QoE performance evaluation of YouTube video streaming in mobile broadband networks

Savvas Argyropoulos
StreamOwl
Athens, Greece
Email: savvas@streamowl.com

Nikos Fotiou
Dept. of Informatics
Athens University of Economics and Business
Email: fotiou@aueb.gr

George C. Polyzos
Dept. of Informatics
Athens University of Economics and Business
Email: polyzos@aueb.gr

***Abstract*—In this paper, a performance evaluation of the Quality-of-Experience (QoE) of YouTube video streaming in mobile broadband networks with active measurements is described. The measurements were collected from a field experiment campaign using the MONROE platform which provides probes in four European countries and enables the benchmarking of three mobile broadband operators. Firstly, we present a framework for the automated collection and processing of the measurements, and then, we analyze the results to identify the cache allocation policy per operator. Additionally, we examine whether the selected cache server has an effect on the delivered video quality and present the results using standardised objective methods for the estimation of the perceived quality.**

I. Introduction

The recent advances in mobile devices and video streaming services have motivated large-scale media consumption in wireless and mobile environments. In this context, the main objective of network operators and Internet Service Providers (*ISPs*) is to improve the Quality-of-Experience (*QoE*) of end-users by providing high-quality services and interactive mechanisms for seamless adaptation to the network conditions of each user. At the same time, the recent advances in mobile broadband (MBB) networks have diminished delays and increased bandwidth to the extent that it is possible to shift to unicast streaming technologies instead of multicast or broadcast. During the last few years, streaming media is predominately provided over-the-top (*OTT*) of existing network infrastructures and several proprietary industry solutions have been deployed, e.g., Apple HTTP Live Streaming (*HLS*), Adobe HTTP Dynamic Streaming (*HDS*), and Microsoft Smooth Streaming. Also, MPEG issued the open international standard on Dynamic Adaptive Streaming over HTTP (MPEG-DASH) and it has been employed by the major Video-on-Demand (VoD) media providers, such as YouTube and Netflix.

The ability to deliver content to individual users has created the need to monitor the performance of video streaming services to each individual client. Recognising the importance of measuring the QoE in adaptive streaming services, the International Telecommunication Union (ITU) issued the Recommendation P.1203 “Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport” [1] in 2016 and a new work item for the inclusion of HEVC-encoded and Ultra-HD video streaming is currently under investigation that is expected to be finalized in 2018[1].

The video delivery chain for typical video streaming services is depicted in Figure 1: the video resides in a video server of the OTT operator and when it is requested by the mobile device: a) it goes through the ISP backbone network directly or b) it transits through a peering connection (with an ISP which the ISP has a service agreement), or c) it is fetched from a cached copy from a Content Delivery Network (CDN). Then, the video is transmitted to the mobile client from (typically the closest) cell of the mobile network. Thus, there can be several bottlenecks in video delivery.

From the perspective of an ISP, the QoE of the video download service is mainly driven by the quality of the Internet connection between the video client and the OTT operator server. A poor connection between the user’s device and the content server can lead to a reduced video download performance which can cause long video start-up times, video freezes, and unwanted video terminations. In this context, the video client plays a major role as it handles the video buffer size and controls the video download; in adaptive bitrate (ABR) video streaming, especially, the video content is encoded at several quality levels (representations) and the video client is responsible for the adaptation strategy and the fetching of the representation that matches the network conditions of the client to avoid buffer underruns.

YouTube issues the “Video Quality Report” which ranks the ISPs based on their ability to stream YouTube video at HD quality for at least 95% of the day, while Netflix issues the “ISP Speed Index”[2] to rank the ISPs based on the average throughput of Netflix video delivery in prime time. Therefore, it is evident that the evaluation of video streaming services is essential in order to identify the bottlenecks in the network and provide insights for the design the network for efficient video delivery.

In this paper, we focus on the performance of YouTube since it is one of the most popular video streaming services and serves more than 1 billion users per month. Indeed, mobile YouTube traffic accounts for more than 30% of the total

[1] http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=14039
[2] https://ispspeedindex.netflix.com/

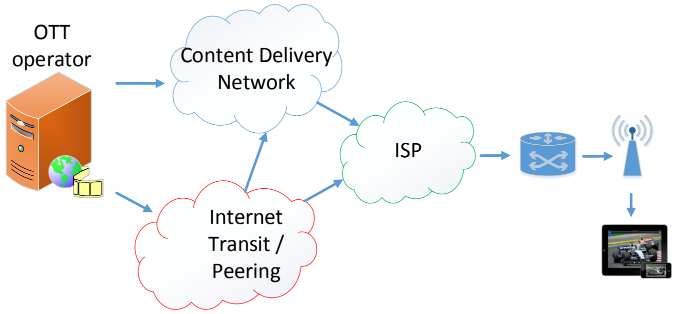

Fig. 1. Video delivery chain in mobile broadband networks.

Internet traffic, and more than 40% of the total YouTube traffic. We present a framework for active YouTube measurements and present the results that have been collected from probes in Norway, Sweden, Spain, and Italy with a variety of ISPs in each country using the MONROE platform [2]. Both Key Performance Indicators (KPIs) at the network layer, e.g. loss rate, and Key Quality Indicators (KQIs), e.g. start-up delay, freezing time, and quality adaptation, are presented for the analysis of the results and the variations in performance during the day is analysed.

II. Related work

The traffic analysis of YouTube video has been in the focus of several works in the past because of the popularity of the service and the constant evolution in the delivery infrastructure by Google. In [3], a characterisation of YouTube traffic in fixed and mobile environments was presented. The study analysed the hosting infrastructure and showed that that the usage of caching in mobile networks provides high benefits in terms of delay as well as downlink throughput. Similar studies of YouTube traffic characterisation were conducted in [4] but the study was confined to a campus network. A specific study of YouTube in cellular networks was presented in [5]; the study corroborates that local content caching by servers at the edge of the ISP improves the streaming performance.

Moreover, the streaming performance, in terms of startup delay and bitrate ratio between download rate and video encoding bitrate was evaluated in [6] and it was shown that the performance is heterogeneous across the different collected data (geographical locations and mobile/PC device), with mobile devices suffering larger delays and that YouTube infrastructure is performing less efficiently when serving requests by mobile clients. The impact of the YouTube infrastructure and the distributed CDN system is the subject of several research studies. In [7], a methodology to unveil CDN changes through passive measurements was presented and it was shown that sudden changes may occur. In [8], YouTube traffic flows were analysed for over a month and showed that the CDN server selection policy had a negative impact on QoE.

From the above studies, it becomes evident that YouTube streaming performance may be different in cellular networks than in fixed networks. In [9], the packet loss characteristics of MBB networks under mobility was investigated, and the impact of Radio Access Technology (RAT) changes, cell handovers and Location Area Codes (LAC) changes was examined and showed that almost 70% of the sessions with a RAT change resulted in packet loss. Additionally, in [10], measurements were conducted on a 4G cellular network in Germany and it was shown that the management and configuration decisions have a substantial impact on the performance since the association of mobile devices to a Point of Presence (PoP) within the operator' s network can influence the end-to-end performance by a large extent. Furthermore, a profiling of MBB coverage along the railway infrastructure in Norway was performed in [11]; it was shown that two main coverage profiles emerge, one where 3G dominates and one where "no-service" dominates. A unique platform for independent, repeatable, multi-homed, large-scale measurements and experiments in operational MBB networks was developed [2],[12]. The platform supports three cellular network connections to enable parallel measurements from different operators. This platform is used for the collection of the measurements in the present study, as described in Section III below.

The relation between low-level network characteristics and video QoE, i.e. the quality as it is perceived by the end-users and reflects their satisfaction, has been studied in the past. More specifically, in [13], the YouTube Performance Monitoring Application (YomoApp) was presented, an Android application which passively monitors KQIs (e.g. player state/events, video quality levels, stalling, etc.) of YouTube adaptive video streaming on end-user smartphones. Another similar tool for YouTube QoE evaluation in Android wireless terminals was presented in [14]. The importance of examining the impact of network-level parameters on human-perceived QoE was highlighted in [15]. This study presented a YouTube flow control mechanism and a model for the video quality as perceived by the viewers. Finally, a measurement campaing in the field was presented in [16], and it was shown that monitoring the network parameters alone is not sufficient to infer the QoE; instead, application-layer parameters need to be collected which show higher correlation with subjective opinions.

This paper presents a framework for automated testing of YouTube video streaming performance and presents the results of a field study in cellular networks in four countries using the MONROE platform [12]. The main contributions with regard to previous studies are: (a) the measurements are conducted with 3 mobile broadband operators in parallel, which enables the evaluation of the service based on the network infrastructure, (b) the measurements are collected over a period of one month using real network traffic and no artificial or emulated environment, (c) network- and application-layer parameters are collected and analysed in terms of QoE, and (d) QoE is objectively assessed using the ITU-T Rec. P.1203 which has been recently standardised for the quality assessment of ABR video.

III. Measurement Setup

A. Probe infrastructure

The experiment was conducted using the MONROE platform, which employs probes, in the form of a mini-computer, and a software framework that is responsible for the orchestration of the measurements and for the collection, analysis, visualization and sharing of the measurements. The probe is connected with 3 MiFis which provide a connection with the three major mobile operators in each country (Norway, Sweden, Italy, and Spain) and a GPS with external antenna. In total, there are 150 mobiles nodes and 100 nodes positioned in fixed places. In this study, 3 nodes in each country were used for the measurements. The experiments can be run in the form of a Linux container (Docker), running on Debian Linux operating system. Thus, there is full access to the resources of the computers and external applications can be run in parallel with the measurements.

B. Active measurements setup

The concept of the proposed experiment is depicted in Figure 2. The probes periodically request a web-page with an embedded YouTube player through a bash script; a set of predefined videos (via the 11-character videoID) is provided with the URL and parsed by the web-page to fetch the respective video from YouTube. We have pre-selected a set of 3 videos with a duration of approximately 2 minutes and typical video content (one movie trailer, one music video clip, and one baseball scene). The videos are fetched in HD-720p (1280×720) resolution, because the CPU of the probes cannot playback HD-1080 videos smoothly. Moreover, the videos are fetched every 40 minutes (thus 3 measurements every two hours) from the same operator. Since the node is connected with 3 MiFi interfaces, the routing is alternated in each video session (with a break of 5 minutes between ISP switches) to use the connection of all 3 mobile operators.

We employ the Selenium framework for automated web-browser testing and the X virtual frame buffer (Xvfb) for headless browser testing. This setup enables the playback with minimal CPU and GPU capabilities, which enables the use of lightweight and economical computers for the active probing. The web-page with the embedded player employs the YouTube IFrame API [3] which enables the constant monitoring of the player status and provides callback functions for a variety of events, such as, the time when the video player was loaded, the playback start time, the playback status (playing/stalling/stopped), and the playback quality (i.e. video resolution). By constantly monitoring these parameters through Javascript, the video playback performance can be monitored and the following application-layer KQIs, which are directly related to the user experience, are collected:

- player load time
- startup delay
- time and duration of stalling events
- quality switches during playback
- buffer fill level (i.e. fraction of video bytes downloaded)

The use of the actual YouTube player provides the opportunity to use the exact setup of a typical viewer, since there are no assumptions about the buffer behaviour (as in the case of youtube-dl[4] for example).

In addition to the above KQIs, the traffic flow is inspected using the proprietary "*StreamOwl OTT probe*", which is a network sniffer that can passively perform Deep Packet Inspection (DPI) on TCP flows and extract the following network-layer parameters:

- throughput
- number of TCP re-transmissions
- number of lost packets

Additionally, the network sniffer can detect the YouTube cache that was selected for delivering the video to the user in order to identify the caching strategy and policy. Using a geo-location service (e.g. telize[5]) we can infer the location of the cache and whether the video is fetched from a transit connection to the specific ISP under investigation.

The raw data from the measurements are collected locally at the end of each video session on each probe and then transferred to a central management system for centralised storage and processing.

C. QoE Assessment

The ITU-T Rec. P.1203 model predicts the impact of audiovisual quality variations and stalling events on quality experienced by the end user in multimedia mobile streaming and fixed network applications using adaptive bitrate streaming, based on a previous estimation of audio and video quality and information on startup delay and stalling events during the media session. The model predicts the Mean Opinion Score (MOS) on a 5-point Absolute Category Rating (ACR) scale as a final audiovisual quality MOS score.

Each individual stalling event has a weight assigned to it depending on the time between evaluation time and when the stalling event occurs. This length of the stalling event

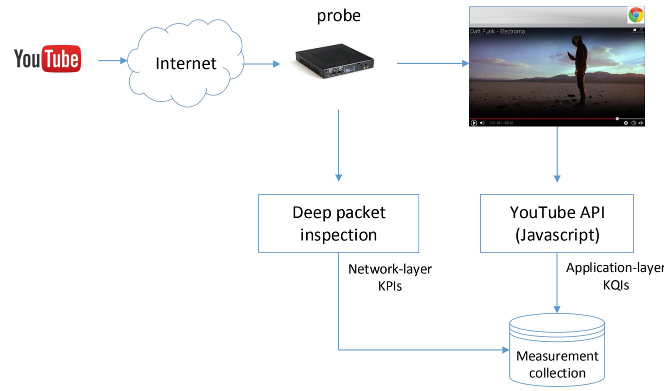

Fig. 2. Block diagram of the active probing infrastructure.

[3]https://developers.google.com/youtube/iframe_api_reference

[4]https://rg3.github.io/youtube-dl/

[5]https://www.telize.com/

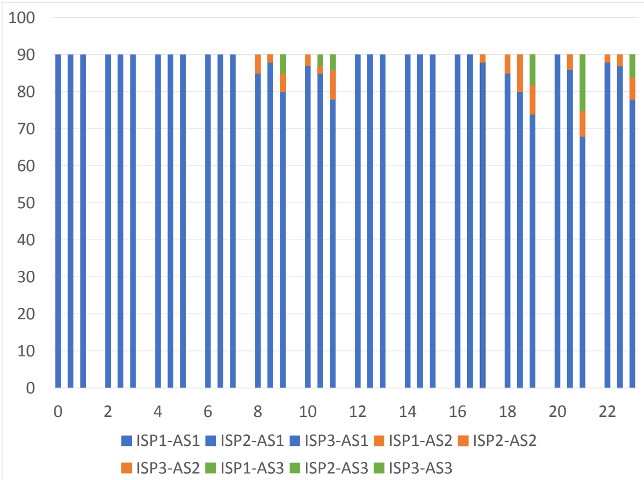

Fig. 3. Cache server selection during the day in Norway.

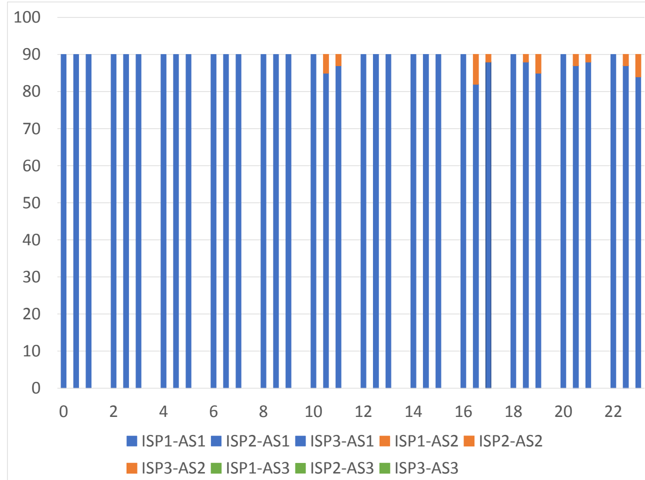

Fig. 4. Cache server selection during the day in Sweden.

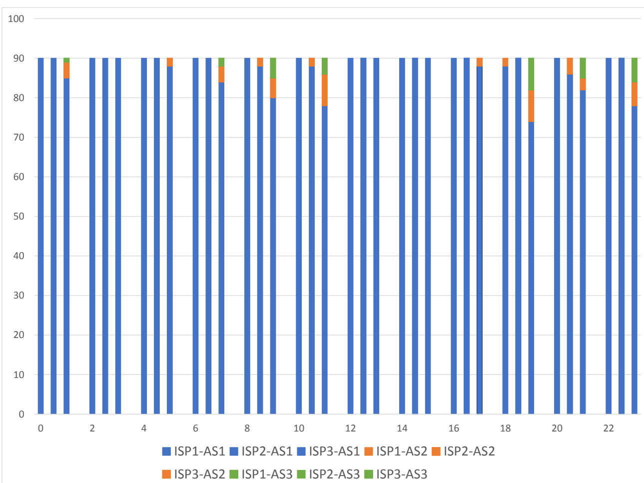

Fig. 5. Cache server selection during the day in Spain.

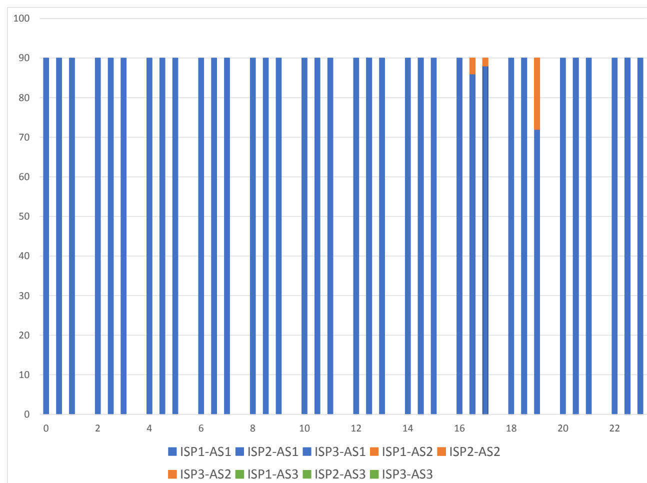

Fig. 6. Cache server selection during the day in Italy.

Fig. 7. Cache server selection during the day per country.

is multiplied with the weight when calculating the the total stalling duration impact on perceived quality. The weight $w_s tal_i$ of the i-th stalling event is computed as:

$$w_s tal_i = a_1 + (1 - a_1) * exp^{-\left(x * (\frac{log(0.5)}{a_2})\right)} \tag{1}$$

where x is the time difference between the timestamp of the stalling event and the end of the video (i.e. if a 20-sec video stalls at 8-th second, then x is equal to 20-8=12), and a_1, a_2 are coefficients equal to 0.484 and 10, respectively.

The model also takes into account the total number of quality direction changes, the difference between the maximum and minimum video quality scores, and the duration in seconds of the longest period without positive or negative quality change (stable quality). Finally, a machine learning algorithm takes into account the following features: number of stalling events, the total duration of all the stalling events, the relative ratio of stalling duration over the video length, the time of the last stalling event, and the weights of each stalling event. For the sake of brevity, the algorithms are only briefly described here, the interested reader is referred to the Recommendations for a detailed analysis of the video quality assessment algorithm. The output of the model is the audio-visual MOS score at the end of the video and is used in this paper to objectively estimate the quality of the video.

IV. Performance evaluation

A. Host infrastructure

Firstly, we investigate whether the cache server which is used to fetch the video to the client has an impact on quality. The cache server is identified by monitoring the redirection of the TCP using DPI on the traffic flow. Since the probes are connected to 3 MBB operators the aim is to investigate if clients from different ISPs get directed to the same cache server or not. Moreover, the aim is to evaluate whether the cache assignment per ISP is constant during the day since previous studies have shown that during peak times the video may be fetched from a different cache.

The results for each country are depicted in Figure 3 for all the days of the measurements; the vertical bars always equal 90, since we perform 3 measurements every two hours (the break between sessions is 40 minutes), for 30 days. The horizontal line depicts the hour of the day; the bars are clustered in even-hours during the day for legibility purposes. The vertical stacked bars give an indication of how stable the assigned cache server per ISP is, as denoted by the Autonomous System (AS) number. If a vertical bar does not contain values of the two other ASes it means that the video is always fetched from the same cache for all the days of the experiment.

In Figure 3, the results for the cache server selection in Norway are presented. AS1 is placed in Oslo, i.e. in the same city as the probe. AS-2 is placed in Oslo, and it belongs to the Google CDN, while AS-3 is placed in London, thus it is obvious that the video is server via a peering connection. The first remark is that the transit connection appears mostly with operator-3 during the peak times 8-12am and 6-10pm. The second remark is that operator-1 almost always delivers the video via AS-1 while operator-2 uses AS-2 for approx. 5% of the videos during peak hours.

In Figure 4, the cache server selection for Sweden is presented. It is noted that AS-1/2/3 in Norway is different than the AS-1/2/3 in Sweden and AS-1/2/3 in Italy or Spain; the notation 'AS-x' is used to denote an Autonomous system in the measurement campaign that was conducted in each country. Thus, AS-x in the Norway campaign corresponds to a different AS than the AS-x in the Sweden campaign. Similarly, ISP-x in Norway is different that ISP-x in Sweden.

It can be seen that the video is always fetched from AS1 (which corresponds to AS- and is placed in Karlstad, the place where the probe is located, while AS-2 belongs to the Google CDN. Moreover, it can be observed that during peak hours the operator-2 and operator-3 deliver some videos from AS-2, while operator-1 always delivers the video from AS1 (n.b. the "operator-x" in one country is obviously different than the "operator-x" in another country, and AS-y in one country is different that AS-y in another country.).

In Spain, as depicted in Fig. 5, there is more volatility during the day for operator-3, which employs 3 ASes both during peak hours and non-peak hours. Instead, operator-1 always uses a local cache, placed in Madrid, corresponding to AS-1 and corresponds to Google CDN, and operator-2 uses AS-2 during the peak hours, in addition to AS-1, which is also placed in Madrid but does not belong to the Google CDN. AS-3 is placed in Paris and is only used by operator-3, which is an indication that there is a peering connection.

Finally, in Italy, as depicted in Fig. 6, there is a much more stable cache server selection. Operator-1 always uses the AS-1 from the Google CDN and the same holds for the vast majority of the cases for operator-2 and operator-3 with a small exception for operator-3 during 6-8pm. The second AS-2 that is used for the delivery of the content is placed in Torino, the same city as the probe.

From these results it is evident that there is a different cache selection policy in each country, and in some of them the video is even fetched via a peering connection from another ISP at another country. Also, peak hours seem to have an effect on the switch between Google CDN and a local cache.

B. QoE performance evaluation

Besides the investigation of the host infrastructure, it is also important to study the impact of the cache server on the delivered quality, calculated as described in Section III-C. The distance of the cache server and its capacity may affect the delivery to a multitude of users and may result to higher startup delay, stalling events, and switching to a lower quality during playback. From the collected results, we observed that the number of video sessions in which the playback stalled during playback are less than 5%. However, there were many cases (ranging from 5% to 30%) where the video playback was not sustained at HD-720p but was scaled down to 480p or even 360p. The MOS values for each of the operators in each country, depending on the cache server are presented in Table 1. To highlight how the cache server affects the delivered quality, we have not aggregated the results per operator, but rather we present the results for each operator and AS individually. The bold figures in the table indicate the hightest MOS, i.e. the highest perceived quality by the viewers.

From this table, it can be seen that, in Norway, AS-1 and AS-2 perform equally, while AS-3 performs slightly worse, which is attributed to the fact that the cache is located in another country and the video is fetched through a peering connection. However, the quality does not drop significantly. In Sweden, AS-1 and AS-2 also perform equally for all operators. It is reminded again that AS-1/2/3 in Norway is different than the AS-1/2/3 in Sweden and AS-1/2/3 in Italy or Spain; the notation 'AS-x' is used to denote an Autonomous system in the measurement campaign that was conducted in each country. In Spain, AS-1 and AS-2 perform equally well, however AS-3 exhibits the lowest MOS value. Again, this is attributed to the fact the the server is located far from the client and the video is served through a peering connection. Finally, in Italy, where there was the most stable cache selection, the

TABLE I
AVERAGE VIDEO MOS PER OPERATOR AND AS.

Norway	operator 1	operator 2	operator 3
AS-1	**4.54**	4.38	4.41
AS-2	4.34	4.46	4.52
AS-3	n/a	3.78	3.28
Sweden	**operator 1**	**operator 2**	**operator 3**
AS-1	4.48	**4.52**	4.39
AS-2	n/a	4.42	4.41
AS-3	n/a	n/a	n/a
Spain	**operator 1**	**operator 2**	**operator 3**
AS-1	**4.54**	4.46	4.50
AS-2	4.48	4.38	4.36
AS-3	n/a	n/a	3.62
Italy	**operator 1**	**operator 2**	**operator 3**
AS-1	4.48	4.52	4.54
AS-2	**4.55**	4.37	4.48
AS-3	n/a	n/a	n/a

performance was equally good for all operators and ASes. From these results, it can be concluded that the main factor that may deteriorate the delivered quality is the fact that the video is served from a distant cache which means that the packets need to traverse several communication links and go through several bottlenecks. This information can be used by network operators in order to optimize the strategy for cache selection and the provision of CDNs for fast delivery of the content to the viewer.

V. CONCLUSIONS AND OUTLOOK

In this paper, we have presented the results of a field study for performance evaluation of YouTube streaming quality which was conducted for a month in four European countries, using probes with a connection to 3 mobile broadband operators. The video QoE was computed using standardised objective methods and was used to identify whether the cache infrastructure and the selection policy may have an impact on quality. Firstly, it was found that there is a different cache selection policy in each country, and in some of them the video is even fetched via a peering connection from another ISP located at another country. It was also found that during peak hours, there is a different selection of cache for a percentage of the videos. Secondly, it was found that the main factor that may deteriorate the delivered quality, in terms of perceived quality, is the fact that the video is served from a distant cache which means that the packets need to traverse several communication links and go through several bottlenecks.

We plan to continue these field measurements over the next six months and correlate them with more low-level network parameters which are collected by the probes. This will assist in identifying what are the bottlenecks in the delivery chain and provide useful insights for the optimization of the service.

ACKNOWLEDGMENT

This work is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644399 (MONROE) through the open call project NESTOR. The views expressed are solely those of the authors.'

REFERENCES

[1] *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*, International Telecommunication Union, Nov 2016.

[2] Ö. Alay, A. Lutu, D. Ros, R. Garcia, V. Mancuso, A. F. Hansen, A. Brunstrom, M. A. Marsan, and H. Lonsethagen, "Monroe - measuring mobile broadband networks in europe," in *Workshop on Research and Applications of Internet Measurements (RAIM) in Cooperation with ACM SIGCOMM*. IRTF ISOC, October 2015.

[3] P. Casas, P. Fiadino, A. Bär, A. D. Alconzo, A. Finamore, and M. Mellia, "Youtube all around: Characterizing youtube from mobile and fixed-line network vantage points," in *European Conference on Networks and Communications (EuCNC)*, August 2014.

[4] M. Zink and et al., "Characteristics of youtube network traffic at a campus network - measurements, models, and implications," in *Computer Networks*, 2009.

[5] P. Casas, P. Fiadino, A. Sackl, and A. D. Alconzo, "Youtube in the move: Understanding the performance of youtube in cellular networks," in *Proceedings of the IFIP Wireless Days Conference*, 2014.

[6] A. Finamore, M. Mellia, M. M. Munafo, R. Torres, and S. Rao, "Youtube everywhere: Impact of device and infrastructure synergies on user experience," in *ACM IMC - Internet Measurement Conference*. IRTF ISOC, November 2011.

[7] D. Giordano, S. Traverso, L. Grimaudo, M. Mellia, E. Baralis, A. Tongaonkar, and S. Saha, "Youlighter: An unsupervised methodology to unveil youtube cdn changes," in *International Teletraffic Congress*. IRTF ISOC, September 2015.

[8] P. Casas, A. D. Alconzo, P. Fiadino, A. Bär, and A. Finamore, "On the analysis of qoe-based performance degradation in youtube traffic," in *Int. Conf. on Network and Service Management (CNSM)*. IEEE, January 2015.

[9] A. K. Dziugas Baltrunas, Ahmed Elmokashfi and Ö. Alay, "Investigating packet loss in mobile broadband networks under mobility," in *IFIP Networking Conference (IFIP Networking) and Workshops*, June 2016.

[10] F. Kaup, F. Michelinakis, N. Bui, J. Widmer, K. Wac, and D. Hausheer, "Assessing the implications of cellular network performance on mobile content access," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 168–180, March 2016.

[11] A. Lutu, Y. R. Siwakot, Ö. Alay, D. Baltrunas, and A. Elmokashfi, "Profiling mobile broadband coverage," in *IFIP International Workshop on Traffic Monitoring and Analysis (TMA)*, April 2016.

[12] Ö. A. et al., "Measuring and assessing mobile broadband networks with monroe," in *Int. Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, June 2016.

[13] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "Yomoapp: a tool for analyzing qoe of youtube http adaptive streaming in mobile networks," in *European Conference on Networks and Communications (EuCNC)*, August 2015.

[14] G. Gomez, L. Hortiguela, Q. Perez, J. Lorca, R. Garca, and M. C. Aguayo-Torres, "Youtube qoe evaluation tool for android wireless terminals," *EURASIP Journal on Wireless Communications and Networking*, vol. 164, 2014.

[15] F. Wamser, P. Casas, M. Seufert, C. Moldovan, P. Tran-Gia, and T. Hossfeld, "Modeling the youtube stack: From packets to quality of experience," *Computer Networks*, vol. 109, no. 2, pp. 211–224, November 2016.

[16] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "Application-layer monitoring of qoe parameters for mobile youtube video streaming in the field," in *IEEE Int. Conf. on Communications and Electronics (ICCE)*, July 2015.