RTSP video video UE2: iPerf3 traffic**					
3.1	4 Mbit/s video and TCP traffic	5G	LAN	TCP	1/4	UL
3.2	4 Mbit/s video and UDP traffic	5G	LAN	TCP	1/4	UL
3.3	4 Mbit/s video and TCP traffic	5G	LAN	UDP	1/4	UL
3.4	4 Mbit/s video and UDP traffic	5G	LAN	UDP	1/4	UL

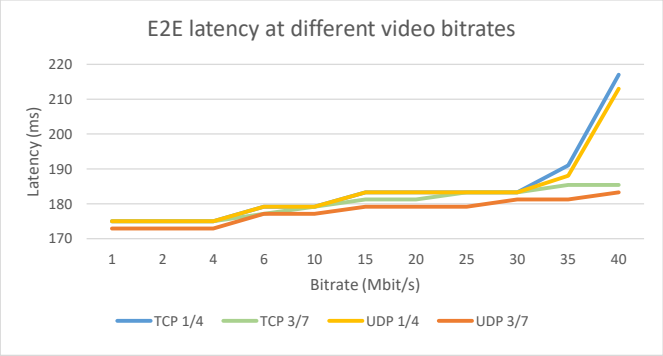

Fig. 4: Measured average E2E latency for the RTSP video stream at different video bitrates.

performs equally well both for UDP and TCP. The difference in jitter between the protocols is even smaller, slightly higher in UL than in DL. 3/7 frame configuration leads to decreased latency mostly due to higher UL capacity.

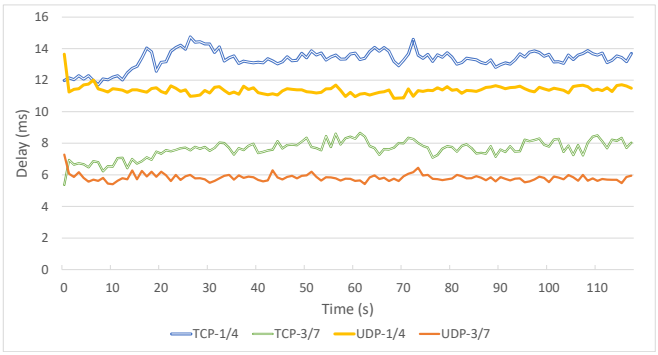

Fig. 5: Measured average UL delays for the 4 Mbit/s RTSP video stream.

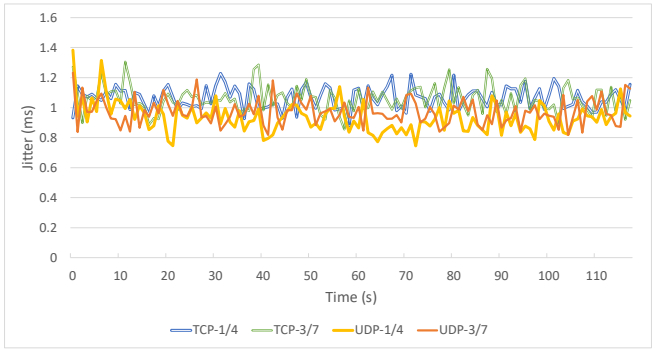

Fig. 6: Measured average UL jitters for the 4 Mbit/s RTSP video stream.

Fig. 8 and Fig. 9 illustrate the effect of increasing video bitrate according to the UL throughput measured with Qosium in Fig. 7. These results comprise the evaluation from the test cases 2.1-2.4 according to the Table I. The throughput is similar both for UDP and TCP, as it should be. No packet losses occurred yet at these bitrates, which was a conscious choice according to the UL capacity. We see clearly the effect of dynamic encoding aka bitrate increase from Fig. 7, where 11 steps of bitrate increase was applied.

The measured UL delay for 1/4 configuration (Fig. 8) stays below 20 ms until 20 Mbit/s bitrate is reached, which basically means half of the UL channel utilisation capacity. Beyond that, the UL delay stays relatively low until 40 Mbit/s is reached, which is near the the total UL limit (approximately 45 Mbit/s). Finally, the TCP delay is higher near the UL capacity limitation. On the other hand, UDP jitter (Fig. 9) is significantly higher and increases almost linearly with respect to the delay and/or bitrate, which can cause problems in remote controlled scenarios dependent of only small delay variations.

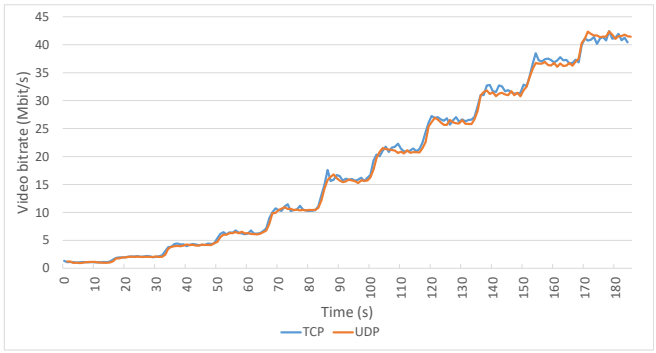

Fig. 7: Measured average bitrate when using different bitrates.

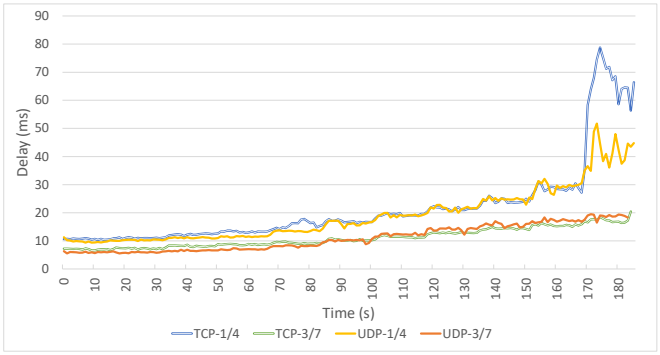

Fig. 8: Measured average UL delays when using different bitrates.

Fig. 10 and Fig. 11 illustrate the test cases 3.1-3.4 where extra traffic was congested into UL from separate UE and access point. In these test cases we wanted to see how another user influences to the video streaming session, how the scheduling in the 5G small cell works, and whether if congestion protocol (UDP or TCP) has an effect to the results. The 4 Mbit/s video stream was initiated before the congestion was generated with timed iPerf3 script.

The results indicate that congestion affects slightly both to UL delay as well as jitter. TCP based video has higher

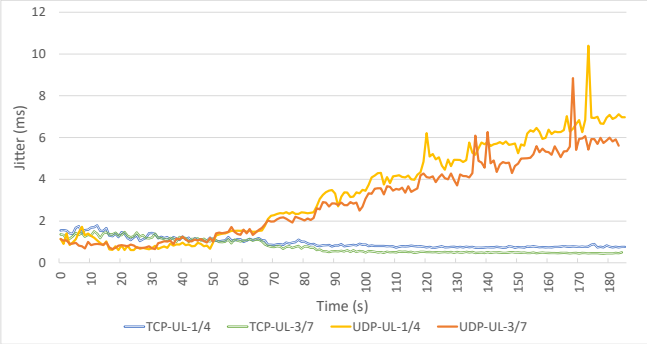

Fig. 9: Measured average UL jitters when using different bitrates.

delay approximately 2 ms regardless of the congestion type. The overall increment to delay due to congestion is 2 ms for TCP and 1 ms for UDP. This is basically explained with the scheduling algorithm in the small cell, which favours the video stream which was initiated before congestion. In the jitters' graph, the values increase similarly like delays according to the amount of congestion. The increase in jitters raise from 0.9 ms to 1.2 ms, which is considered low.

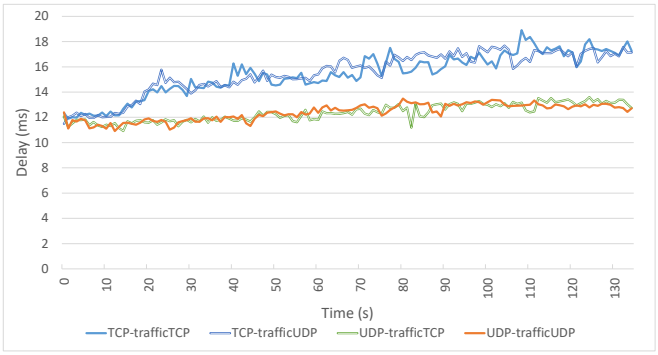

Fig. 10: Measured average UL delays when creating extra congestion with iPerf3.

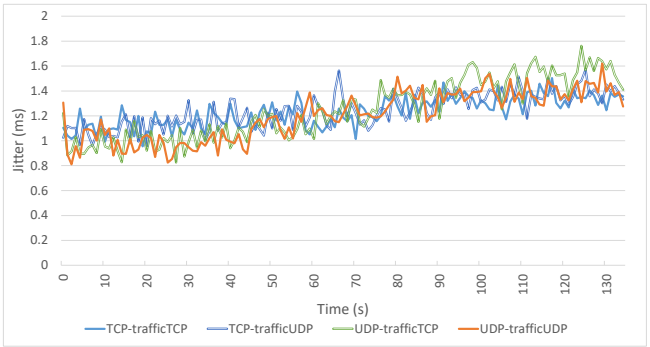

Fig. 11: Measured average UL jitters when creating extra congestion with iPerf3.

V. Conclusion

This paper presented the comparative evaluation measurements using 5G SA with UDP and TCP based video streaming targeting for low latency. The emphasis was on evaluating network delay on UL direction, and E2E latency in UL critical streaming scenarios, such as in remote operations for unmanned vehicles. In such cases delay variation aka jitter can play a critical role. The results indicate that use of UDP and UL/DL configuration has significant impact on latency when designing private networks. On the other hand according to the measurements, its usage in heavily congested UL can lead to higher latency penalty and eventually to packet losses, which needs to be taken into account when designing quality-dependent streaming solutions. The next steps for the authors are to investigate and focus on 5G SA parameters, which can not only improve the UL performance in terms of latency and capacity, but also to form private networks in dedicated mobile environments.

Acknowledgment

This work was carried out and supported in by the European Commission in the framework of the H2020-ICT-19-2019 project 5G-HEART (Grant agreement no. 857034), and DEDICAT 6G project funded under the European Union H2020 research and innovation programme (Grant Agreement No. 101016499). The contents of this publication are the sole responsibility of the authors and do not in any way reflect the views of the European Union. The authors would like to thank for the support.

References

[1] G. Liu, Y. Huang, Z. Chen, L. Liu, Q. Wang, and N. Li, "5g deployment: Standalone vs. non-standalone from the operator perspective," *IEEE Communications Magazine*, vol. 58, no. 11, pp. 83–89, 2020.

[2] 5GTN. (2019) 5GTN - 5G test network. [Online]. Available: https://5gtn.fi/

[3] S. K. Rao and R. Prasad, "Impact of 5g technologies on industry 4.0," *Wireless Personal Communications*, vol. 100, pp. 145–159, 2018.

[4] M. Uitto and A. Heikkinen, "Evaluation of live video streaming performance for low latency use cases in 5g," in *2021 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, 2021, pp. 431–436.

[5] C. Bachhuber, E. Steinbach, M. Freundl, and M. Reisslein, "On the minimization of glass-to-glass and glass-to-algorithm delay in video communication," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 238–252, 2018.

[6] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.

[7] N. Makris, V. Passas, C. Nanis, and T. Korakis, "On minimizing service access latency: Employing mec on the fronthaul of heterogeneous 5g architectures," in *2019 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, 2019, pp. 1–6.

[8] K. Pedersen, D. Chandramouli, M. Baker, A. Toskala, S. Nielsen, and J. Moilanen, "5g-advanced: Expanding 5g for the connected world. white paper," Nokia, Tech. Rep., 2022.

[9] S. Lee. (2022) Open5gs - open source project of 5gc and epc (release-16). [Online]. Available: https://open5gs.org/

[10] Kaitotek. (2020) Qosium. [Online]. Available: https://www.kaitotek.com/qosium

[11] iPerf3. (2021) iperf3. [Online]. Available: https://iperf.fr/